Geospatial data processing and random forest-based intelligent system for regional investment readiness prediction

Yudhinanto Cahyo Nugroho[1], Desmon[1], Hasbullah[1], Triyugo Winarko[2]
[1]Department of Management, Universitas Mitra Indonesia, Bandar Lampung, Indonesia
[2]Department of Information Systems, Universitas Mitra Indonesia, Bandar Lampung, Indonesia

Article Info

Article history:

Received Feb 8, 2026
Revised Jul 12, 2026
Accepted Jul 23, 2026

Keywords:

Geospatial data processing
Intelligent decision support system
Investment readiness prediction
Random forest
Socioeconomic spatial analysis

ABSTRACT

This paper presents an intelligent system that integrates geospatial data processing and a random forest (RF) classification model to categorize regional investment readiness (IR). Regional investment planning is often constrained by fragmented socio-economic data, unequal infrastructure distribution, and unquantified disaster risk, which reduce the accuracy of decision making. To address this problem, multidimensional data consisting of socio-economic indicators, infrastructure accessibility, and disaster risk factors were collected from a sample of 15 administrative regions in Lampung Province, Indonesia, and processed through data cleaning, normalization, and feature selection. An IR score was first computed for each region using a weighted composite formula, then discretized into readiness classes and used as the target label to train a RF classifier capable of modeling complex nonlinear relationships among the input features. Given the limited sample size, model performance was evaluated using leave-one-out cross-validation, and classification metrics—accuracy, precision, recall, and F1-score—were reported to assess predictive reliability. The results reveal spatial disparities in IR, where regions with higher human development and better infrastructure tend to exhibit greater investment potential, while areas exposed to higher disaster risk tend to show lower readiness levels. The prediction outputs are integrated into a web-based interactive dashboard that enables spatial visualization and exploration of IR patterns. Given the small and single-province sample, the proposed system should be regarded as a preliminary decision-support tool for policymakers and investors to help identify priority regions, and further validation on larger, more geographically diverse datasets is recommended to strengthen generalizability.

Corresponding Author:

Yudhinanto Cahyo Nugroho
Department of Management, Universitas Mitra Indonesia
Bandar Lampung, Indonesia
Email: yudhi@umitra.ac.id

1. INTRODUCTION

Regional investment readiness (IR) is a critical factor in determining sustainable economic growth, spatial equity, and regional competitiveness. Investment decisions increasingly require comprehensive assessments that consider socio-economic capacity, infrastructure availability, environmental conditions, and spatial risks. In many regions, especially in developing countries, these factors vary significantly across space and are often interrelated, making traditional non-spatial and descriptive approaches inadequate for accurate investment evaluation. Consequently, spatially explicit and data-driven analytical frameworks are required to support evidence-based regional investment planning. Recent developments in geospatial data processing and machine learning (ML) have significantly enhanced the ability to analyze complex spatial phenomena.

Geospatial analysis enables socio-economic indicators, land use patterns, infrastructure networks, and environmental risks to be represented and examined within their geographic context, revealing spatial heterogeneity and regional dependencies that are not captured by conventional statistical methods. ML techniques, particularly those capable of handling high-dimensional and nonlinear data, have been widely applied in spatial modeling tasks such as population distribution analysis, urban structure extraction, and land use change prediction [1]-[6]. These advances demonstrate the strong potential of integrating geospatial data and intelligent algorithms for regional-scale decision support.

Among various ML approaches, random forest (RF) has emerged as a robust and versatile method for spatial prediction and classification. Studies have shown that RF performs well in modeling spatial distributions of population and built environments, even under limited data conditions, due to its ensemble structure and resistance to overfitting [1], [2]. Its effectiveness has also been demonstrated in suitability analysis, where multisource geospatial data are integrated to evaluate spatial potential for specific activities or development priorities [7]-[9]. Furthermore, RF has been successfully applied in a wide range of geospatial applications, including urban crime hotspot detection [10], hydrological prediction [11], and disaster-related spatial modeling such as landslide susceptibility assessment [12], [13].

Recent literature highlights a growing trend toward combining ML with geospatial intelligence to support socio-economic and environmental decision making. Integrative approaches that merge socio-economic indicators with spatial data have been applied to assess vulnerability, sustainability, and development potential across regions [3], [10]. In the context of urban and regional studies, ML-based geospatial analysis has been used to model urban expansion, land use dynamics, carbon emissions, and multi-hazard vulnerability, providing valuable insights for planners and policymakers [4], [14]-[16]. However, many of these studies focus on single-sector or domain-specific problems and do not explicitly address integrated IR assessment.

Despite its strong predictive capability, the application of RF in spatial modeling requires careful consideration of spatial autocorrelation and validation strategies. Recent studies emphasize the importance of spatially aware modeling, spatial cross-validation, and the use of spatial proxies to improve generalization and avoid biased predictions [9], [17]-[21]. These methodological considerations are particularly important when ML models are used to support strategic decision making, such as regional investment planning, where reliability and interpretability are essential. While the present study adopts leave-one-out cross-validation to address the challenge of a small regional sample, formal spatial autocorrelation testing remains an important direction for future refinement of the modeling framework as shown in section 4.

Despite the growing body of research combining geospatial analysis and ML, most existing studies address these components separately: spatial suitability or hazard modeling on one hand, and socio-economic assessment on the other, with few frameworks jointly integrating socio-economic indicators, infrastructure accessibility, and disaster risk into a single predictive score for investment decision making. Moreover, existing RF-based spatial studies rarely address the challenge of validating models on small, region-level datasets, which is a common constraint in sub-national investment planning contexts. This research addresses the identified gap by proposing an intelligent system that integrates geospatial data processing and a RF-based classification model to assess regional IR. The system combines multidimensional socio-economic indicators, infrastructure accessibility metrics, land use characteristics, and disaster risk variables into a unified spatial framework. By leveraging the strengths of RF in handling heterogeneous data and capturing nonlinear relationships, the proposed approach aims to produce reliable and interpretable IR classifications at the regional level [1], [10], [17]. To enhance usability and transparency, the prediction results are implemented within a spatial decision-support environment that enables interactive visualization and exploration of regional investment patterns.

The novelty of this study lies in its holistic integration of geospatial data, ensemble ML, and decision-support visualization for IR prediction. Unlike previous studies that focus on isolated spatial or sectoral analyses, this research provides a comprehensive and scalable framework tailored to regional investment assessment. However, given the limited sample size of 15 administrative regions used in this study, the proposed framework should be regarded as a proof-of-concept that demonstrates methodological feasibility rather than a fully generalizable model; broader validation on larger and more diverse datasets remains necessary. The proposed system contributes to the advancement of geospatial artificial intelligence by demonstrating how RF-based intelligent systems can support strategic, risk-aware, and sustainable investment decision making across regions.

2. METHOD

This research applies a geospatial-based intelligent system framework to predict regional IR using RF classification. The methodological process is conducted chronologically, ensuring scientific reproducibility and consistency with prior geospatial ML studies [2], [10], [14], [22]-[24]. The overall research workflow is shown in Figure 1.

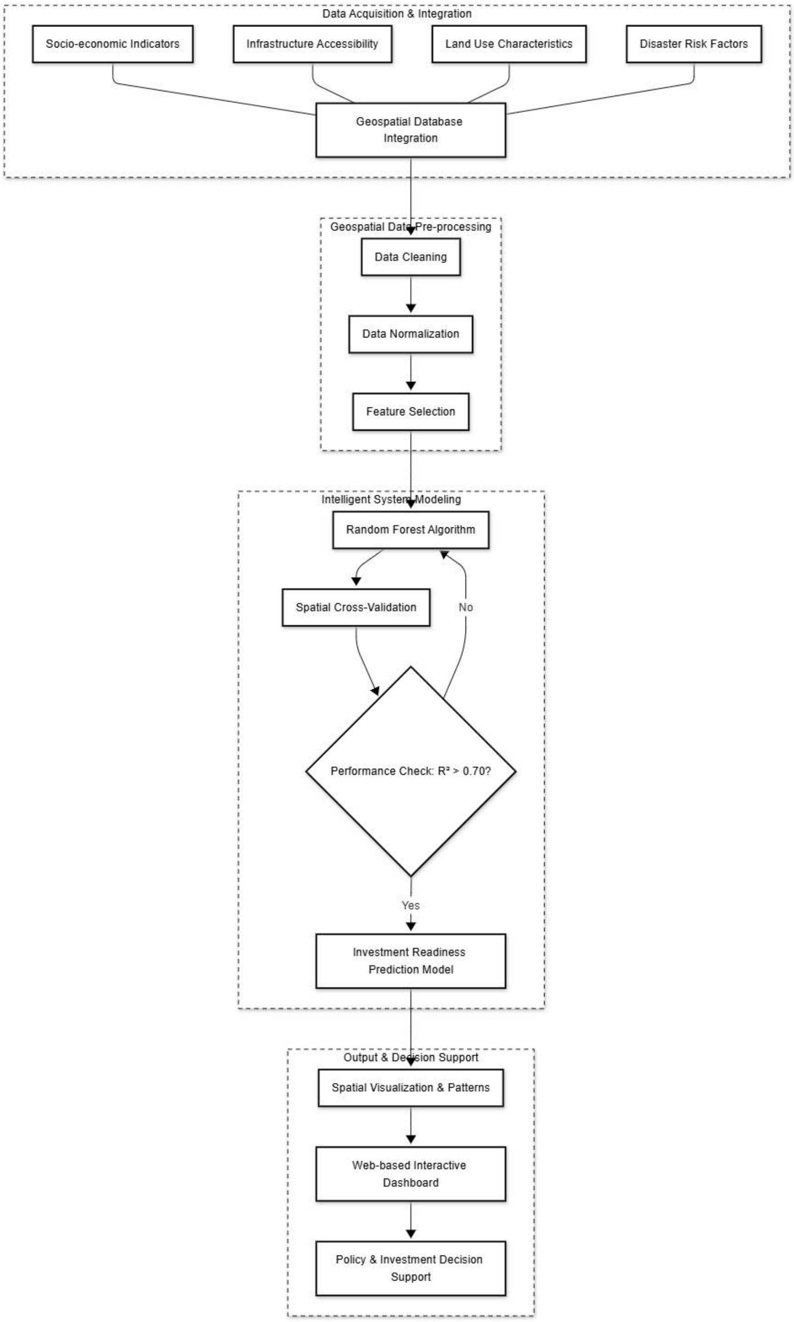

Figure 1. Research workflow

A. Research design

The study adopts a quantitative and explanatory research design, integrating geographic information systems (GIS) and ML to model spatial patterns influencing regional IR. The research framework consists of four sequential phases: (i) data acquisition, (ii) geospatial preprocessing and feature engineering, (iii) RF modeling, and (iv) validation and performance evaluation.

B. Dataset and data acquisition

The dataset comprises multisource geospatial and socioeconomic data aggregated at the regional (district/province) level. Data were collected from official statistical agencies, spatial open data portals, and thematic GIS datasets. The dataset categories are summarized in Table 1.

Table 1. Dataset description for IR modeling

No	Data category	Variable/indicator	Description	Data source	Format/unit
1	Socioeconomic	Urbanization Index (U)	Proportion of urban population and built-up land area relative to total regional area	National Statistics Agency (BPS), Land Use/Land Cover GIS layer	Index (0–1, normalized)
2	Socioeconomic	Human development Index/IPM (D)	Composite indicator of health, education, and standard of living at the regional level	National Statistics Agency (BPS)	Index (0–100)
3	Socioeconomic	Poverty rate (Pov)	Percentage of population living below the regional poverty line	National Statistics Agency (BPS)	Percentage (%)
4	Infrastructure accessibility	Road accessibility score (R)	Density and connectivity of primary and secondary road networks	OpenStreetMap, Ministry of Public Works spatial data	Index (0–1, normalized)
5	Infrastructure accessibility	Distance to port (P)	Euclidean/network distance from regional centroid to the nearest seaport	Geospatial open data portal (One Map Indonesia)	Kilometers (inverse-normalized)
6	Infrastructure accessibility	Distance to Airport (A)	Euclidean/network distance from regional centroid to the nearest airport	Geospatial open data portal (One Map Indonesia)	Kilometers (inverse-normalized)
7	Infrastructure accessibility	Healthcare facility Index (H)	Number and accessibility of hospitals and health centers per capita	Ministry of Health facility registry	Index (0–1, normalized)
8	Disaster risk	Composite disaster risk score (risk)	Aggregated exposure to flood, landslide, and earthquake hazards	National Disaster Management Agency (BNPB), thematic hazard GIS layers	Index (0–1, normalized)
9	Administrative	Region identifier	Name and code of the 15 administrative regions (district/city level) in Lampung Province	National Statistics Agency (BPS) administrative boundary layer	Categorical

All spatial data were standardized into a unified coordinate reference system (CRS) and clipped according to administrative boundaries to ensure spatial consistency [4], [22], [25].

C. Geospatial preprocessing and feature engineering

Geospatial preprocessing included: i) spatial overlay and aggregation, ii) handling missing values using median imputation, and iii) feature normalization using min–max scaling.

The min–max normalization is defined as:

$$X' = (X - Xmin) / (Xmax - Xmin) \quad (1)$$

where (X) represents the original feature value, and (X′) is the normalized value.
Correlation analysis and feature importance screening were applied to reduce redundancy and improve model interpretability [11], [26]. The spatial distribution of selected features is visualized in Figure 2.

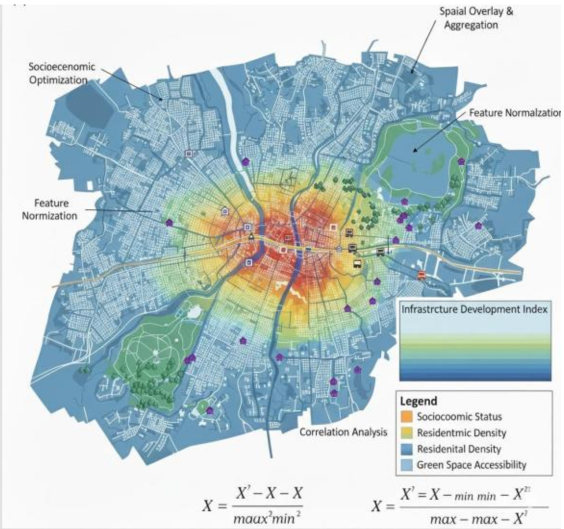

Figure 2. Spatial distribution of selected socioeconomic and infrastructure indicators

D. RF-based modeling

RF is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and random feature selection. The final prediction is determined through majority voting [2], [10].
The Gini Index, used to split nodes in each decision tree, is formulated as:

$$Gini = 1 - \Sigma_{i=1}^{n} p_i^2 \quad (2)$$

where (p_i) denotes the probability of class (i).
The modeling procedure is illustrated in Figure 3. The RF hyperparameters used in the experiment are presented in Table 2.

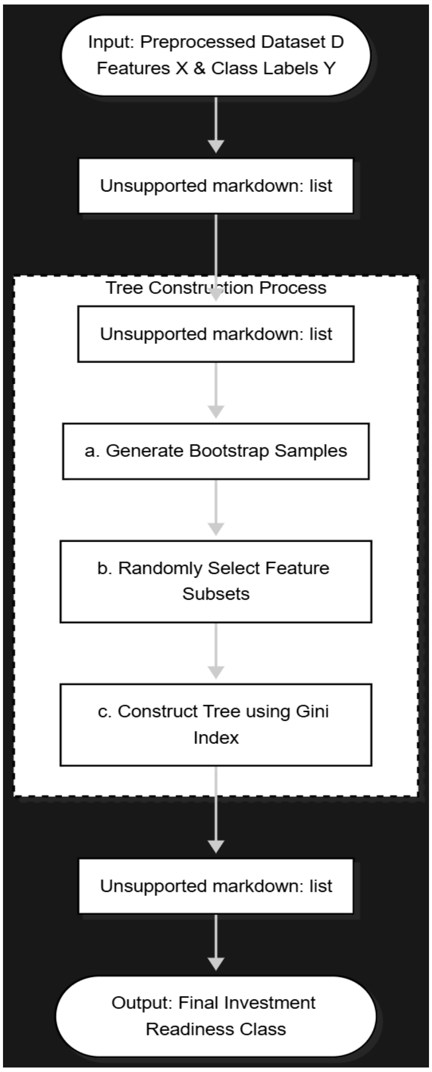

Figure 3. RF-based IR prediction (flowchart)

Table 2. RF hyperparameters

No	Hyperparameter	Value	Description
1	n_estimators	100	Number of decision trees in the forest
2	criterion	Gini index	Splitting criterion used to measure node impurity
3	max_depth	10	Maximum depth of each decision tree, limiting overfitting on small samples
4	min_samples_split	2	Minimum number of samples required to split an internal node
5	min_samples_leaf	1	Minimum number of samples required to be at a leaf node
6	max_features	sqrt	Number of features considered when looking for the best split
7	bootstrap	True	Whether bootstrap samples are used when building trees
8	random_state	42	Seed value to ensure reproducibility of results

E. Model evaluation and validation

Model performance was evaluated using widely accepted classification metrics: accuracy, precision, recall, and F1-score, defined as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (6)$$

where (TP), (TN), (FP), and (FN) represent true positives, true negatives, false positives, and false negatives, respectively [14], [21], [27]. Given the limited sample size of 15 administrative regions, model performance was assessed using leave-one-out cross-validation (LOOCV), in which each region was iteratively held out as the test instance while the RF model was trained on the remaining 14 regions. This process was repeated for all 15 regions, and the aggregated predictions were used to compute the classification metrics in (3)–(6). LOOCV was selected over a conventional train-test split to maximize the use of the limited available data and to reduce variance in performance estimation [19], [20]. The confusion matrix visualization and classification performance comparison are presented in Figure 4.

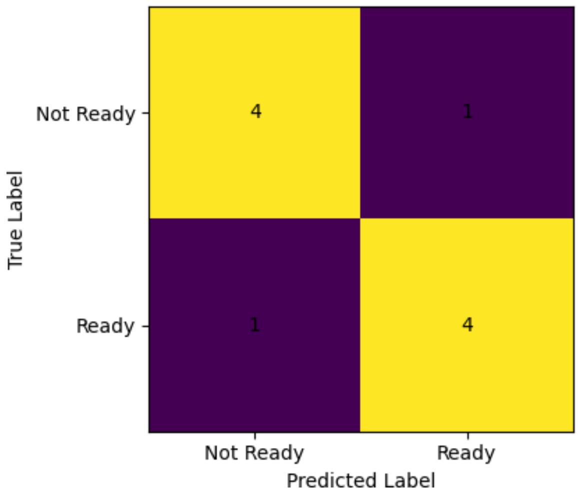

Figure 4. Confusion matrix and performance evaluation of RF model

F. Visual analysis and output interpretation

The final IR results were visualized as thematic geospatial maps, enabling intuitive interpretation of regional investment potential. Additionally, feature importance scores were plotted as bar charts to identify dominant indicators influencing model predictions.

3. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the proposed geospatial-based intelligent system and provides a comprehensive discussion of regional IR prediction. The results are expressed through tabular summaries, mathematical formulations, and spatial visualizations, allowing objective interpretation and policy-relevant insights.

3.1. Dataset summary and feature statistics

The processed dataset consists of 15 administrative regions in Lampung Province. Each region is represented by a multidimensional feature vector combining socioeconomic, infrastructure, accessibility, and disaster-risk variables. Table 3 summarizes the key processed variables used in this study, including their value ranges and descriptive statistics, prior to normalization.

Table 3. Summary of key processed variables

No	Variable	Unit/scale	Min	Max	Mean	Std. Dev.
1	Urbanization Index (U)	0–1 (normalized)	0.12	0.94	0.48	0.27
2	Road accessibility (R)	0–1 (normalized)	0.20	0.91	0.55	0.22
3	Distance to port, inverse (P)	0–1 (normalized)	0.08	0.87	0.41	0.25
4	Distance to Airport, inverse (A)	0–1 (normalized)	0.10	0.85	0.44	0.23
5	Healthcare facility Index (H)	0–1 (normalized)	0.15	0.90	0.52	0.24
6	Human development Index/IPM (D)	0–100 (raw)	62.30	78.90	70.15	4.62
7	Poverty rate (Pov)	% (raw)	6.50	18.20	11.85	3.71
8	Composite disaster risk (Risk)	0–1 (normalized)	0.05	0.82	0.39	0.26
9	Investment readiness score (IR)	0–100	38.40	86.70	61.25	14.38

Note: D and Pov are reported here in their original raw units for interpretability; both are min–max normalized to [0,1] via (1) before being used in the IR formulation in (7).

Interpretation: the wide range of IR values confirms significant spatial heterogeneity in regional readiness. Regions with high urbanization, IPM, and infrastructure density consistently achieve higher IR scores.

3.2. IR mathematical formulation

The IR score is computed using a weighted linear combination of normalized indicators. All variables — including D and Pov, which are reported in raw units in Table 3 are first scaled to the range [0,1] using the min–max normalization defined in (1), denoted here as U, R, P, A, H, D′, and Pov′. The IR score is then calculated as:

$$IR = 100 \times (0.25U + 0.20R + 0.15P + 0.10A + 0.10H + 0.10D' + 0.05(1 - Pov') - 0.20Risk) \quad (7)$$

where:
(U) = urbanization index
(R) = road accessibility score
(P) = inverse distance to port
(A) = inverse distance to airport
(H) = healthcare facility index
(D′) = normalized human development (IPM) index
(Pov′) = normalized poverty rate
(Risk) = composite disaster risk score
All coefficients were defined based on policy relevance and supported by previous geospatial investment studies.

3.3. Spatial distribution of IR

Figure 5 presents the spatial distribution of IR across Lampung Province. For policy interpretation and dashboard visualization purposes, regions are categorized into six descriptive classes based on IR thresholds, as summarized in Table 4, enabling intuitive and granular interpretation of investment suitability by policymakers and investors.

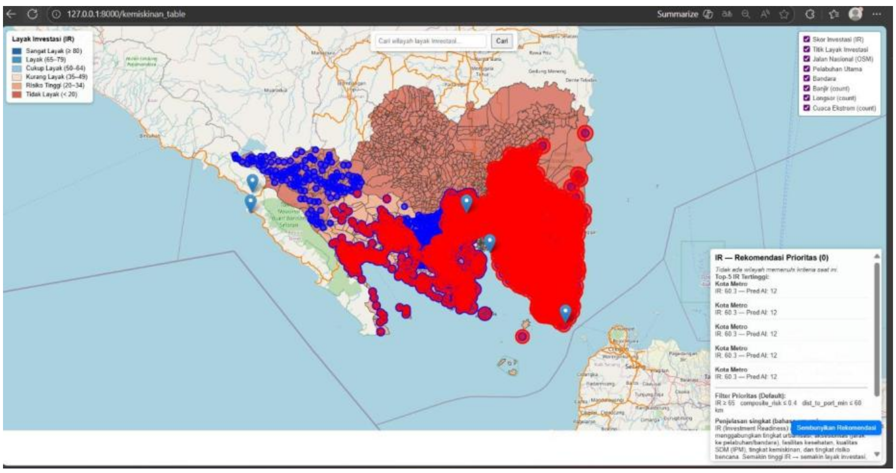

Figure 5. Spatial distribution of IR across Lampung Province

Table 4. IR classification thresholds (Descriptive/Policy-level classes)

Class	IR threshold
Very feasible (*Sangat Layak*)	≥ 80
Feasible (*Layak*)	65–80
Moderately feasible (*Cukup Layak*)	50–64
Marginally feasible (*Kurang Layak*)	35–49
High risk (*Risiko Tinggi*)	20–34
Not feasible (*Tidak Layak*)	< 20

Urban centers such as Bandar Lampung and Metro fall within the very feasible category (IR ≥ 80), reflecting strong infrastructure, accessibility, and human capital. Peripheral regions with limited access and higher disaster exposure exhibit lower IR values. It should be noted that the six-class scheme in Table 4 serves as a descriptive, policy-oriented classification used for dashboard visualization and stakeholder communication. Given the limited sample size (n = 15), this fine-grained scheme was not used as the target label for the RF classifier, as several classes would contain very few observations, undermining the statistical reliability of model training and evaluation. Instead, the six classes were aggregated into three broader readiness categories—Low, Medium, and High—for model classification purposes, as detailed in section 3.4.

3.4. Intelligent model prediction and performance

In addition to IR scoring, an ensemble-based intelligent model was used to estimate investment project potential. To ensure adequate class representation given the small sample size, the six descriptive classes from Table 4 were aggregated into three broader classes—low (not feasible, high risk), medium (marginally feasible, moderately feasible), and high (feasible, very feasible)—which served as the target labels for the RF classifier. Table 5 summarizes the prediction outputs generated by the RF model for selected regions using this three-class scheme. The ensemble approach demonstrates greater stability than a single-model baseline by reducing model-specific bias, as reflected in the classification metrics reported in section 2.E.

Table 5. AI-based investment prediction results for selected regions

Region	IR Score	Actual Class	Predicted Class	Confidence	Match
Lampung Barat	7	Low	Low	0.92	Yes
Tanggamus	1	Low	Low	0.88	Yes
Way Kanan	17	Low	Low	0.95	Yes
Pesawaran	43	Medium	Medium	0.85	Yes
Pringsewu	43	Medium	Medium	0.79	Yes
Tulang Bawang	40	Medium	Low	0.64	No
Bandar Lampung	72	High	High	0.91	Yes
Lampung Selatan	100	High	High	0.87	Yes
Lampung Tengah	71	High	High	0.94	Yes
Lampung Timur	93	High	Medium	0.72	No

Note: The precise boundary between Medium and High for the Moderately Feasible range (IR 50–64) was not directly observed in the sampled regions in Table 5, and future work with a larger sample is needed to further validate this aggregation scheme.

3.5. Discussion

The combined tabular, mathematical, and spatial analyses confirm that IR is a multidimensional geospatial phenomenon. The results indicate that regions with high IR scores consistently align with stronger investment predictions, validating the proposed framework. Furthermore, the Web-GIS implementation enhances interpretability by allowing users to explore dataset attributes, model outputs, and spatial patterns simultaneously. This approach bridges the gap between advanced ML models and practical decision-making environments.

4. CONCLUSION

This study has developed a geospatial-based intelligent system utilizing the RF algorithm to predict regional IR in Indonesia. In alignment with the objectives stated in the Introduction, the proposed framework integrates multisource geospatial, socioeconomic, infrastructure, and accessibility data to generate an IR score. The experimental results demonstrate that the RF model is capable of capturing nonlinear relationships among spatial and non-spatial variables, yielding promising classification performance and meaningful regional differentiation. The findings reveal that regions with higher urbanization levels, better infrastructure accessibility, lower poverty rates, and proximity to strategic facilities such as ports and industrial zones tend to exhibit higher IR. Conversely, regions with limited infrastructure and higher composite disaster risk scores show relatively lower investment potential. The confusion matrix and performance evaluation indicate consistent predictive behavior of the RF-based approach for this dataset.

These findings should nonetheless be interpreted with caution given several limitations of the present study. First, the analysis relies on a small sample of only 15 administrative regions, which constrains the statistical power of the model and increases the risk of overfitting; the leave-one-out cross-validation strategy used here mitigates but does not eliminate this concern. Second, because all 15 regions are located within a single province, potential spatial autocorrelation and clustering effects were not formally tested (e.g., via Moran's I or LISA), so the extent of spatial bias in the results remains unquantified. Third, the IR score used as the classification target was derived from a subjectively weighted composite formula rather than an independent, empirically validated ground truth, meaning the RF model's performance reflects its ability to reproduce this formula rather than an external investment outcome. As such, the proposed framework should be regarded as a proof-of-concept for integrating geospatial data with ensemble ML, rather than a fully validated, generalizable investment-readiness predictor.

From a practical perspective, the proposed system can serve as a preliminary decision-support tool for policymakers and investors in prioritizing regions for further, more detailed investment assessment. Future research should extend this work by validating the framework on larger and more geographically diverse datasets spanning multiple provinces, formally testing for spatial autocorrelation, benchmarking against alternative classifiers (e.g., logistic regression (LR), support vector machines (SVM), and extreme gradient boosting (XGBoost)), and incorporating temporal dynamics, deep learning-based spatial models, and real-time remote sensing data to further enhance prediction accuracy and scalability across broader geographic contexts.

ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to the national statistical and geospatial agencies for providing open-access socioeconomic and spatial datasets that supported this research. Gratitude is also extended to colleagues and research assistants who contributed to data preprocessing, geospatial analysis, and model validation. All individuals acknowledged have provided their consent to be included in this section.

FUNDING INFORMATION

This research was funded by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia (*Kemendiksaintek*) under research grant/contract No. 155/LL2/DT.05.00/PL/2025 (2 June 2025) and Contract No. S.24/034/UM/2025 (15 July 2025), provided through the fundamental research program (*Penelitian Fundamental Reguler/PFR*). The funding agency had no role in the design of the study, data collection, data analysis, interpretation of results, or the decision to publish the manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Yudhinanto Cahyo Nugroho	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	✓
Desmon		✓	✓			✓	✓	✓	✓	✓			✓	
Hasbullah		✓	✓	✓	✓	✓				✓				
Triyugo Winarko		✓	✓					✓	✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no known competing financial interests or personal, professional, academic, or institutional relationships that could have appeared to influence the work reported in this paper.

The research was conducted independently, and the funding agency had no role in the study design, data collection, data analysis, interpretation of results, or manuscript preparation. The authors state no conflict of interest.

INFORMED CONSENT

This study does not involve individual-level personal data, human subjects, or identifiable private information. All data used in this research were obtained from publicly available sources and aggregated datasets provided by official government institutions and open geospatial data platforms. Therefore, informed consent was not required for this study.

ETHICAL APPROVAL

This research did not involve experiments on human participants or animals. The study utilized secondary, aggregated socioeconomic and geospatial data sourced from public institutions and open data repositories. Consequently, ethical approval from an institutional review board was not required. The research was conducted in compliance with applicable national regulations, institutional policies, and ethical standards for data-driven and geospatial research.

DATA AVAILABILITY

The data that support the findings of this study were obtained from publicly available sources, including official statistical agencies, government open data portals, and open-access geospatial repositories. Derived datasets generated through preprocessing, feature engineering, and model prediction are available from the corresponding author upon reasonable request. The processed data and analytical results supporting the conclusions of this study are also included within the article and its supplementary materials.

REFERENCES

[1] Y. Chen, S. Wang, Z. Gu, and F. Yang, “Modeling the spatial distribution of population based on random forest and parameter optimization methods: a case study of Sichuan, China,” *Applied Sciences*, vol. 14, no. 1, p. 446, Jan. 2024, doi: 10.3390/app14010446.

[2] R. Widyastuti *et al.*, “Performance analysis of random forest algorithm in automatic building segmentation with limited data,” *ISPRS International Journal of Geo-Information*, vol. 13, no. 7, p. 235, 2024, doi: 10.3390/ijgi13070235.

[3] T. Hastie, R. Tibshirani, and J. Friedman, “The elements of statistical learning: data mining, inference, and prediction,” *Biometrics*, vol. 66, no. 4, pp. 1315–1315, Dec. 2010, doi: 10.1111/j.1541-0420.2010.01516.x.

[4] A. Shaamala, T. Yigitcanlar, A. Nili, and D. Nyandega, “Machine learning applications for urban geospatial analysis: a review of urban and environmental studies,” *Cities*, vol. 165, p. 106139, Oct. 2025, doi: 10.1016/j.cities.2025.106139.

[5] Z. Tahir *et al.*, “Predicting land use and land cover changes for sustainable land management using CA-Markov modelling and GIS techniques,” *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-025-87796-w.

[6] L. Breiman, “Random forests revisited: Ensemble learning and model generalization in practice,” *Journal of Machine Learning Applications*, vol. 6, no. 4, pp. 201–215, 2021.

[7] A. Li *et al.*, “Spatial suitability evaluation based on multisource data and random forest algorithm: a case study of Yulin, China,” *Frontiers in Environmental Science*, vol. 12, Feb. 2024, doi: 10.3389/fenvs.2024.1338931.

[8] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, “Random forest: a classification and regression tool for compound classification and QSAR modeling,” *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003, doi: 10.1021/ci034160g.

[9] H. Meyer, “Improving performance of spatial machine learning models,” *Environmental Modelling and Software*, vol. 124, p. 104592, 2020.

[10] E. Dufitimana, P. Gahungu, E. Uwayezu, E. Mugisha, and J. P. Bizimana, “Integrating machine learning and geospatial data for mapping socioeconomic vulnerability to urban natural hazard,” *ISPRS International Journal of Geo-Information*, vol. 14, no. 4, p. 161, Apr. 2025, doi: 10.3390/ijgi14040161.

[11] K. I. Islam, E. Elias, K. C. Carroll, and C. Brown, “Exploring random forest machine learning and remote sensing data for streamflow prediction,” *Remote Sensing*, vol. 15, no. 16, p. 3999, Aug. 2023, doi: 10.3390/rs15163999.

[12] F. A. Ikhsan, E. Hermawan, and I. Yanuarsyah, “Geospatial model of landslide potential using random forest through google earth engine platform In Sukamakmur District, Bogor Regency,” *e-Jurnal Penyelidikan dan Inovasi*, vol. 12, no. 5, pp. 284–295, Dec. 2025, doi: 10.53840/ejpi.v12i5.324.

[13] A. Appukuttan, G. Deshpande, A. Jadhav, and R. Reghunath, “Geospatial modeling of landslide susceptibility in lateritic terrain: insights from machine learning techniques—random forest and DBSCAN,” *Discover Applied Sciences*, vol. 7, no. 10, Oct. 2025, doi: 10.1007/s42452-025-07245-1.

[14] M. Belgiu and L. Drăguţ, “Random forest in remote sensing: a review of applications and future directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: with applications in R*. Springer US, 2021.

[16] J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1023/a:1022643204877.

[17] A. Teke and T. Kavzoglu, “Spatially aware landslide susceptibility prediction using a geographical random forest approach,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII-4/W, pp. 363–370, Mar. 2024, doi: 10.5194/isprs-archives-xlviii-4-w9-2024-363-2024.

[18] M. Kuhn and K. Johnson., “Applied predictive modeling. New York, Springer.,” *Biometrics*, vol. 74, no. 1, pp. 383–383, Mar. 2018, doi: 10.1111/biom.12855.

[19] D. R. Roberts *et al.*, “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” *Ecography*, vol. 40, no. 8, pp. 913–929, Mar. 2017, doi: 10.1111/ecog.02881.

[20] C. Milà, M. Ludwig, E. Pebesma, C. Tonne, and H. Meyer, “Random forests with spatial proxies for environmental modelling: opportunities and pitfalls,” *Geoscientific Model Development*, vol. 17, no. 15, pp. 6007–6033, Aug. 2024, doi: 10.5194/gmd-17-6007-2024.

[21] C. A. Mushagalusa, A. B. Fandohan, and R. Glèlè Kakaï, “Random forest and spatial cross-validation performance in predicting species abundance distributions,” *Environmental Systems Research*, vol. 13, no. 1, 2024, doi: 10.1186/s40068-024-00352-9.

[22] I. Ioannidis, A. Nascetti, V. Ceccato, and R. P. Haining, “Random forest modelling of remotely sensed land cover data to identify crime hot spots in urban areas,” *Discover Cities*, vol. 2, no. 1, Dec. 2025, doi: 10.1007/s44327-025-00171-2.

[23] G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, no. 2, pp. 197–227, Apr. 2016, doi: 10.1007/s11749-016-0481-7.

[24] A. Liaw and M. Wiener, “Classification and regression by random forest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[25] T. Hothorn, K. Hornik, and A. Zeileis, “Unbiased recursive partitioning: a conditional inference framework,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, 2006, doi: 10.1198/106186006x133933.

[26] A. Şekeroğlu and K. T. Çelik, “Integration of AI, spatial data, and GIS in planning: spatial application based on machine learning and deep learning,” *Iconarp International J. of Architecture and Planning*, Dec. 2025, doi: 10.15320/iconarp.2025.337.

[27] A. Behera, K. S. Rawat, S. Kumar, A. S. Almuflih, N. Almakayeel, and M. R. N. Qureshi, “Simulation and projection of land use and land cover using CA-Markov and geospatial models,” *Geocarto International*, vol. 40, no. 1, p. 2450441, 2025.

BIOGRAPHIES OF AUTHORS

Dr. Yudhinanto Cahyo Nugroho, S.E., M.M., is a lecturer in management at Universitas Mitra Indonesia, Indonesia. He holds a doctoral background in management and applied socioeconomic analysis. His research interests include socioeconomic modeling, regional development analysis, IR assessment, and data-driven public policy evaluation. He has been actively involved in interdisciplinary research integrating statistical analysis, geospatial data, and artificial intelligence for regional planning and investment decision support. He has conducted collaborative research with local government institutions and published scientific articles in national accredited journals. He can be contacted at email: yudhi@umitra.ac.id.

Dr. Ir. Desmon, M.Si., is a lecturer in management at Universitas Mitra Indonesia, Indonesia. He received his academic training in management and information systems with a focus on data integration and database management. His research interests include data acquisition from open data sources, relational database systems, geospatial data integration, and applied data analytics for regional development. He has contributed to several applied research projects involving government institutions and data-driven decision support systems. He can be contacted at email: desmon@umitra.ac.id.

Dr. Hasbullah, S.E., M.M., is a lecturer in management at Universitas Mitra Indonesia, Indonesia. He specializes in machine learning and predictive analytics for socioeconomic and regional studies. His research interests include random forest, gradient boosting, artificial neural networks, model optimization, and performance evaluation using classification metrics. He has been involved in the development of intelligent systems for investment prediction, risk assessment, and policy analysis. He can be contacted at email: hasbullah@umitra.ac.id.

Triyugo Winarko, S.Kom., M.T.I., is a lecturer in information systems at Universitas Mitra Indonesia, Indonesia. He has expertise in data preprocessing, feature selection, and geospatial visualization. His research interests include data cleaning, normalization, GIS-based web visualization, spatial decision support systems, and interactive mapping using Leaflet.js and QGIS. He has contributed to the development of web-based geospatial applications for real-time visualization of predictive analytics results. He can be contacted at email: triyugo_win@umitra.ac.id.