International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 2, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n2.02
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Integration of Artificial Intelligence and GIS: An Application Study Based on Machine Learning Random Forest and SHAP

Feng Li *

College of Xi'an International Studies University, Xi'an710100, China

ABSTRACT

The integration of Geographic Information System (GIS) technology with artificial intelligence opens up innovative pathways for geospatial analysis. This study provides theoretical support and practical guidance for the deep integration of GIS and artificial intelligence, contributing to the advancement of geospatial analysis towards intelligence-driven approaches. With the evolution of GIS technology and the widespread adoption of artificial intelligence, The integration of Geographic Information System (GIS) technology and artificial intelligence has opened innovative paths for geographic spatial analysis. This study provides theoretical support and practical guidance for the deep integration of GIS and artificial intelligence, facilitating the move towards intelligent geographic spatial analysis. With the evolution of GIS technology and the widespread adoption of artificial intelligence, this research explores their innovative fusion in geographic spatial analysis. It focuses on the potential applications of the SHAP model in land use change prediction, environmental risk assessment, and urban planning, covering stages such as data preparation, model construction, interpretation, and spatial visualization. The results show that the SHAP model can effectively analyze the decision-making mechanisms of machine learning models and intuitively present the impact of geographic factors through spatial visualization.

KEYWORDS

Artificial Intelligence; Geographic Information Systems (GIS); Machine Learning; SHAP Model; Random Forest

1. INTRODUCTION

Random Forest is an ensemble learning algorithm that improves the accuracy and stability of classification and regression by constructing multiple decision trees and combining their results. The basic idea is to introduce randomness into decision trees to generate multiple models, and then perform voting (for classification problems) or averaging (for regression problems) to achieve the final prediction result [1].

SHAP (Shapley Additive Explanations) is a model interpretation method based on the Shapley values in game theory. It aims to quantify the contribution of each feature to the prediction results of a machine learning model. The core concept is to fairly allocate the "credit" or "responsibility" of the model output to each input feature through mathematical distribution principles [2]. Whether it is global interpretation analyzing the overall model's feature importance (e.g., the average impact of features on predictions) or local interpretation of individual sample prediction results (e.g., the specific contribution of each feature in a given prediction), SHAP can clearly present the results.

Although GIS technology has unique advantages in processing and analyzing geographic spatial data, it faces limitations when handling complex data relationships and large-scale datasets. Artificial intelligence, particularly machine learning techniques, can uncover complex patterns and

relationships in the data. Combining the two can enhance the efficiency and accuracy of geographic spatial analysis, providing more scientific evidence for decision-making in related fields. This study aims to explore the integration of GIS and artificial intelligence, innovatively introducing the SHAP model into geographic spatial analysis to improve the interpretability of machine learning models, and providing new ideas for research and practice in related fields.

2. ORGANIZATION OF THE TEXT

2.1. Characteristics of Random Forest in Geographic Applications

2.1.1. High Accuracy

The Random Forest algorithm effectively reduces the variance of a single model by integrating the predictive results of multiple decision trees, thereby enhancing the model's generalization capability [3]. In the field of geography, Random Forest is capable of accurately capturing nonlinear relationships and interactions between features within the data when faced with complex and variable terrain characteristics and the interactions of multiple environmental factors. For instance, in land use type classification, numerous complex factors must be considered, including terrain (e.g., elevation, slope, aspect), climate (e.g., precipitation, temperature), and soil types. By comprehensively analyzing these multi-source geographic information, Random Forest can more accurately classify the land use types in different regions, providing reliable support for geographic research and land planning. Compared to traditional methods, it demonstrates higher accuracy and applicability.The Random Forest algorithm effectively reduces the variance of a single model by integrating the prediction results of multiple decision trees, thereby enhancing the model's generalization ability [3]. In the field of geography, where there are complex and variable terrain features and interactions among various environmental factors, Random Forest can accurately capture the nonlinear relationships and interactions between features in the data. For instance, when classifying land use types, it is necessary to consider multiple complex factors such as terrain (e.g., elevation, slope, aspect), climate (e.g., precipitation, temperature), and soil type. Random Forest, by comprehensively analyzing this multi-source geographic information, can more accurately classify land use types in different regions, providing reliable support for geographic research and land planning, demonstrating higher accuracy and applicability compared to traditional methods.

2.1.2. Strong Anti-overfitting Ability

In geographic research and applications, decision tree models often face the risk of overfitting, which means they fit the training data closely but have poor generalization on new data. The Random Forest algorithm addresses this issue by integrating multiple decision trees and introducing randomness, reducing this problem fundamentally. When constructing each decision tree, the algorithm proceeds with splits based on only part of the randomly selected samples and features, ensuring the uniqueness of each tree while reducing over-reliance on specific training data [4]. In land use classification tasks, where vast geographic features such as terrain, climate, and soil are involved, a single decision tree might overfit. However, Random Forest leverages the advantage of multiple trees, integrates prediction results through voting or averaging mechanisms, offsets mutual errors, corrects biases, and ultimately produces more robust and generalized classification outcomes, ensuring reliable application in practical geographic scenarios.

2.1.3. Ability to Handle High-dimensional Data and Large-scale Data

With its unique algorithmic mechanism, Random Forest can efficiently process large-scale datasets containing numerous features and samples, showing significant advantages in geographic research. In scenarios of geographic data analysis, such as processing hyperspectral remote sensing images, a single pixel may contain hundreds or even thousands of band information, leading to extremely high dimensionality. Random Forest employs ensemble learning strategies, effectively avoiding the "curse

of dimensionality" issue through random sampling and feature subset selection [5]. When classifying these high-dimensional data, Random Forest can accurately differentiate various land feature types, such as forests, water bodies, and farmland in land cover classification tasks, providing powerful tools for quantitative geographic research and spatial analysis, promoting the development of geographic information science in complex data processing.

2.1.4. Relative Interpretability

As an ensemble learning model, Random Forest exhibits relatively significant interpretability advantages compared to complex deep learning architectures like multilayer neural networks [6]. Researchers can thoroughly analyze each decision tree within the forest, exploring its node splitting criteria, feature selection preferences, and information distribution patterns. By delving into each decision tree's internal structure, analyzing its decision path from root to leaf node, and assessing each tree's influence on the final voting or averaging step, one can understand the overall decision mechanism of the model. In practical geographic applications such as land use classification tasks, visualization techniques and feature importance assessment tools can clearly display how Random Forest makes decisions based on multiple geographic features like terrain, climate, and soil, clarifying the contribution of each feature to the classification result. This provides geographic researchers with an easy-to-understand and logically clear decision-making basis, strongly supporting the scientific validation of related research conclusions.

3. CHARACTERISTICS OF SHAP IN GEOGRAPHIC APPLICATIONS

3.1. Strong Interpretability

SHAP values can globally quantify the contribution of each feature to the prediction results of the model, helping geographic researchers fully understand which geographic factors play critical roles in areas such as land use change prediction, environmental risk assessment, and urban planning. For example, in land use change prediction, the SHAP model can clearly determine the overall impact of factors like terrain slope, accessibility, and surrounding land use types on land use change trends, providing a basis for reasonable land use planning. For each specific prediction sample, such as predicting land use changes in a particular region or assessing environmental risks at a specific location, SHAP values can specifically show the size and direction (positive or negative) of each geographic feature's contribution. This enables researchers to gain deep insights into how the model makes specific decisions in different geographic contexts, further enhancing the model's interpretability and credibility in real-world geographic problem analysis.

3.2. Good Compatibility

The SHAP model can be combined with various commonly used machine learning algorithms, such as Random Forest and Gradient Boosting Trees. In geographic research, regardless of the machine learning model adopted for data mining and analysis, SHAP can serve as an effective interpretation tool that reveals the geographic logic and patterns behind the model [8]. For example, when using Random Forest for land cover classification, SHAP can explain the classification basis of each decision tree and the role of various features in the combined classification result, helping researchers better understand the model's decision-making process.

The output results of the SHAP model, such as feature importance and contribution values, can be seamlessly integrated with GIS software. By combining SHAP interpretation results with geographic spatial data, researchers can intuitively display the spatial distribution of the influence of different geographic factors on model predictions on maps. For instance, in environmental risk assessment, overlaying SHAP values with geographic spatial data can clearly show which regions are at higher risk due to the combined influence of factors such as terrain, river distribution, and soil type,

providing intuitive geographic information support for environmental management and decision-making.

3.3. Broad Applicability

Whether in land use change prediction, environmental risk assessment, or urban planning, SHAP models can play essential roles. In urban planning, SHAP interpretation models can help understand how features like population density, economic development level, and transportation network layout impact the direction and form of urban expansion, providing a basis for scientifically reasonable urban planning. In environmental risk assessment, such as flood risk or landslide risk, SHAP can help identify key risk factors in high-risk areas and their contribution levels, thereby providing robust support for the formulation of prevention and response measures.

Geographic research involves a wide variety of data types, including spatial data (e.g., remote sensing images, terrain data) and non-spatial data (e.g., socioeconomic statistics, environmental monitoring data), with substantial differences in feature quantity and dimensions [10]. The SHAP model can effectively handle these different types and scales of geographic data, whether the data is high-dimensional or low-dimensional, continuous or discrete, providing reliable interpretation results and helping researchers further explore the geographic information and patterns hidden in the data.

3.4. Intuitive Visualization

SHAP provides various intuitive visualization tools, such as SHAP value summary plots, SHAP value-feature relationship plots, and dependency plots. In geographic applications, these visualization charts can present complex model interpretation results to researchers in a straightforward, comprehensible manner. For instance, through the SHAP value summary plot, researchers can quickly understand the relative importance of each geographic feature across all sample predictions. In contrast, the SHAP value-feature relationship plot can display how variations in a specific geographic feature affect model predictions, aiding researchers in exploring the relationships between geographic factors and phenomena more deeply [2].

Combined with GIS technology, SHAP's visualization results can be extended into spatial visualization formats. By mapping SHAP interpretation results to the geographic space, researchers can intuitively observe spatial distribution patterns of feature importance and contribution values in different geographic regions [4]. For example, in land use change prediction, spatial visualization can clearly show the significant impact of transportation accessibility features on land use change in urban peripheral areas and the limiting role of terrain features on land use types in mountainous areas, providing robust support for localized geographic decision-making.

3.5. High Flexibility

SHAP models allow researchers to flexibly choose the level and granularity of interpretation according to specific research questions and needs. In geographic research, researchers can choose to interpret the entire model globally or perform local interpretations for specific regions, time periods, or geographic phenomena. For example, when studying land use changes during urban expansion, SHAP interpretation can be conducted for different stages of urban expansion to understand changes in key driving factors at different stages. Alternatively, land use changes in urban central areas and suburban areas can be interpreted separately to analyze the differences and characteristics between regions [11].

SHAP models have high flexibility and can be combined with other geographic analysis methods to form a more formidable comprehensive analysis framework. For instance, during the exploratory analysis phase of geographic spatial data, traditional statistical analysis methods can first be used for preliminary data processing and feature selection, after which SHAP models can interpret prediction

models based on machine learning algorithms. Alternatively, SHAP interpretation results can be combined with spatial autocorrelation analysis, geographically weighted regression, and other methods to further explore the spatial structure and non-stationary characteristics of geographic data, providing new avenues for deepening and refining geographic theories.

4. DISCUSSION

Xie et al. [11] used Sentinel-2 remote sensing images as a data source and adopted a hierarchical strategy based on the characteristics of phenological changes to first interpret primary land cover types and then secondary types. In primary land cover type interpretation, three machine learning models—Random Forest, Support Vector Machine, and Decision Tree—were constructed to categorize the land in the Ordos Irrigation District into forestland, water bodies, cropland, buildings, and others. After selecting the optimal model, manual adjustments were made to further improve the accuracy of cropland classification. For secondary land cover type interpretation, the manually adjusted results were further subdivided into corn, sunflower, wheat, and other crops [7]. The results show that Random Forest performed the best in primary land cover classification, with an overall accuracy of 92.27%, a Kappa coefficient of 0.85, and cropland classification accuracy of 87.60%, which increased to 98.90% after manual adjustment. The overall accuracy of secondary land cover interpretation was 87.93%, with a Kappa coefficient of 0.81. This hierarchical interpretation strategy provides a new approach for high-precision regional survey and statistical work in agriculture, forestry, and other fields.

In land use change prediction, ensemble machine learning models can accurately predict changes in land use types in different regions based on complex factors such as terrain, climate, and soil types. The SHAP model can explain the decision-making processes of these machine learning models and visually present the influence of geographic factors through spatial visualization [12]. For example, utilizing SHAP to interpret the Random Forest model can reveal the contribution size and direction of various geographic features, such as accessibility and surrounding land use types, to the prediction outcomes in land use change prediction, thus providing researchers with valuable insights into the model's decision-making logic, which can serve as a valuable reference for land planning and policy formulation [9].

5. CONCLUSION

This study delves into the integration of artificial intelligence and geographic information systems (GIS), particularly the application of Random Forest and SHAP models in geography. Random Forest demonstrates significant advantages in land use change prediction, environmental risk assessment, and urban planning due to its accuracy, anti-overfitting ability, capability to handle high-dimensional data, and relatively high interpretability. It effectively processes high-dimensional geographic data like hyperspectral remote sensing images and accurately identifies various land feature types. The SHAP model provides powerful explanatory capabilities, quantifying the contribution of each feature to the prediction results and helping researchers understand the mechanisms of key geographic factors comprehensively. Its compatibility allows it to integrate with various machine learning algorithms and GIS software, facilitating spatial visualization to intuitively present the influence of geographic factors. Furthermore, the SHAP model has broad applicability, capable of handling different types and scales of geographic data. Its rich visualization tools can present complex model interpretation results in a straightforward manner. This study concludes that the combination of Random Forest and SHAP models provides theoretical support and practical guidance for the deep integration of GIS and artificial intelligence, promoting the intelligent development of geographic spatial analysis and offering more scientifically grounded decision-making bases in related fields.

REFERENCES

[1] He, Wenmin, Liu, Xuanyuan, Zhou, Qihai, et al. Accuracy optimization of Random Forest interpretation based on terrain data. Journal of Guangxi Normal University (Natural Science Edition), 2025.

[2] Lin, Na, Quan, Hailin, Li, Shuangtao, et al. Remote sensing extraction of abandoned farmland based on SHAP-explainable feature selection. Transactions of the Chinese Society for Agricultural Engineering, 2025.

[3] Guo, Song, Yang, Dongwei, Yin, Xiaoxing, et al. Land use classification of Poyang Lake Nanji Wetland based on feature selection. Yangtze River, 2025.

[4] Wu, Peiyu, Zhang, Xiaoli, Bi, Yuxin, et al. Study on spatiotemporal distribution and dynamic analysis of aboveground biomass of coniferous forests in Yunnan Province. Journal of Southwest Forestry University (Natural Science), 2025.

[5] Zhang, Ping, Tang, Xiaolu, Yang, Zhihan, et al. Analysis of land use change and landscape patterns in Luding County MS6.8 earthquake based on machine learning. Journal of Natural Disasters, 2025, (03).

[6] Hu, Xiangxiang, Shi, Yaya, Hu, Liangbai, et al. Loess landslide susceptibility evaluation based on the coupling model of InSAR and information content-machine learning. Northwestern Geology, 2025, (02): 159-171.

[7] Zhang, Huangfan, Xie, Zhengyi, Wen, Xiaojuan, et al. Driving mechanisms of traffic CO_2 and O_2 based on interpretable machine learning. China Environmental Science, 2025.

[8] Li, Guangyu, Ding, Guosheng, He, Shaoyao, et al. Study on the accessibility and influencing factors of rural educational facilities in Xiangxi Prefecture based on path planning data and RF-SHAP algorithm. Journal of Human Settlements in West China, 2025.

[9] Kong, Kunfeng, Chen, Yiling, Chen, Feng, et al. Multi-model decision-level fusion soil slope stability prediction model and interpretability analysis. Journal of Railway Science and Engineering, 2025.

[10] Gong, Yi, Du, Qiuyue, Zhang, Limei. Estimation of road surface adhesion coefficient based on XG-Boost SHAP dynamic information fusion. Era Car, 2025, (13): 96-98.

[11] Xie, Mei, Han, Congying, Duan, Ximing, et al. Land use classification research based on machine learning. Journal of China Agricultural Science and Technology Review, 2025.

[12] Wang, Huogen, Hu, Mengting, Liu, Xiaochun. Reconstruction and interpretability analysis of China's food security measurement based on machine learning and SHAP algorithm. Journal of China Agricultural University, 2025, (07): 264-274.