Journal of Sustainable Renewable Energy Innovations and Practices

ISSN: XXXX-XXXX

Volume 01 Issue 01 (November) 2024

Ecological Risk Identification and Assessment in Geothermal Energy Projects Using GIS and Random Forest Algorithm for Advanced Environmental Analysis

Rizal Mustafa[1], Aishah Noraini[2]

[1]Department of Anthropology and Sociology, Universiti Malaya, Malaysia
[2]Department of Computer Applications, Universiti Kebangsaan Malaysia, Malaysia

Abstract:

Geothermal energy (GE) is an alternative renewable energy source with a low carbon footprint but has ecological risks if not managed with due care. Therefore, this research identifies and accesses the potential environmental risks associated with geothermal energy projects that would impact local ecosystems through subsurface disturbances, water contamination, and land-use changes. Most traditional approaches to risk assessment lack the required precision to handle such complex spatial data and the environmental variables that may be involved in geothermal projects. Therefore, a new framework, GISRF-GE, combining advanced environmental analysis with Geospatial Information Systems (GIS) technology and the Random Forest (RF) algorithm, is proposed in this study for handling complex environmental data, which is multilayered, on the level of soil pollution, geography, land use patterns, and ecological indicators. Combining the spatial analysis capability of GIS with the predictive power of RF allows the model to identify and assess potential ecological risks around geothermal sites accurately. The GISRF-GE approach enables the analysis of complex risk patterns, indicating unexpected high-risk areas in such vulnerable ecosystems. Additionally, GIS visualizations enhance these patterns to better understand them for more site-specific and targeted planning in environmental management. Compared to traditional risk assessment models, this method offers a 30% increase in accurately predicting contamination hotspots to ensure a far better appreciation of any prevailing risk. This new framework gives stakeholders a better way to conduct an ecological risk assessment of geothermal projects and enhances the development of mitigation strategies that are ecologically effective but also cost-effective. These strategies would improve practical effectiveness in the restoration of sites and management of resources.

Keywords: Geothermal energy, Ecological Risk Identification and Assessment, Geospatial Information Systems, Random Forest

1. Introduction

The insight and guardianship of Earth's natural ecosystems is contingent on the nature reserve. To name a few of the important ecosystem services that nature reserves provide—clean water, timber, biodiversity conservation: all essential to human life [1]. Landscape pattern formation is affected by both natural and anthropogenic impacts that are unfavourable for their environment. They also pose ecological risks to landscapes, altering certain landforms' structure, functioning, and composition [2]. Sediments in coastal and estuarine ecosystems are where you can frequently come across heavy metals (HMs). But these same rock types can serve as a source for HMs towards various aquatic organisms when their habitat transforms. Then these are metals that can make their way into our diet through food chains and potentially harm us as well [3]. The evaluation of heavy metals risk assessment by utilizing plant biotic response measures can provide insights into metal bioavailability and its effect on the natural state of aquatic ecosystems, particularly [4]. Sediment can serve as one of the largest sources of HM contamination and provide a vehicle for transporting HM contamination. There is a growing ecological concern for aquatic creatures and human beings as HMs enter the water column from sediments through chemical and biological processes [5]. Sustainable human-earth system growth and ecological security cannot be achieved without the scientific control of these environmental threats [6].

Evaluating the likelihood and severity of potential negative impacts that human actions could have on the environment is the primary goal of an environmental risk assessment (ERA) [7]. Opportunities for effective land management can be identified by redeveloping numerous brownfield sites. The pressing necessity to transform and rejuvenate these neglected areas is highlighted by the rapid increase in land utilized for human habitation, which is anticipated to double by the year 2050, as reported by the United Nations [8]. Contaminated land remediation (CLR) through nature-based solutions (NbS) has recently become more prominent due to its various benefits, extending beyond merely minimizing human exposure to pollutants. One definition of NbS remediation systems is "strategies inspired and supported by nature, simultaneously providing human well-being and biodiversity benefits" [9]. Generating predictive maps using geostatistics relies heavily on machine learning, efficiently analysing data patterns. These maps influence decisions on additional exploration or sample procedures, which are crucial in establishing the extent and magnitude of soil contamination [10]. Quantitative molecular structure descriptions and biological activity predictions are made using a random forest model. When compared to decision trees, partial least squares (PLS), and support vector machines (SVM) without parameter optimization, the random forest model performed better using six available data sets [11]. With robust generalizability, random forest can handle big and multi-dimensional learning sets. Overfitting is less common in random forests than in other statistical learning models [12]. Figure 1 shows the image data of soil pollution (heavy metal) in Europe [22].

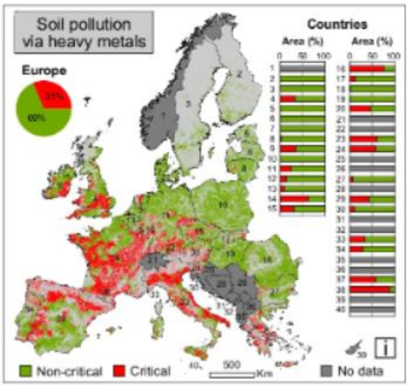

Fig.1 Soil pollution via heavy metal [22]

Spatial models of polluted sites, pinpointing problem regions, and assisting with remediation plan formulation are all possible using GIS. The capacity to manage data with multiple layers is a major strength of GIS. Soil type, distance from pollution sources, past land use, and other ecological markers are some environmental inputs that RF can use to forecast contamination levels and locate ecological danger zones. Since contamination and its environmental effects are caused by many factors, ecological risk assessment is an ideal application for the algorithm's data-processing and interaction-aware capabilities. To improve the accuracy and completeness of ecological risk assessment in Geothermal energy projects, the GISRF-GE methodology combines GIS with RF. The initial phase involves collecting comprehensive environmental data, including levels of soil contamination, geographical characteristics, land utilization trends, and ecological indicators. Subsequently, the GIS platform organizes and spatially analyzes this data to pinpoint potential concerns. The following phase entails applying the RF algorithm to the GIS data to predict ecological risks and pollutant concentrations.

The key importance of this study lies in the following aspects:

- To introduce a robust GISRF-GE framework combining GIS and Random Forest, enabling advanced environmental analysis for geothermal energy projects.
- To improve risk assessment accuracy by 30% compared to traditional methods, providing precise identification of ecological risk hotspots around geothermal sites.
- To effectively handle multi-layered environmental data—such as soil pollution, geography, and ecological indicators—for comprehensive geothermal project risk assessment.
- To identify previously unexpected high-risk areas in sensitive ecosystems, helping stakeholders proactively address the environmental impacts of geothermal activities.
- To leverage GIS visualization tools within the framework for better interpretation, allowing targeted and site-specific planning in environmental management.
- To develop ecologically effective and cost-efficient mitigation strategies, supporting sustainable site restoration and responsible resource management in geothermal energy projects.

The GISRF-GE framework combines GIS with the Random Forest algorithm to enhance the ecological risk assessment of geothermal energy projects. The framework integrates robust spatial analysis using GIS with the predictive solid power of RF to improve the accuracy of risk assessment by as much as 30% compared to traditional models, pinpointing high-risk areas. The model handles multilayer environmental data, including soil pollution, geography, and ecological indicators for comprehensive impact analysis. GIS visualization further enables the understanding of multi-complex risk patterns, which assist in planning a target. This, in turn, supports sustainable mitigation strategies to ensure value-for-money principles in enhancing site restoration and responsible management of resources within geothermal projects.

2. Literature review

The potential for large-scale in-situ cleanup of crude oil polluted locations in Nigeria was addressed by Adesipo, A. A. et al. [13]. Agronomic measurements, regulatory standards, plant characteristics, cost estimation, site conditions, maintenance and operation, and the outcome of harvest plants were also detailed as practical considerations. Nevertheless, when conventional clean-up methods have been exhausted, phytoremediation can serve as a last "polishing step" on severely polluted soils. It can be mixed with vermiremediation and other similar methods for greater efficacy.

From 1986–2016, researchers in Iran's dry regions used a model created by Taghizadeh-Mehrjardi et al. [14] to forecast the amount of heavy metals that soil may absorb. A random forest (RF) model was used to investigate the association between soil-absorbed heavy metal georeferenced values and a collection of geographical predictors from digital elevation models and remote sensing data. The findings showed that the RF model could effectively map the heavy metal distribution with Fe(0.53), Mn(0.59), Ni(0.45), Pb(0.45), and Zn(0.60) as coefficients of determination, correspondingly.

Huang H. et al. [15] also suggested assessing human activity's influence on possibly hazardous soil constituents via multivariate statistical techniques (Spearman correlation analysis, Random Forest Analysis (RFA), and Principal Component Analysis (PCA)) to address the issue of soil quality deterioration resulting from human activity-led impacts. The results advanced knowledge of pollutant sources and their contributions to soil contamination while reaffirming the presence of solid relationships between human activity and metals such as Cd and Hg, as well as As, Pb, and Cr.

Torabi Haghighi et al. [16] developed a quantitative method for mapping land degradation (LD) by integrating benchmark models of human and socio-environmental factors and using machine learning methods, e.g., Generalized Linear Model (GLM), Support Vector Machine (SVM), Dragonfly Algorithm (DA), and Multivariate Adaptive Regression Splines (MARS). This method evaluated various algorithms using the Taylor diagram, receiver functioning characteristic, and Kappa index. A very high degradation risk was found in 19.16% of the entire area in the Pole-Doab watershed in LD risk maps produced using SVM, 19.29% in GLM, 21.76% in MARS, and 22.40% in DA.

Anifowose, B., and Anifowose, F. [17] suggested improving soil pollution management through the incorporation of machine learning (ML) techniques into Environmental Impact Assessments (EIA). This approach tackles the issue of inaccurate results caused by insufficient use of ML in soil research, which is especially prevalent in underdeveloped nations. By showing that ML models, particularly Random Forest, greatly enhance soil contaminant prediction accuracy, it elucidates complicated nonlinear correlations and provides superior insights for environmental governance.

Research conducted by Ai, J., et al. [18] examined the best spatial scale for analyzing changes in ecological risk to the landscape on Haitan Island from 2000 to 2020. As a result of urbanization, the impermeable ground has largely replaced crops and woodland on Haitan Island, altering the spatial patterns of land use and creating ecological risk. On Haitan Island, the sector with the lowest danger rose steadily by 68.53% and eventually became the most important.

To evaluate ecological hazards in dry regions, Gan L. et al. [19] suggested a methodology called the Ecological Risk Index (ERI). It makes use of agricultural and socio-economic data spanning from 1980 to 2020. It uses the PLUS model to examine ecological risk patterns over space and time, as well as their underlying causes and spatial variability, and to forecast these risks in the future under different conditions. Among the independent variables influencing this increasing ecological concern, results show human activity and changes in land use, demographics, GDP, distance from water and highways, and concentrations of government.

Majemite, M. T. et al. [20] reviewed current data analytics methodologies, synthesized their findings and provided critical feedback about how those mitigate geoenvironmental risks resulting from geological activities. The scope of the study comprehensively covered all development analyses of contemporary methodologies that focused on applying GIS, big data, predictive analytics, and machine learning in geological prediction and risk management. This integration enhances the risk assessments' reliability, efficiency, and comprehensiveness.

Li W. et al. [21] suggested one model for assessing environmental ecological risks, which can be served to model potential shifts in land usage in the future in the Selenga River Basin. This model addresses the lack of information about the interaction of land use change and its ecological effects. By simulating land use using the PLUS model and calculating environment ecological risk indices in different spaces from 1990 to 2040, we can obtain important variations in ecological risk.

3. Proposed Work

The proposed GISRF-GE framework combines GIS and Random Forest (RF) data analysis approaches to solve complex multi-layered environmental data issues for ERIA in LRPs. Figure 2 shows the proposed method's working flow in the following sections.

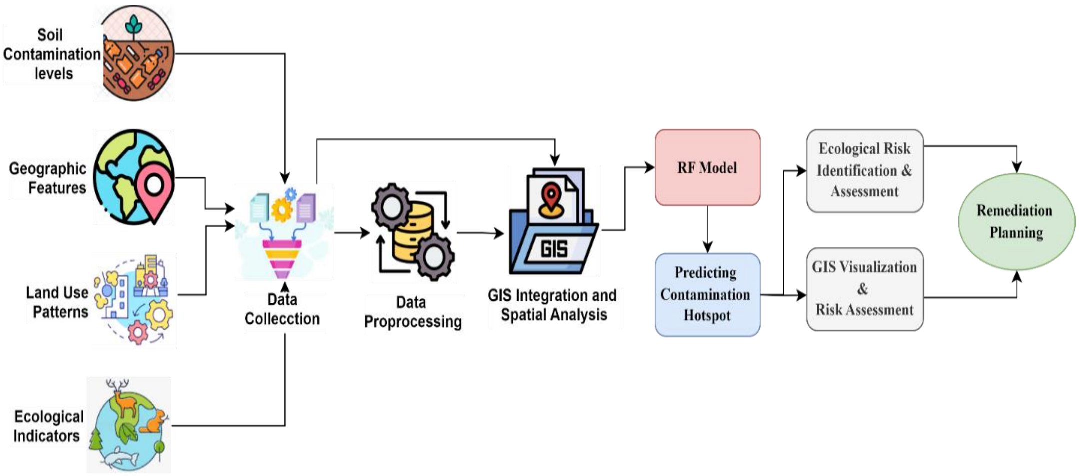
Soil Contamination levels
Geographic Features
Land Use Patterns
Ecological Indicators
Data Collecction
Data Proprocessing
GIS
GIS Integration and Spatial Analysis
RF Model
Predicting Contamination Hotspot
Ecological Risk Identification & Assessment
GIS Visualization & Risk Assessment
Remediation Planning

Fig.2 Working flow of the proposed GISRF-GE method

a) Data Collection

Soil contamination levels: They range from heavy metals, arsenic, and volatile organic compounds (VOCs) in the soil, which are environmental and health hazards from industrial processes.

Geographic Features: Topography, hydrology, slope, and elevation may be considered geographic features that affect soil erosion, water flow, and environmental contamination.

Land Use Patterns: Land usage and land cover comprise the past and current utilization of the land for industrial, agricultural, and residential purposes, which influence contamination levels and resource management.

Ecological Indicators: Factors in this category include species richness and composition measures, vegetation canopy, and the distance to water sources, which confirm ecosystems' status and susceptibility to pollution.

These data are obtained from monitoring programs in these environments, field surveys, or records.

b) Data Pre-processing

Data Organization in GIS Platform: Information sources of environmental data include field surveys, environmental monitoring systems, and current and historical records on aspects of soil contamination levels, geographical features, land use, and ecological features. This data is arranged in spatial layers in the GIS working platform, each covering a specific environmental aspect. These layers are georeferenced using longitude, latitude and, where appropriate, altitude or elevation. For instance, there could be a layer showing the quantity of heavy metals or VOCs in contaminated soil or another that would depict rivers or lakes in the area. By creating this spatial framework, other features can be integrated and analyzed to understand their relationships with the GIS platform.

Handling Missing Data: Environmental databases are usually full of such gaps because of factors such as parts of a survey being incomplete, some pieces of equipment breaking down, or simply the fact that some locations cannot easily be accessed, and such missing records can form quite a big portion of the total records, which if not well handled can greatly affect GIS models. There are measures used in handling missing data, as explained below. Interpolation, therefore, makes a forecast of values at a point using nearby others, for instance, using contamination levels of the neighbouring areas. Other missing value handling approaches, such as median imputation, mean imputation, or regression imputation, fill the missing values per some trends seen in the set. Where missing data are too numerous, exclusion may be used to eliminate bias in the analysis.

Data Normalization: Quantitative environmental data usually comes from different units and scales. For instance, the concentration of soil pollution is expressed in parts per million, whereas altitude is expressed in meters. Normalization scales these variables to the same measure in the GIS model.

Outlier detection and removal: Entries in a dataset far from the rest of the values are termed outliers. Errors in measurement or data recording may cause outliers or arise as a rare extreme case. In the context of energy conservation, this would mean an unusually high level of contamination, likely to skew the analysis results.

c) GIS Integration and Spatial Analysis

Layer Combination: In the Geographic Information Systems course, there is a combination of layers of diverse data in a spatial manner. For instance, a layer showing soil contamination can be overlaid with a layer showing hydrological features to study how bodies of water may affect or spread contaminants. This layering technique enables advanced spatial queries, such as identifying contaminated areas within a specific distance from water bodies or regions that exhibit high biodiversity.

Spatial Interpolation and Hotspot Analysis: Geographical information systems apply higher-order statistical methods to determine the degree of pollution in the environment and the vulnerability of the zones. This can be done using techniques like Kriging and Inverse Distance Weighting, which will estimate the contamination levels at unsampled locations from data collected in proximal locations. This method enables the creation of contamination surface maps as a continuous function while providing insights into regions with unavailable measurements. Also, GIS assists in identifying hotspots that are statistically more contaminated or have greater ecological risks. These methods also identify significant spatial relationships by pointing out specific areas of concern within the data set. Integrating these methods and techniques will allow the environment to identify areas of need for intervention or remediation and make efficient decisions on resource allocation to further the cause of environmental management and conservation.

Data visualization: Environmental mapping software develops illustrations of contamination, danger areas, and geographic factors. These maps are important in presenting large amounts of information to key decision-makers. For instance, a GIS map may overrepresent contamination intensity in an area, as the high-risk areas will be represented by red colours, which will assist the planners in determining where to begin the treatment processes.

Statistical Analysis in GIS: GIS RF is a statistical model that GIS can accommodate when integrated into the system. Remote-sensing data is used as the input data, allowing one to model contamination levels and risk factors in a spatial context. Thus, using spatial and statistical analysis integrated within GIS helps improve energy risk evaluations.

d) Model for Random Forests (RF)

An effective ML method for modelling and predicting contamination risk in contexts without direct measurements is the Random Forest (RF) algorithm. To forecast possible contamination levels, it uses environmental parameters and previous data. The RF model builds a "forest" of hundreds or thousands of decision trees. The final prediction is an average of the projections made by each tree. The input features are randomly divided among the trees. This helps avoid overfitting by introducing variety among the trees. A replacement-drawn random subset of the training data is used to train each tree. Bagging is a method that further improves tree diversity. Equation 1 shows the calculation for the average forecast of all decision trees.

$$p = \frac{1}{N}\sum_{i=1}^{N} T_i(X) \qquad \text{(Eq.1)}$$

where N is the total number of trees, $T_i(X)$is the prediction of ith tree in the forest.

Training and Testing Data

1. *Training Phase:* Environmental variables and known contamination levels are used to train the RF algorithm. The data is divided into two sets, one for training and one for testing, with an 80/20 split. Soil characteristics, topography, and land use are some of the features that are supplied as inputs.

2. *Cross-Validation:* Applying 10-fold cross-validation helps prevent overfitting. Each of the ten training iterations of the RF model uses one subset for testing and nine for training, dividing the dataset into ten smaller subsets.

e) Prediction of Contamination Hotspots

Soon after training, the RF model uses environmental variables and spatial patterns to predict the contamination levels in unmeasured locations. The product is probability maps of contamination hotspots. The following algorithm 1 shows the RF model training.

Algorithm 1 RF model for ecological risk identification

```
    Input: Environmental features (X1, X2, ..., Xn), known contamination levels (Y)
    Output: Predicted contamination levels and risk zones

    1. Initialize the Random Forest model with N trees.
    2. For each tree Ti in the forest:
        a. Randomly sample data from the training set (bootstrap sampling).
        b. Train a decision tree on the subset.
        c. Repeat for N trees.
    3. For each test sample X in the dataset:
        a. Pass X through each trained tree Ti.
        b. Record the prediction Ti(X).
    4. Compute the final prediction for each sample as the average prediction of all trees:
                        $p = \frac{1}{N}\sum_{i=1}^{N} T_i(X)$
    5. Generate a contamination risk map based on predicted contamination levels.
    6. Overlay GIS maps for spatial analysis and visualization of high-risk zones.
Return contamination hotspot map and risk analysis.
```

f) GIS-Based Visualization

After the Random Forest model predicts contamination levels, stakeholders are given powerful tools to examine and understand contamination risk zones. These outputs are seamlessly integrated with GIS visualization tools. Geographic information systems use numerous visualization methods to portray this data accurately. Heatmaps visually represent the likelihood of contamination using colour coding, usually ranging from green (low risk) to red (great danger). Contour maps demarcate regions with uniform contamination levels across various geographic locations to identify pollution gradients. Environmental sensitivity indicators and pollution levels are only two examples of the many data types that can be layered using GIS. This multi-faceted method allows for a thorough assessment of the environmental situation by showing temporal patterns of possible contaminant diffusion and pinpointing locations where vulnerable ecosystems meet high contamination risk. These visualization tools, when combined, form a powerful tool capable of imparting to ecological managers and remediation planners a sophisticated knowledge of hazards from contamination.

g) Ecological Risk Identification and Assessment (ERIA)

Coupling GIS spatial modeling with the prediction capability of RF algorithms enables developing an integrated approach for risk assessment in environmentally sensitive areas. Such combined method leverages a complex system of risk assessment that identifies ecological risk zones that were earlier unidentified. This is illustrated in the computation involved in risk prediction as follows: Equation 2.

$$Risk\ Score = \sum_{i=1}^{n} \omega_i \cdot X_i \qquad \text{(Eq.2)}$$

where X_i is the environmental factor, and ω_i is the weight assigned to each factor based on its impact on risk.

The RF model can predict the variable values for unmeasured locations using GIS. This model is trained on historical data and environmental features. Traditional risk assessment approaches may have missed certain zones of possible ecological peril. Still, the resultant risk map shows both those places and others already recognised as high-risk. These newly identified potentially dangerous regions are prioritised since they result from the intricate spatial interconnections of several environmental elements. Ecological risk management tactics may be made much more efficient and successful using this method, which allows for preventative actions to safeguard delicate ecosystems and forestall possible environmental harm.

4. Results and Discussion

a) Dataset explanation

Researchers and analysts in [23] working in the renewable energy sector have access to a wealth of information in the Global Renewable Energy and Indicators Dataset. This dataset includes renewable energy output, socioeconomic variables, and environmental indicators from around the world. Key features include:

Renewable Energy Data: This dataset details the production, installed capacity, and investments in renewable energy across multiple nations and years. It covers many renewable energy sources, including solar, wind, hydro, and geothermal. The data is presented in GWh, MW, and USD.

Socio-Economic Indicators: This category includes statistics on population, gross domestic product, energy consumption, carbon dioxide emissions, employment in the renewable energy sector, government policies, research and development spending, and renewable energy goals.

Environmental Factors: This section details the typical yearly weather conditions, including the average temperature, rainfall, wind speed, solar irradiance, hydro potential, geothermal potential, and biomass availability.

b) Performance metrics

In this section, the proposed method is compared with conventional methods like SVM [16], PLUS [19], and EIA [17]. Performance evaluation in binary risk classification will involve several key performance indicators for each model, including the Root Mean Square Error, Mean Absolute Error (MAE), R-Squared (R^2) and the Area Under the Receiver Operating Characteristic curve (AUC-ROC).

Root Mean Square Error (RMSE)

Perhaps the most common statistic used in evaluating model fit, or the differences between model predictions versus actual data, is the RMSE—Root Mean Square Error. This statistic can be particularly useful in quantifying the accuracy of models that predict ecological risks and other events that depend on spatial data from sources such as geographic information systems (GIS). The RMSE can be calculated using Equation 3.

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{k=1}^{n}(x_k - x_k')^2} \quad \text{(Eq.3)}$$

where x_k the actual observed value, x_k' predicted value from the model.

Figure 3 compares Root Mean Square Error (RMSE) between EIA, GISRF-LRP, SVM, and PLUS. Regarding ecological risk assessment, the GISRF-GE model demonstrates the highest prediction accuracy and the lowest RMSE, establishing it as the clear leader. It integrates geographical data with the Random Forest algorithm for more precise risk detection; the GISRF-GE method improves decision-making in geothermal energy projects by decreasing errors.

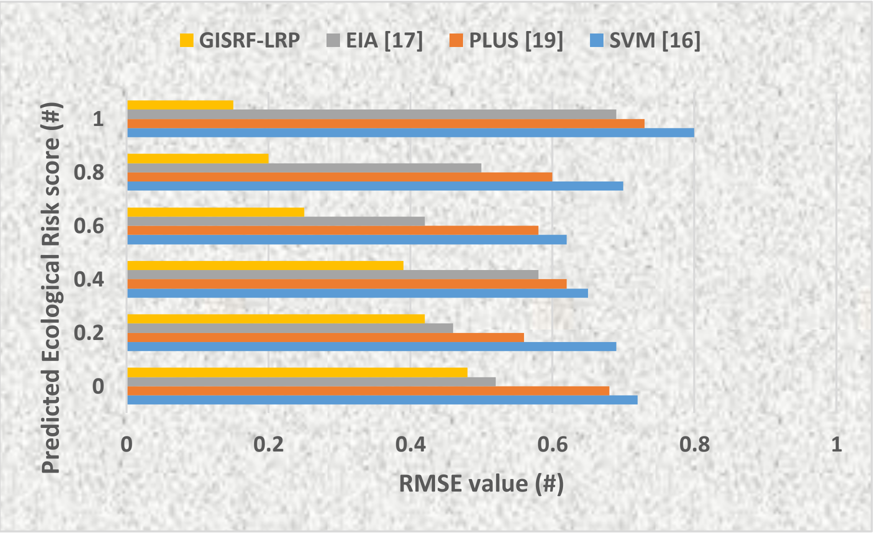

Fig.3 RMSE analysis

Mean Absolute Error (MAE)

The statistical measure known as MAE calculates the typical size of the disparities between predicted and actual values, ignoring the path of the errors. MAE may be used to measure prediction accuracy easily and linearly. This is achieved by equation 4.

$$MAE = \left(\frac{1}{n}\right)\sum_{k=1}^{n}|x_k - x_k'| \qquad \text{(Eq.4)}$$

where x_k is the true measured value, x_k' is the model's anticipated value, n is an aggregate count of observations, and $|x_k - x_k'|$ absolute difference between predicted and actual value.

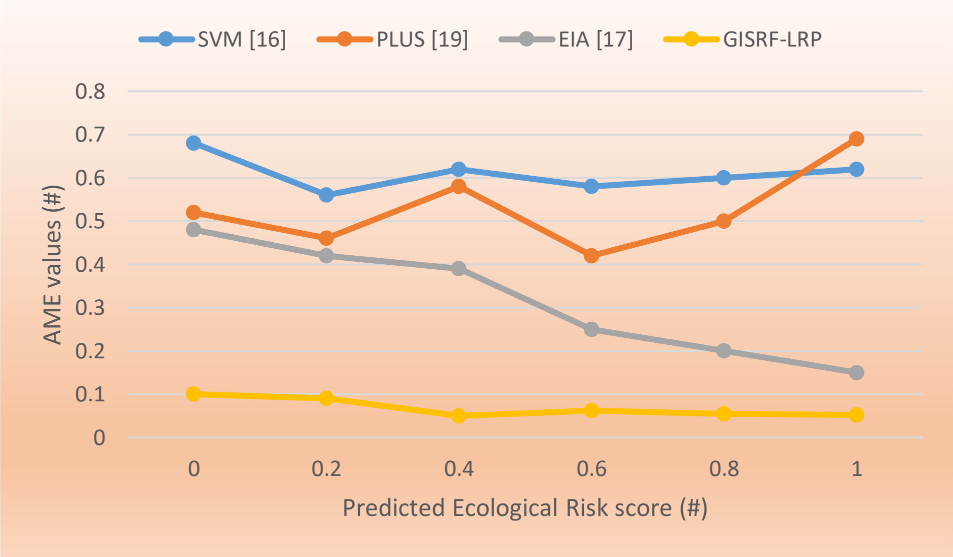

Fig.4 MAE analysis

Figure 4 compares the Mean Absolute Error (MAE) for ecological risk prediction of several models, including SVM, PLUS, EIA, and GISRF-LRP. GISRF-GE repeatedly proves it is the most effective at reducing average prediction errors by maintaining its position as the model with the lowest MAE. This proves that, compared to conventional methods, the model is superior at generating accurate risk estimations. More accurate identification of high-risk regions and better allocation of resources for focused environmental management are made possible by GISRF-LRP's reduction of mistakes, which improves decision-making in geothermal energy projects.

R-Squared (R2)

An important statistic for any statistical model is the R-squared (R²), sometimes called the coefficient of determination. This statistic demonstrates how effectively the independent variables (predictors) account for the variation in the dependent variable (outcome). This means that the independent factors can explain the dependent variable's variance to a certain extent. This can be calculated using the equation 5.

$$R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}}\right) \qquad \text{(Eq.5)}$$

where SS_{res} is the sum of squared residuals, SS_{tot} is the total sum of squares. Figure 5 shows the findings of the R-squared (R²) study, which evaluates the eco-risk variance explanation models provided by SVM, PLUS, EIA, and GISRF-LRP. The GISRF-GE model surpasses the others in capturing data variability, as indicated by its highest R² values. This suggests that GISRF-GE more effectively uses GIS spatial analysis and the predictive abilities of Random Forest, leading to a deeper understanding of environmental risk factors. The model's increased accuracy and predictive strength enable better decisions regarding land cleanup and targeted interventions, which eventually result in more reliable ecological risk assessments.

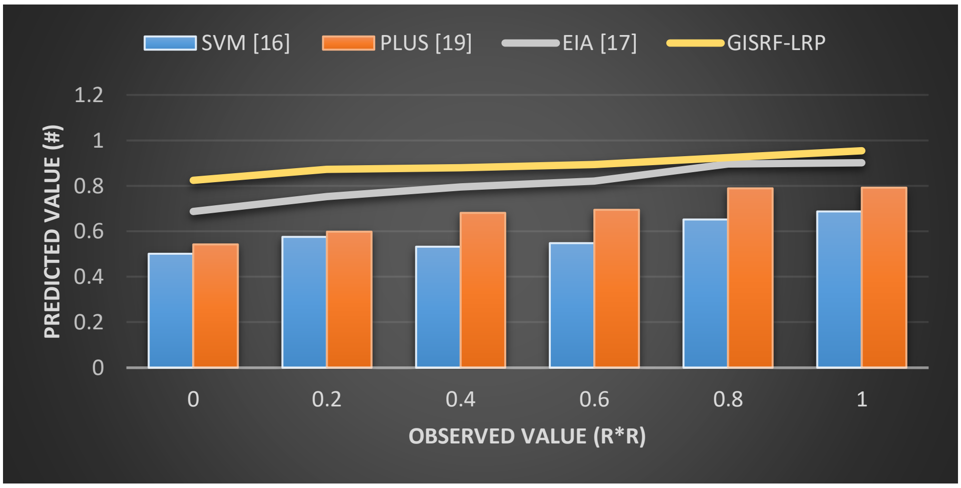

Fig.5 R² Analysis

AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

AUC-ROC has been applied to several studies assessing the performance of various binary classification algorithms. This metric gives insight into how well a model can separate two classes. In ecological risk assessment studies, the AUC-ROC is utilized to make decisions about a model's ability to classify areas into high and low-risk zones.

A receiver operating characteristic curve is a method applied to compare sensitivity for a range of thresholds, the sensitivity of classification against specificity related to the false positive rate. A ROC curve plots different cutoffs or thresholds to determine whether to classify a prediction as positive or negative.

Sensitivity or true positive rate: the ratio of high-risk areas that were identified as positive by the algorithm. It was computed using equation 6.

$$TPR = \frac{True\ positive}{True\ Positive + False\ Negative} \qquad \text{(Eq.6)}$$

False Positive Rate (FPR): Percentage of low-risk areas which model gets it wrong and expects to be positive. Obtained by equation 7.

$$FPR = \frac{False\ positive}{False\ Positive + True\ Negative} \qquad \text{(Eq.7)}$$

Figure 6 shows the outcome of the AUC-ROC analysis, presenting the performance of various models developed for environmental risk evaluation. Indeed, the highest AUC-ROC values are obtained by the GISRF-GE model, with a high reliability in mapping the separation between high-and low-risk zones. It outperforms other models such as SVM, PLUS, and EIA by a large margin. In this respect, GISRF-GE is the best model for ecological risk mapping, especially under complex conditions.

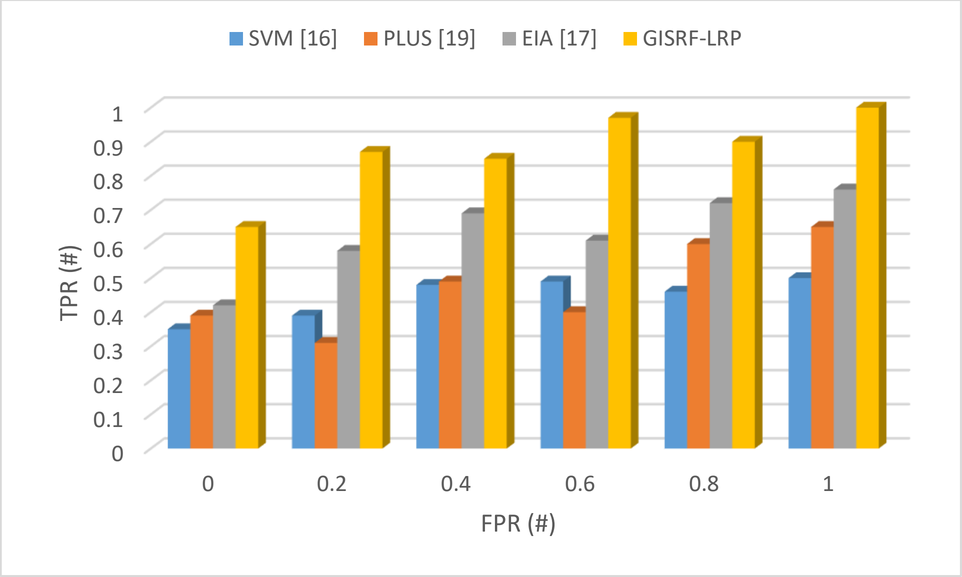

Fig.6 AUC-ROC Analysis

5. Conclusion

The proposed method, GISRF-GE, integrates GIS into the Random Forest algorithm to enhance ERIA in geothermal energy projects. Integrating the spatial analysis conducted in GIS with the predictive capabilities of RF allows for an in-depth and more accurate assessment of the contamination levels across different environmental layers, such as soil strata, geographical formations, and land use patterns. Specifically, GISRF-GE is helpful for ecological risk assessment, and exhibit an improvement of up to 30% if compared to results obtained using more conventional methodologies. By allowing more accurate risk assessments, better resource allocations, and advancement in sustainable and cost-effective solutions concerning ecological risk management, GISRF-GE enhances decision-making in geothermal energy. The difficulty with this approach may arise when environmental data is unavailable or incomplete for areas. That lack of data could reflect on the precision of the models themselves. For more precise predictions to be made and to make those relevant under a wide range of environmental conditions, future studies should cover both the extension toward new machine learning algorithms and the enlargement of the data sources used.

References

[1]. Wang, H., Liu, X., Zhao, C., Chang, Y., Liu, Y., & Zang, F. (2021). Spatial-temporal pattern analysis of landscape ecological risk assessment based on land use/land cover change in Baishuijiang National Nature Reserve in Gansu Province, China. *Ecological Indicators*, *124*, 107454.

[2]. Tan, L., Luo, W., Yang, B., Huang, M., Shuai, S., Cheng, C., ... & Hu, C. (2023). Evaluation of landscape ecological risk in key ecological functional zone of South–to–North Water Diversion Project, China. *Ecological Indicators*, *147*, 109934.

[3]. Tian, K., Wu, Q., Liu, P., Hu, W., Huang, B., Shi, B., ... & Wang, T. (2020). Ecological risk assessment of heavy metals in sediments and water from the coastal areas of the Bohai Sea and the Yellow Sea. *Environment international*, *136*, 105512.

[4]. Aljahdali, M. O., & Alhassan, A. B. (2020). Ecological risk assessment of heavy metal contamination in mangrove habitats, using biochemical markers and pollution indices: A case study of Avicennia marina L. in the Rabigh lagoon, Red Sea. *Saudi journal of biological sciences*, *27*(4), 1174-1184.

[5]. Xiao, H., Shahab, A., Xi, B., Chang, Q., You, S., Li, J., ... & Li, X. (2021). Heavy metal pollution, ecological risk, spatial distribution, and source identification in sediments of the Lijiang River, China. *Environmental Pollution*, *269*, 116189.

[6]. Wang, K., Zheng, H., Zhao, X., Sang, Z., Yan, W., Cai, Z., ... & Zhang, F. (2023). Landscape ecological risk assessment of the Hailar River basin based on ecosystem services in China. *Ecological Indicators*, *147*, 109795.

[7]. Kaikkonen, L., Parviainen, T., Rahikainen, M., Uusitalo, L., & Lehikoinen, A. (2021). Bayesian networks in environmental risk assessment: A review. *Integrated environmental assessment and management*, *17*(1), 62-78.

[8]. Hou, D., Al-Tabbaa, A., O'Connor, D., Hu, Q., Zhu, Y. G., Wang, L., ... & Rinklebe, J. (2023). Sustainable remediation and redevelopment of brownfield sites. *Nature Reviews Earth & Environment*, *4*(4), 271-286.

[9]. Alshehri, K., Gao, Z., Harbottle, M., Sapsford, D., & Cleall, P. (2023). Life cycle assessment and cost-benefit analysis of nature-based solutions for contaminated land remediation: A mini-review. *Heliyon.*

[10]. Han, H., & Suh, J. (2024). Spatial Prediction of Soil Contaminants Using a Hybrid Random Forest–Ordinary Kriging Model. *Applied Sciences*, *14*(4), 1666.

[11]. Tan, K., Wang, H., Chen, L., Du, Q., Du, P., & Pan, C. (2020). Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. *Journal of hazardous materials*, *382*, 120987.

[12]. Liu, W., Zhang, Y., Liang, Y., Sun, P., Li, Y., Su, X., ... & Meng, X. (2022). Landslide risk assessment using a combined approach based on InSAR and random forest. *Remote Sensing*, *14*(9), 2131.

[13]. Adesipo, A. A., Freese, D., & Nwadinigwe, A. O. (2020). Prospects of in-situ remediation of crude oil contaminated lands in Nigeria. *Scientific African*, *8*, e00403.

[14]. Taghizadeh-Mehrjardi, R., Fathizad, H., Ali Hakimzadeh Ardakani, M., Sodaiezadeh, H., Kerry, R., Heung, B., & Scholten, T. (2021). Spatio-temporal analysis of heavy metals in arid soils at the catchment scale using digital soil assessment and a random forest model. *Remote Sensing*, *13*(9), 1698.

[15]. Huang, H., Zhou, Y., Liu, Y., Li, K., Xiao, L., Li, M., ... & Wu, F. (2020). Assessment of anthropogenic sources of potentially toxic elements in soil from arable land using multivariate statistical analysis and random forest analysis. *Sustainability*, *12*(20), 8538.

[16]. Torabi Haghighi, A., Darabi, H., Karimidastenaei, Z., Davudirad, A. A., Rouzbeh, S., Rahmati, O., ... & Klöve, B. (2021). Land degradation risk mapping using topographic, human-induced, and geo-environmental variables and machine learning algorithms, for the Pole-Doab watershed, Iran. *Environmental Earth Sciences*, *80*, 1-21.

[17]. Anifowose, B., & Anifowose, F. (2024). Artificial Intelligence and Machine Learning in environmental impact prediction for soil pollution management–Case for EIA Process. *Environmental Advances*, 100554.

[18]. Ai, J., Yu, K., Zeng, Z., Yang, L., Liu, Y., & Liu, J. (2022). Assessing the dynamic landscape ecological risk and its driving forces in an island city based on optimal spatial scales: Haitan Island, China. *Ecological Indicators*, *137*, 108771.

[19]. Gan, L., Halik, Ü., Shi, L., & Welp, M. (2023). Multi-scenario dynamic prediction of ecological risk assessment in an arid area of northwest China. *Ecological Indicators*, *154*, 110727.

[20]. Majemite, M. T., Dada, M. A., Obaigbena, A., Oliha, J. S., Biu, P. W., & Henry, D. O. (2024). A review of data analytics techniques in enhancing environmental risk assessments in the US Geology Sector. *World Journal of Advanced Research and Reviews*, *21*(1), 1395-1411.

[21]. Li, W., Lin, Q., Hao, J., Wu, X., Zhou, Z., Lou, P., & Liu, Y. (2023). Landscape Ecological Risk Assessment and Analysis of Influencing Factors in Selenga River Basin. *Remote Sensing*, *15*(17), 4262.

[22]. https://esdac.jrc.ec.europa.eu/themes/land-degradation
https://www.kaggle.com/datasets/anishvijay/global-renewable-energy-and-indicators-dataset