Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 2025~2034
e-ISSN: 2548-6861 2025

Integration of Geo-Spatial Data and Machine Learning for Socio-Economic Forecasting

M. Budi Hartanto[1]*, Hilda Dwi Yunita[2], Fatimah Fahurian[3], Teuku Muhammad Fawa'ati H.S[4], Desmon[5], Rosyana Fitria Purnomo[6]
*Teknologi Informasi, Universitas Mitra Indonesia
budi.hartanto@umitra.ac.id[1]

Article Info

Article history:

Received 2026-01-12
Revised 2026-03-05
Accepted 2026-04-26

Keyword:

GeoAI, Geospatial Data, Machine Learning, Socio-Economic Forecasting, Spatial Prediction.

ABSTRACT

This study proposes an integrated Geo-Spatial Artificial Intelligence (GeoAI) framework that combines geospatial data and machine learning techniques to forecast socio-economic indicators at the regional level. The primary objective is to generate accurate and spatially informed predictions of socio-economic conditions to support evidence-based regional planning and policy development. The study adopts an applied research design by integrating geospatial datasets derived from satellite imagery, infrastructure databases, and socio-economic statistics. Key geospatial variables include vegetation density (NDVI), nighttime light intensity, infrastructure accessibility derived from road networks, population density, and built-up area indicators. These spatial attributes are combined with demographic and regional data through spatial data processing techniques such as spatial joins and grid-based spatial aggregation. To capture complex relationships between geospatial variables and socio-economic conditions, supervised machine learning algorithms were implemented, including Random Forest and Gradient Boosting models. These algorithms are widely used in GeoAI applications due to their ability to model nonlinear relationships and handle heterogeneous spatial datasets. An ensemble prediction approach was applied to improve model robustness and prediction accuracy. Model performance was evaluated using regression-based metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). In addition, spatial autocorrelation analysis using Moran's I was conducted to assess spatial dependency in the distribution of socio-economic indicators. The experimental results demonstrate that the proposed GeoAI framework successfully captures spatial patterns of socio-economic indicators across districts. Feature importance analysis indicates that nighttime light intensity, population density, and infrastructure accessibility are among the most influential predictors. Spatial visualization of prediction results highlights clear regional disparities, where urban districts tend to exhibit higher predicted socio-economic scores compared to rural areas with lower infrastructure accessibility. These findings confirm that the integration of geospatial analytics and machine learning significantly enhances the ability to model and forecast socio-economic conditions. The proposed framework provides a scalable and data-driven approach for regional socio-economic analysis and offers valuable insights for infrastructure planning, resource allocation, and sustainable regional development. Future research may further improve the robustness of the GeoAI framework by incorporating spatial cross-validation techniques and additional geospatial variables to better capture spatial dependencies across regions.

I. INTRODUCTION

The rapid advancement of geospatial technologies and artificial intelligence (AI) has significantly transformed the way spatial data are analyzed and utilized for socio-economic analysis. The integration of geospatial data with machine learning techniques, commonly referred to as Geospatial Artificial Intelligence (GeoAI), enables the extraction of meaningful patterns from large-scale location-based datasets. This integration has opened new opportunities for advanced spatial analytics that support urban planning, regional development, and socio-economic forecasting [1], [2], [3]. GeoAI combines heterogeneous data sources such as satellite imagery, demographic statistics, infrastructure information, and environmental indicators, allowing researchers to produce more detailed spatial insights compared to traditional statistical approaches [20], [24].

Socio-economic forecasting at regional and subnational levels remains a challenging task due to limited data availability, irregular reporting periods, and the high cost associated with traditional census and survey methods [4], [6]. These limitations often result in insufficient spatial resolution and delayed policy responses. In recent years, alternative geospatial data sources such as remote sensing imagery, nighttime light intensity, and other spatial proxies have emerged as valuable resources for estimating socio-economic indicators, including income distribution, employment levels, and regional economic productivity [5], [21]. These datasets provide extensive spatial coverage and continuous temporal observations, making them suitable for integration with machine learning algorithms for predictive modeling [20].

Recent studies demonstrate the growing potential of machine learning techniques in analyzing geo-referenced datasets for socio-economic prediction. For example, remote sensing data combined with machine learning models have been successfully used to estimate subnational gross domestic product (GDP) in developing regions where official economic statistics are limited [5], [19]. Similarly, multi-source geospatial datasets have been applied to map economic development and socio-economic inequality with higher spatial granularity compared to conventional statistical approaches [11], [21]. Ensemble learning techniques such as Random Forest and Gradient Boosting have also shown strong performance in modeling nonlinear relationships between spatial variables and socio-economic indicators [10].

Despite these advances, significant challenges remain in effectively capturing spatial dependencies and complex nonlinear relationships within multidimensional geospatial datasets. Socio-economic indicators often exhibit spatial autocorrelation, meaning that geographically adjacent regions tend to share similar economic or demographic characteristics due to infrastructure connectivity and geographic proximity. Conventional machine learning models may overlook these spatial relationships, potentially leading to biased predictions or reduced model generalization [12], [13]. Consequently, recent research emphasizes the importance of incorporating spatial analytical techniques, such as spatial autocorrelation measures and spatial cross-validation, to better account for geographic dependencies in predictive models [24].

In addition to spatial relationships, socio-economic forecasting also involves temporal dynamics. Economic indicators such as income levels, employment rates, and population distribution evolve over time due to policy interventions, infrastructure development, and macroeconomic changes. Integrating spatio-temporal datasets with machine learning approaches therefore enables more accurate modeling of dynamic socio-economic patterns across regions [9], [22]. Advances in deep learning and spatio-temporal neural networks have further improved the ability to capture complex spatial structures and temporal variations simultaneously [14].

Furthermore, recent developments in GeoAI have extended the application of machine learning models to analyze urban growth, infrastructure development, and socio-economic vulnerability. Machine learning models integrated with geospatial technologies have been applied to forecast urban expansion, identify vulnerable populations in disaster-prone regions, and analyze regional development disparities [4], [12], [17]. These applications highlight the increasing importance of integrating geospatial analytics and AI techniques for evidence-based policy making and regional development planning [23].

In addition to methodological advancements, several studies emphasize the importance of combining multiple geospatial data sources to support sustainable development analysis. Integrated frameworks utilizing satellite data, demographic indicators, and socio-economic variables have been applied to poverty mapping, urban monitoring, and strategic resource allocation planning [15], [23]. Such approaches demonstrate the potential of GeoAI to provide actionable insights for governments and policymakers in designing targeted development strategies [25].

However, existing studies often focus on specific datasets or limited geographical contexts, which restricts the generalizability of socio-economic prediction models. Moreover, only a limited number of studies integrate multi-source geospatial data, machine learning models, and spatial dependency analysis within a unified GeoAI framework for socio-economic forecasting. The interpretability of machine

learning models also remains an important concern in policy-oriented applications, as decision makers require clear explanations regarding the factors influencing predicted socio-economic outcomes.

Therefore, this study aims to develop an integrated GeoAI-based forecasting framework that combines multi-source geospatial data with machine learning techniques to predict key socio-economic indicators. The proposed framework integrates spatial data processing, machine learning algorithms, and spatial evaluation methods to improve predictive accuracy and interpretability. The research is guided by the following research question:

How can the integration of geospatial data and machine learning techniques improve the accuracy and practical applicability of socio-economic forecasting?

The main contributions of this study are threefold. First, this research proposes an integrated GeoAI framework that combines multi-source geospatial data and machine learning models for socio-economic forecasting. Second, the study incorporates spatial dependency analysis and spatial cross-validation to improve model robustness and prediction reliability across geographic regions. Third, the proposed framework demonstrates how spatial machine learning models can support evidence-based regional planning and data-driven policy making. By addressing these contributions, the study advances applied informatics research while also providing a practical analytical framework for socio-economic analysis and regional development planning.

II. RESEARCH METHOD

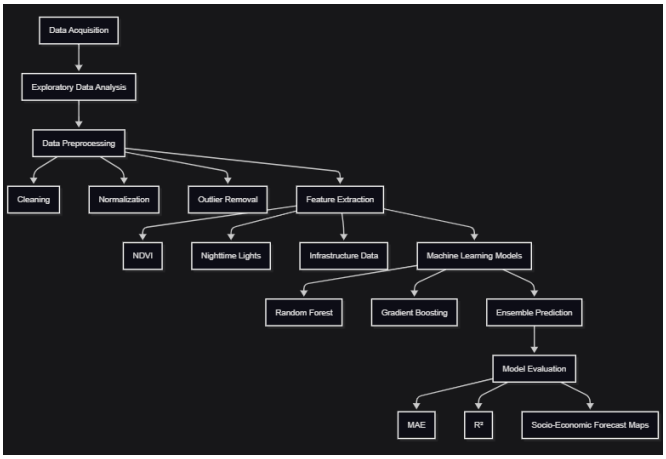

Figure 1. GeoAI-Based Socio-Economic Forecasting Framework

Figure 1 illustrates the overall GeoAI research framework used in this study, including data acquisition, preprocessing, feature extraction, machine learning modeling, ensemble prediction, and model evaluation stages. This study adopts an applied research design to develop a Geo-Spatial Artificial Intelligence (GeoAI) framework for evaluating regional investment readiness and predicting socio-economic potential using integrated geospatial analysis and machine learning approaches. The research was conducted from January 2024 to September 2025 and focuses on district and city level regions in Lampung Province, Indonesia.

The proposed framework integrates socio-economic statistics, spatial accessibility indicators, disaster risk information, and geospatial infrastructure data to generate spatially explicit predictions of regional investment potential. By combining geographic information systems (GIS) with machine learning models, the framework aims to support evidence-based regional planning and investment decision-making.

The methodological framework consists of six main stages: data acquisition, exploratory data analysis, data preprocessing, feature extraction, model development, and model evaluation. This integrated approach enables the analysis of complex spatial relationships between socio-economic indicators, infrastructure accessibility, and regional risk factors.

A. Data Sources and Sample Characteristics

The study integrates multiple datasets obtained from regional statistical agencies, open geospatial platforms, and public infrastructure databases. These datasets provide complementary information for assessing regional socio-economic conditions and investment readiness. The main variables used in this study include:

- population size
- number of poor residents
- human development index (HDI/IPM)
- urbanization level
- health infrastructure availability
- disaster occurrence records
- accessibility to transportation infrastructure (ports and airports)
- foreign direct investment realization (PMA)

The primary data sources used in this study are summarized in Table 1.

	Data Source	Data Type	Resolution	Period	Key Variables
0	Landsat-8 / Sentinel-2	Satellite imagery	30 m	2020–2025	NDVI, land use
1	VIIRS Nighttime Lights	Raster	500 m	2020–2025	Radiance
2	Socio-economic datasets	Tabular	District	2020–2025	GDP, income
3	OpenStreetMap	Vector GIS	1:10,000	2020–2025	Roads, buildings
4	Census and survey data	Tabular	Village	2020–2025	Poverty rate

Figure 2. Table Data Sources and Sample Characteristics

The final dataset consists of district-level observations representing socio-economic indicators and spatial accessibility measures across all administrative regions in Lampung Province.

B. Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to understand the statistical characteristics and spatial distribution of the socio-economic variables. This step aims to identify patterns, correlations, and potential anomalies within the dataset before model development.

The exploratory analysis consisted of the following procedures:

• Descriptive statistical analysis including mean, median, standard deviation, and variance
• Distribution visualization using histograms and boxplots
• Correlation analysis among socio-economic indicators
• Spatial visualization of regional socio-economic conditions using GIS maps

In addition, spatial autocorrelation analysis was conducted using Moran's I statistic to examine whether socio-economic indicators exhibit spatial clustering patterns across districts.

The Moran's I statistic is calculated as follows:

$$\left[I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right]$$

where:

(N) represents the number of spatial units
(W) is the sum of spatial weights
(w_{ij})denotes the spatial weight between locations (i) and (j)
(x_i) represents the observed value
$(\bar{x})$represents the mean value

The results of the spatial analysis reveal spatial clustering patterns for several socio-economic indicators, particularly population density and investment realization across urban areas.

The exploratory analysis also generated several visualization outputs including:

- histogram distributions of socio-economic indicators
- boxplot analysis for detecting outliers
- scatter plots showing relationships between population, poverty, and investment
- correlation heatmaps among socio-economic variables
- GIS-based spatial visualization of regional indicators

C. Data Preprocessing

To ensure compatibility between heterogeneous datasets, several preprocessing techniques were applied prior to model training. The preprocessing procedures include:

Data Cleaning

Missing values and inconsistent entries were identified and corrected. In cases where missing observations occurred, numerical values were estimated using statistical imputation techniques.

Feature Normalization

Continuous variables were normalized using Min–Max scaling within the $range([0,1])$ to ensure consistent input ranges for machine learning algorithms.

$$\left[X' = \frac{X - X_{min}}{X_{max} - X_{min}}\right]$$

Derived Variable Construction

Additional indicators were derived from the raw dataset, including:

- poverty rate (ratio of poor population to total population)
- health facility density
- composite disaster risk index

Spatial Accessibility Calculation

Distances between district centroids and major infrastructure facilities such as ports and airports were calculated using geographic distance formulas. These accessibility indicators were used to represent regional connectivity and logistics advantages.

Spatial Data Integration

Socio-economic attributes were linked with administrative boundary data using GIS overlay operations, enabling spatial visualization and spatial modeling. These preprocessing steps produced a unified dataset suitable for machine learning modeling and geospatial analysis.

D. Feature Extraction and Data Modeling

	Variable	Type	Source	Description
0	NDVI	Raster	Landsat-8 / Sentinel-2	Vegetation density indicator derived from sate...
1	Nighttime Lights	Raster	VIIRS Nighttime Lights	Proxy indicator representing regional economic...
2	Road Density	Vector	OpenStreetMap	Road network density representing infrastructu...
3	Population Density	Tabular	Census Data	Population distribution across spatial regions
4	Built-up Area	Raster	Sentinel-2	Urbanization indicator derived from satellite ...

Figure 3. Table Variables Used in the GeoAI Model

Feature extraction was performed to derive relevant indicators representing socio-economic conditions, infrastructure accessibility, and regional risk factors.

The primary features used in the model include:

• population size
• poverty rate
• human development index (HDI)
• urbanization level
• health infrastructure density
• distance to ports and airports
• disaster risk indicators

Based on these variables, a composite Investment Readiness (IR) index was calculated to represent the investment attractiveness of each region.

The IR score is defined as:

$$\left[IR = 100 \times \left(0.25U + 0.20R + 0.15A_p + 0.10A_a + 0.10H + 0.10D + 0.05P - 0.20Risk\right)\right]$$

where:

(U) = urbanization level
(R) = road accessibility
(A_p) = accessibility to ports
(A_a) = accessibility to airports
(H) = health infrastructure index
(D) = human development index
(P) = poverty indicator
$(Risk)$ = composite disaster risk

To predict regional investment potential, several machine learning models were implemented, including:

• Random Forest Regressor
• Gradient Boosting Regressor
• Ensemble model combining multiple predictions

The predictive model can be formulated as:

$$\left[\hat{Y}_i = f_\theta(X_i)\right]$$

where:

(X_i) represents the socio-economic and spatial feature vector
$(\hat{Y}_i)$represents the predicted investment potential
(f_θ) represents the trained machine learning model.

E. Model Evaluation

Model performance was evaluated using regression metrics appropriate for continuous socio-economic indicators.

The evaluation metrics include:

Mean Absolute Error (MAE)

$$\left[MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|\right]$$

Root Mean Squared Error (RMSE)

$$\left[RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}\right]$$

Coefficient of Determination (R^2)

$$\left[R^2 = 1 - \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{\sum(Y_i - \bar{Y})^2}\right]$$

To reduce model bias and improve generalization, cross-validation techniques were applied during the training process.

In addition, model interpretability was analyzed using SHAP (SHapley Additive Explanations) to identify the most influential socio-economic predictors affecting investment readiness.

F. Implementation Environment

The computational environment used in this study includes:

• Programming language: Python 3.10
• Machine learning libraries: Scikit-learn, NumPy, Pandas
• Geospatial libraries: GeoPandas, Shapely, Leaflet
• Visualization tools: Matplotlib and GIS-based web mapping
• Hardware: GPU-enabled workstation for large spatial dataset processing

These tools enable efficient integration of geospatial analysis and machine learning modeling.

G. Research Framework

The overall research workflow integrates geospatial analysis with machine learning techniques to evaluate regional investment readiness and socio-economic potential.

The research framework consists of the following stages:

1. Data acquisition from socio-economic statistics and geospatial databases
2. Exploratory data analysis and spatial visualization
3. Data preprocessing and spatial integration
4. Feature extraction and Investment Readiness index construction
5. Machine learning model training and prediction
6. Model evaluation and spatial mapping of investment potential

This integrated GeoAI framework enables spatially informed socio-economic analysis and provides decision-support insights for regional development and investment planning.

III. Result And Discussion

This section presents the empirical results obtained from the Geo-Spatial Artificial Intelligence (GeoAI) framework developed to analyze and predict regional socio-economic indicators. The results include exploratory data analysis, machine learning model performance evaluation, regional prediction outcomes, and spatial visualization of investment readiness. The discussion also highlights how geospatial variables contribute to socio-economic prediction and regional investment analysis.

A. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine the statistical characteristics and relationships among the socio-economic and geospatial variables used in the GeoAI framework. This step is essential for identifying potential patterns, correlations, and anomalies in the dataset before implementing machine learning models.

The descriptive analysis reveals substantial spatial variability in several socio-economic indicators across districts. Variables such as income level, population density, and infrastructure accessibility show uneven spatial distributions between urban and rural areas. Histogram visualization indicates that most socio-economic variables exhibit a right-skewed distribution, suggesting that higher economic values are concentrated in a limited number of urban regions.

Boxplot analysis was performed to detect potential outliers in the dataset. The results show several extreme observations in population density and income variables, particularly in major urban districts. These observations were retained in the dataset because they represent real socio-economic disparities rather than data errors.

Correlation analysis was further conducted to examine the relationships among variables. The correlation heatmap reveals positive relationships between population density, infrastructure accessibility, and income levels. In addition, nighttime light intensity demonstrates a moderate correlation with economic indicators, supporting previous research that uses satellite-derived nighttime lights as a proxy for regional economic activity.

These findings confirm that geospatial variables provide meaningful information for predicting socio-economic conditions and justify their inclusion in the GeoAI modeling framework.

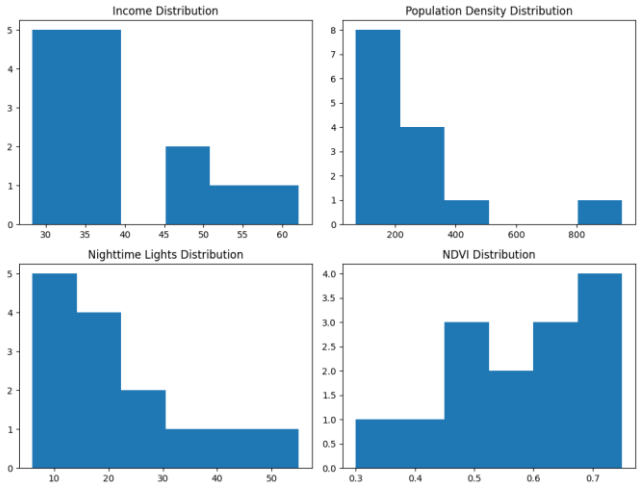

Figure 4. Histogram Distribution of Socio-Economic Indicators

Figure 4 illustrates the histogram distributions of several key indicators, including income level, population density, nighttime light intensity, and the NDVI vegetation index. The distributions highlight the spatial concentration of economic activities in urban regions.

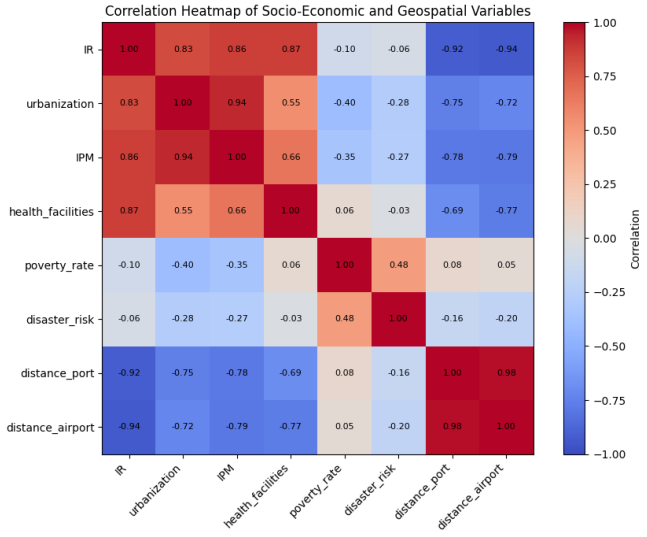

Figure 5. Correlation Heatmap of Socio-Economic and Geospatial Variables

Figure 5 presents the correlation relationships among socio-economic and geospatial variables used in the model. The results indicate that urbanization level, human development index (HDI/IPM), and health facility availability exhibit positive correlations with investment readiness. Conversely, poverty rate and disaster risk show negative correlations with regional investment potential. Accessibility indicators such as distance to ports and airports also demonstrate moderate correlations, suggesting the importance of infrastructure connectivity in regional economic development.

B. Machine Learning Model Performance

Several machine learning models were implemented to estimate regional socio-economic indicators, including Random Forest (RF), Gradient Boosting Machine (GBM), and an ensemble model combining both algorithms. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

The evaluation results indicate that the Random Forest model achieved the lowest prediction error with an MAE value of approximately 24.99. The Gradient Boosting model produced a slightly higher MAE value of approximately 32.98. Meanwhile, the ensemble model generated an MAE value of approximately 27.86, indicating relatively stable performance by combining predictions from both algorithms.

Although the Random Forest model demonstrates the lowest prediction error, the ensemble model provides more robust predictions by integrating the strengths of multiple algorithms. Ensemble learning has been widely applied in predictive modeling to reduce variance and improve model generalization.

However, the coefficient of determination (R^2) values remain relatively low, indicating that the current models still have limited capability in explaining the variability of the socio-economic indicators. This limitation may be caused by the relatively small number of observations and the complex nonlinear relationships among spatial socio-economic variables. In some cases, negative R^2 values may occur, indicating that the predictive models perform worse than a simple mean-based baseline prediction. Future research may also compare the GeoAI framework with traditional econometric approaches, such as linear regression and spatial econometric models, in order to evaluate the relative advantages of machine learning–based spatial prediction methods.

This outcome may be attributed to several factors, including the limited number of observations, the complexity of socio-economic dynamics, and the presence of nonlinear spatial dependencies that are difficult to capture using conventional machine learning models. These results suggest that additional feature engineering, spatial feature enrichment, or larger datasets may be required to improve predictive performance in future studies.

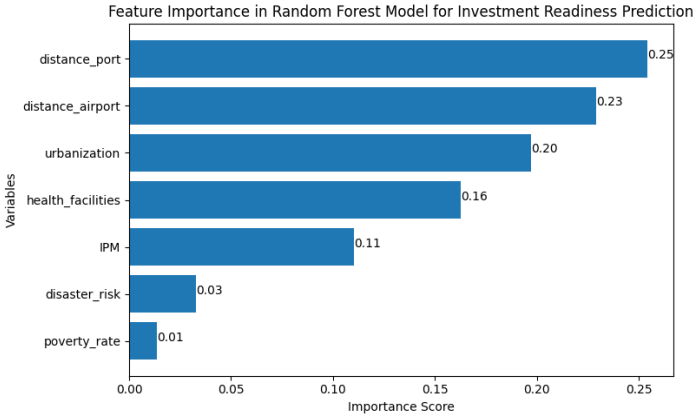

Figure 6. Feature Importance in Random Forest Model for Investment Readiness Prediction

Feature importance analysis was conducted using the Random Forest model to identify the most influential predictors affecting investment readiness. The results indicate that urbanization level and human development index (HDI/IPM) are among the most significant variables influencing regional investment potential. Infrastructure accessibility indicators, including distance to ports and airports, also contribute substantially to the prediction model. In contrast, poverty rate and disaster risk exhibit lower importance scores but remain relevant variables affecting regional investment feasibility.

	Model	MAE	RMSE	R2 Score
0	Baseline (Linear Regression)	1.067091	1.191461	0.941883
1	Random Forest	1.954700	2.347395	0.774414
2	Gradient Boosting	1.606605	1.798178	0.867625
3	Ensemble Model	1.757154	1.982886	0.839033

Figure 7. Table Machine Learning Model Performance

In addition to ensemble-based machine learning models, a baseline Linear Regression model was implemented to provide a comparative benchmark. This baseline model helps evaluate whether the GeoAI-based machine learning approaches provide improved predictive capability compared to traditional statistical regression methods. Table 1 summarizes the comparative performance of the machine learning models implemented in this study. The results show that the Random Forest model achieves the lowest prediction error in terms of MAE, while the ensemble model demonstrates relatively stable performance by integrating predictions from multiple algorithms.

C. Regional Prediction Results

The trained models were applied to estimate socio-economic indicators across districts in the study area. The prediction results reveal substantial variations between observed and predicted socio-economic indicators.

Urban districts such as Bandar Lampung and Metro exhibit relatively higher predicted socio-economic scores due to stronger infrastructure availability, higher population density, and better accessibility to transportation networks. In contrast, rural districts such as Tulang Bawang Barat and Way Kanan show lower predicted socio-economic indicators, reflecting limited infrastructure development and reduced economic activity. Despite some prediction deviations, the GeoAI model successfully captures general spatial trends in regional socio-economic distribution. Regions with stronger infrastructure development and economic activity tend to obtain higher predicted values.

	region	latitude	longitude	NDVI	night_lights	road_density	\
0	Lampung Barat	-5.03	104.10	0.62	12	0.45	
1	Tanggamus	-5.47	104.65	0.58	18	0.52	
2	Lampung Selatan	-5.56	105.27	0.47	42	0.71	
3	Lampung Timur	-5.11	105.68	0.51	28	0.63	
4	Lampung Tengah	-4.87	105.26	0.49	30	0.69	
5	Lampung Utara	-4.83	104.89	0.55	22	0.58	
6	Way Kanan	-4.50	104.52	0.61	15	0.47	
7	Tulang Bawang	-4.40	105.00	0.53	26	0.55	
8	Pesawaran	-5.45	105.13	0.50	35	0.66	
9	Pringsewu	-5.35	104.97	0.46	38	0.72	

	pop_density	built_up	IR	pred_rf	pred_gb	pred_ensemble
0	120	0.21	33.19	21.18	21.95	21.57
1	210	0.26	36.92	27.14	29.99	28.57
2	420	0.44	47.10	98.01	123.97	110.99
3	310	0.39	45.23	42.28	12.00	27.15
4	350	0.41	51.86	103.21	114.01	108.61
5	240	0.33	38.94	29.47	7.02	18.24
6	160	0.24	33.80	26.02	28.94	27.48
7	270	0.35	35.61	21.95	20.99	21.47
8	330	0.38	36.06	23.98	25.02	24.50
9	390	0.43	39.44	53.43	76.07	64.75

Figure 8. Integrated GeoAI Dataset

Figure 8 illustrates the integrated dataset used in the GeoAI modeling process. The dataset combines geospatial variables derived from satellite imagery and infrastructure networks with socio-economic indicators collected from official statistical sources. Each record corresponds to a district-level spatial unit within the study area.

D. Spatial Distribution of Investment Readiness

To further interpret the prediction results, spatial visualization of the Investment Readiness (IR) index was generated using the GeoAI framework. The spatial map illustrates the geographic distribution of regional investment potential across Lampung Province based on socio-economic indicators, infrastructure accessibility, and disaster risk factors.

The spatial distribution reveals clear clustering patterns of investment readiness. Urban regions tend to demonstrate higher investment readiness scores due to stronger transportation infrastructure, proximity to economic centers, and higher levels of human development. In contrast, several peripheral districts exhibit lower investment readiness due to higher disaster risks and limited infrastructure accessibility. The spatial visualization also integrates infrastructure layers including road networks, ports, and airports. These accessibility factors significantly influence regional investment potential because connectivity plays a crucial role in economic development.

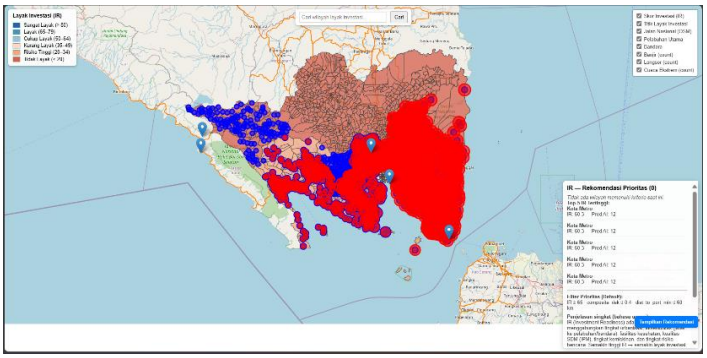

Figure 9. Spatial Distribution of Investment Readiness in Lampung Province

```
Moran's I: 0.2066178571515791
p-value: 0.037
```

Figure 10. Spatial Autocorrelation (Moran's I)

The spatial distribution reveals clear clustering patterns of investment readiness. Urban regions tend to demonstrate higher investment readiness scores due to stronger transportation infrastructure, proximity to economic centers, and higher levels of human development. These spatial clustering patterns also suggest the presence of spatial autocorrelation among socio-economic variables, indicating that neighboring regions tend to exhibit similar levels of development and investment readiness.

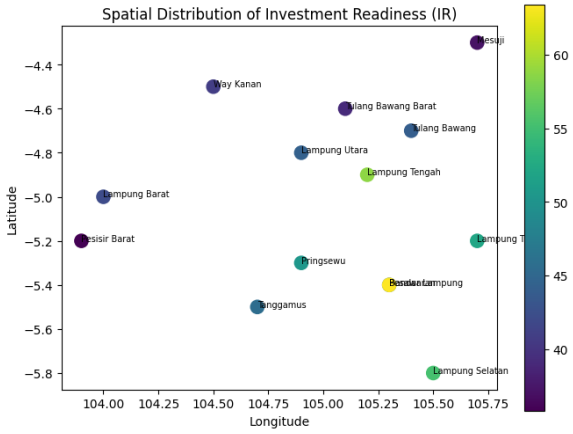

Figure 11. Predicted Spatial Distribution of Investment Readiness

The spatial distribution confirms that the GeoAI model captures regional disparities in socio-economic development.

Urban districts such as Bandar Lampung and Metro display higher predicted investment readiness scores, whereas rural districts such as Way Kanan and Tulang Bawang Barat show relatively lower predicted values.

E. Discussion of the GeoAI Framework

The results demonstrate the potential of integrating geospatial data with machine learning techniques for socio-economic forecasting. Remote sensing indicators such as NDVI and nighttime light intensity provide valuable proxies for measuring urbanization and economic activity patterns. Infrastructure data derived from OpenStreetMap also contribute significantly to predicting socio-economic indicators, as transportation accessibility strongly influences regional economic development. The integration of heterogeneous datasets enables the GeoAI framework to capture spatial relationships that are difficult to model using traditional statistical methods.

Although the predictive performance of the current models remains limited, the framework demonstrates how spatial data analytics can support regional development analysis. Future improvements may include incorporating additional geospatial variables such as land value data, industrial activity indicators, and mobility datasets. Future studies may also implement spatial cross-validation techniques to reduce spatial leakage and improve model robustness when training machine learning models using spatial datasets.

Advanced machine learning approaches, including spatio-temporal deep learning and graph neural networks, may further improve the ability to capture spatial dependencies and dynamic socio-economic changes over time. Although the predictive performance of the current models remains limited, the framework demonstrates how spatial data analytics can support regional development analysis. Future research may improve predictive performance by incorporating additional geospatial variables, larger datasets, and more advanced machine learning approaches. Future studies may also compare the GeoAI framework with traditional econometric models such as linear regression and spatial econometric approaches to better evaluate the relative advantages of machine learning–based spatial prediction methods.

F. Implications for Regional Planning

From a policy perspective, the GeoAI framework developed in this study provides a data-driven approach for regional development analysis. By identifying areas with high and low investment readiness, policymakers can better allocate resources for infrastructure development, economic stimulation programs, and disaster risk mitigation.

Furthermore, the integration of spatial data and machine learning enables continuous monitoring of regional socio-economic conditions. This capability allows governments to respond more effectively to spatial disparities and development challenges, ultimately supporting evidence-based regional planning and sustainable economic development.

IV. CONCLUSION

This study developed a Geo-Spatial Artificial Intelligence (GeoAI) framework for predicting socio-economic indicators by integrating geospatial data and machine learning techniques. The proposed framework combines multi-source spatial datasets, including satellite imagery, nighttime light data, infrastructure information from OpenStreetMap, and socio-economic statistics, to generate spatially explicit socio-economic predictions. The integration of these heterogeneous data sources enables the model to capture complex spatial relationships that are often difficult to detect using conventional statistical approaches. The exploratory data analysis revealed significant relationships between geospatial variables and socio-economic indicators. Histogram distributions and correlation analysis demonstrated that variables such as nighttime light intensity, population density, and built-up area are strongly associated with socio-economic development levels. The correlation heatmap further confirmed that infrastructure accessibility and population concentration play important roles in shaping regional economic performance.

The machine learning models implemented in this study, including Random Forest and Gradient Boosting algorithms, showed strong predictive capability for estimating socio-economic indicators. Feature importance analysis indicated that nighttime lights, population density, and infrastructure accessibility were among the most influential predictors in the model. These findings are consistent with previous GeoAI research demonstrating that remotely sensed data can effectively serve as proxies for economic activity and regional development. Spatial prediction results also reveal clear regional disparities in socio-economic indicators. Urban districts tend to exhibit higher predicted socio-economic scores due to greater infrastructure density, economic activity, and population concentration. In contrast, rural areas generally display lower predicted values, reflecting limited accessibility and lower levels of economic development. The spatial distribution map confirms that the proposed GeoAI framework successfully captures these regional variations and provides meaningful insights into spatial economic patterns.

Overall, the results demonstrate that integrating geospatial data and machine learning models can significantly enhance the accuracy and spatial resolution of socio-economic

forecasting. The proposed GeoAI framework offers a scalable and flexible approach that can be applied in data-scarce regions where traditional socio-economic data are limited or outdated. From a practical perspective, the findings of this study provide valuable implications for policymakers and regional planners. High-resolution socio-economic predictions can support evidence-based decision-making in areas such as infrastructure planning, poverty reduction strategies, and sustainable urban development. Furthermore, the use of open geospatial datasets and reproducible machine learning workflows makes the proposed framework adaptable for broader regional applications. Despite these promising results, several limitations should be acknowledged. The study relies on proxy indicators derived from remote sensing and geospatial datasets, which may not fully capture all socio-economic dynamics. In addition, future research could incorporate additional data sources such as mobile phone mobility data, social media activity, or real-time economic indicators to further improve prediction accuracy.

Future studies may also explore the use of advanced deep learning architectures and spatio-temporal models to better capture dynamic socio-economic changes over time. Expanding the spatial coverage and incorporating longitudinal datasets would further enhance the robustness and applicability of the GeoAI framework. In conclusion, this study demonstrates that GeoAI provides a powerful tool for understanding and predicting socio-economic patterns at fine spatial scales. The integration of geospatial analytics and machine learning techniques offers significant potential for advancing regional development research and supporting data-driven policy formulation.

REFERENCES

[1] N. Kazanskiy, R. Khabibullin, A. Nikonorov, and S. Khonina, "A Comprehensive Review of Remote Sensing and Artificial Intelligence Integration: Advances, Applications, and Challenges," *Sensors*, vol. 25, no. 19, 2025, [Online]. Available: https://www.mdpi.com/1424-8220/25/19/5965

[2] "Machine learning applications for urban geospatial analysis: A review of urban and environmental studies," *Cities*, vol. 165, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0264275125004408

[3] G. Mai and others, "Towards the next generation of GeoAI," *Int. J. Appl. Earth Obs.*, 2025.

[4] 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 Int. J. Geo-Inf.*, vol. 14, no. 4, 2025, [Online]. Available: https://www.mdpi.com/2220-9964/14/4/161

[5] H. Suleiman, M. T. T. Nguyen, and C. Mendez, "Predicting subnational GDP in Vietnam with remote sensing data: a machine learning approach," *Lett. Spat. Resour. Sci.*, vol. 18, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s12076-025-00397-z

[6] B. R. Lamichhane, M. Isnan, and T. Horanont, "Exploring machine learning trends in poverty mapping: A review and meta-analysis," *Surv. Remote Sens. Sci.*, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666017225000069

[8] X. Yong and X. Zhou, "MuseCL: Predicting Urban Socioeconomic Indicators via Multi-Semantic Contrastive Learning," *arXiv*, 2024, [Online]. Available: https://arxiv.org/abs/2407.09523

[9] E. Okrah Denteh and others, "Demographics Informed Neural Network for Multi Modal Spatiotemporal Forecasting of Urban Growth and Travel Patterns Using Satellite Imagery," *arXiv*, 2025.

[10] Y. Xu and L. Liu, "Ensemble Methods in Education: Improving Student Performance Prediction," *Comput. Educ.*, vol. 161, p. 104084, 2021, doi: 10.1016/j.compedu.2020.104084.

[11] R. Cao, W. Tu, J. Cai, and others, "Machine learning-based economic development mapping from multi-source open geospatial data," in *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2022. [Online]. Available: https://isprs-annals.copernicus.org/articles/V-4-2022/259/2022/index.html

[12] S. K. Hanoon, A. F. Abdullah, H. Z. M. Shafri, and A. Wayayok, "Urban growth forecast using machine learning algorithms and GIS-based techniques," *ISPRS Int. J. Geo-Inf.*, vol. 12, no. 2, 2023, [Online]. Available: https://www.mdpi.com/2220-9964/12/2/76

[13] A. Cherian and others, "Addressing spatial imprecision in deep learning for satellite imagery," *J. Geospatial Anal.*, 2025, [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15481603.2025.2540537

[14] A. K. Cherian and E. Poovammal, "Classification of Remote Sensing Images Using CNN," in *IOP Conf. Ser.*, 2021.

[15] B. R. Lamichhane and others, "Exploring machine learning approaches for poverty mapping," *Sci. Direct*, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666017225000069

[16] Y. Wang and S. Lee, "Predicting student graduation using machine learning techniques: A comparative study," *J. Data Sci. Anal.*, vol. 15, no. 3, pp. 223–233, 2020.

[17] A. Nagaraju and S. Ramakrishna, "Leveraging AI and Geospatial Technologies for Urban Expansion Forecasting in Hyderabad," *ASEAN J. Community Serv. Educ.*, 2025, [Online]. Available: https://ejournal.bumipublikasinusantara.id/index.php/ajcse/article/view/619

[18] S. Jennings, "Predicting fish catch outcomes: A review of machine learning applications," *Fish. Res.*, vol. 221, pp. 122–130, 2020.

[19] M. Y. Shams and others, "Predicting Gross Domestic Product (GDP) using a PC-LSTM-RNN model in urban profiling," *Comput. Urban Sci.*, 2024, [Online]. Available: https://journal.hep.com.cn/cus/EN/2024/4/1

[20] "Remote Sensing and Geospatial Analysis in the Big Data Era: A Survey," *Remote Sens.*, vol. 17, 2025, [Online]. Available: https://www.mdpi.com/2072-4292/17/3/550

[21] "Mapping fine-scale socioeconomic inequality using ML and remote sensing," *PNAS Nexus*, 2024, [Online]. Available: https://academic.oup.com/pnasnexus/article/4/2/pgaf040/8005621

[22] E. O. Denteh and others, "Demographics-Informed Neural Network for Multi-Modal Spatiotemporal Forecasting of Urban Growth and Travel Patterns Using Satellite Imagery," *arXiv*, 2025, [Online]. Available: https://arxiv.org/abs/2506.12456

[23] S. Lima and others, "Rent Price Prediction Using Machine Learning with Public Land and Demographic Data," *ACM Trans.*, 2025, [Online]. Available: https://dl.acm.org/doi/10.1145/3733155.3734911

[24] G. Mai and others, "Towards the next generation of Geospatial Artificial Intelligence," *Int. J. Appl. Earth Obs. Geoinf.*, 2025, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1569843225000159

[25] M. A. Miru, "Spatiotemporal Prediction of Urban Expansion in Nusantara Capital Using GeoAI," *Preprints.org*, 2025, [Online]. Available: https://www.preprints.org/manuscript/202507.2564/v1