

Land Subsidence Prediction and Geo Environmental Predictor Importance Assessment in Jakarta, Semarang, and Pekalongan Using InSAR and Ensemble Machine Learning

Luthfy Azh-Zhaviery Puary¹, Panca Dewi Pamungkasari¹, Endah Tri Esti Handayani¹, Rendi Handika², Babag Purbantoro², Joko Widodo², Damy Matheus Heezen³

¹Universitas Nasional Faculty of Communication and Information Technology, Jakarta, Indonesia

²Research and Innovation Agency (BRIN), Jakarta 10340, Indonesia

³Changemakers Network B.V., Netherlands

*Correspondence: luthfyzzhaviery@gmail.com

SUBMITTED: 5 June 2026; REVISED: 3 September 2026; ACCEPTED: 6 September 2026

ABSTRACT: Land subsidence along the northern coast of Java (Pantura) poses a significant threat to coastal infrastructure and flood resilience in densely populated areas. This study integrates 12 geo-environmental variables with Persistent Scatterer Interferometric Synthetic Aperture Radar (PS-InSAR) data to predict land subsidence in DKI Jakarta, Semarang City, and Pekalongan Regency using Random Forest (RF) and XGBoost regression models. Model performance was evaluated using mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2) on an independent 20% holdout dataset. Across the three study areas, the optimized ensemble models demonstrated strong predictive performance. RF outperformed XGBoost in DKI Jakarta (image, image) and Semarang City (image, image), whereas both models showed comparable performance in Pekalongan Regency (image). Overall, RF consistently achieved equal or superior predictive accuracy, although the best performance in Pekalongan Regency was obtained using RF without hyperparameter tuning. Feature importance analysis identified GEOLOGY_CODE as the dominant predictor in DKI Jakarta, while Land Surface Temperature (LST) was consistently important in Semarang City and Pekalongan Regency. Spatial analysis further identified Penjaringan, Genuk, and Kedungwuni as areas experiencing high levels of subsidence. The proposed framework provides a reproducible approach for district-scale land-subsidence assessment and supports risk-informed spatial planning across the Pantura coastal corridor.

KEYWORDS: Land subsidence; InSAR; machine learning; XGBoost; random forest; pantura corridor.

1. Introduction

Land subsidence along the northern coast of Java (Pantura) is a widespread geo-environmental hazard that threatens coastal infrastructure, livelihoods, and vulnerable communities [1]. Although gradual surface lowering can occur naturally through geological processes, such as

the compaction of sediments over geological timescales [2], the alarming rates of present-day subsidence are increasingly driven by human activities. Rapid urban expansion, intensive building development, and excessive groundwater extraction can disrupt aquifer systems and accelerate land subsidence [3]. Without effective intervention, continued subsidence may result in severe and potentially irreversible consequences, including long-term tidal inundation and increased exposure to coastal flooding [4].

In this context, Persistent Scatterer Interferometric Synthetic Aperture Radar (PS-InSAR) is a well-established geodetic technique capable of detecting surface displacement with millimeter-scale accuracy. Recent InSAR monitoring studies have demonstrated that multi-temporal surface displacement measurements can reliably capture subsidence patterns across vulnerable Indonesian coastal landscapes [5]. Nevertheless, geodetic radar methods are primarily diagnostic rather than prognostic, making them well suited to monitoring and reconstructing historical displacement but limited in their ability to predict future subsidence. Furthermore, the complex and nonlinear relationships among multiple geo-environmental drivers make future subsidence patterns difficult to predict using individual variables or conventional approaches. Machine learning has therefore been increasingly employed to address this limitation [6]. Tree-based ensemble algorithms, particularly Random Forest and XGBoost, have demonstrated strong capabilities in modelling complex geospatial relationships and mapping hazard susceptibility [7, 8].

Despite these computational advances, several empirical research gaps remain. Most machine-learning studies of land subsidence have focused on individual cities [3, 8, 9], limiting broader comparisons across multiple urban areas within the same coastal region. In addition, preprocessing procedures for controlling structural noise are often insufficiently addressed in existing workflows [3, 9, 10], allowing localized uplift artifacts to potentially influence regression-tree learning and subsequent predictions. Direct comparisons between sequential boosting algorithms and parallel bagging approaches under comparable coastal conditions also remain limited [8, 9, 11]. These gaps highlight the need for a consistent multi-city framework that enables robust comparison of machine-learning models and their associated geo-environmental drivers.

To address these gaps, this study presents a regional benchmarking analysis comparing XGBoost and Random Forest across three major land-subsidence hotspots along the northern coast of Java: Jakarta Special Capital Region (DKI Jakarta), Semarang City, and Pekalongan Regency. The analysis integrates twelve geo-environmental variables representing 2022 conditions with multi-temporal InSAR datasets covering the 2021–2022 period. A systematic data-cleansing protocol based on statistical thresholds was applied to minimize the influence of localized uplift artifacts and ensure that subsidence-related ground motion was appropriately represented in the predictive analysis. The study aims to compare the predictive performance of XGBoost and Random Forest at the urban scale, identify the relative importance of geo-environmental predictors of land subsidence in each study area, and evaluate the spatial distribution and regional consistency of prediction errors to support coastal risk management and spatial planning across the Pantura corridor.

2. Materials and Methods

2.1. Research Location

The regional benchmarking study was conducted across three coastal hotspots experiencing severe land subsidence along the northern coast of Java (Pantura), Indonesia: DKI Jakarta, Semarang City, and Pekalongan Regency (Figure 1). These study areas were selected to evaluate the generalizability and predictive performance of tree-based ensemble algorithms across different urban settings and geo-environmental conditions. DKI Jakarta (6.1750° S, 106.8275° E) is a densely developed megacity characterized by substantial infrastructure loads and intensive groundwater extraction from local alluvial aquifers, both of which contribute to ongoing surface deformation. In contrast, Semarang City (6.9926° S, 110.4285° E) has complex topography, comprising hilly volcanic landforms in the south and highly compressible coastal alluvial plains in the north. In this area, infrastructure loading and natural sediment compaction contribute to land subsidence. Pekalongan Regency (6.9685° S, 109.6151° E) is a low-lying coastal district characterized by mangrove–fishpond systems and decentralized home-based textile industries, with substantial reliance on groundwater resources. As illustrated in Figure 1, the three study areas represent different spatial and geo-environmental settings along the Pantura coastal corridor. This multi-city regional framework enables the machine-learning models to be evaluated across contrasting urban and environmental conditions and provides a basis for assessing the spatial transferability and stability of land-subsidence predictions.

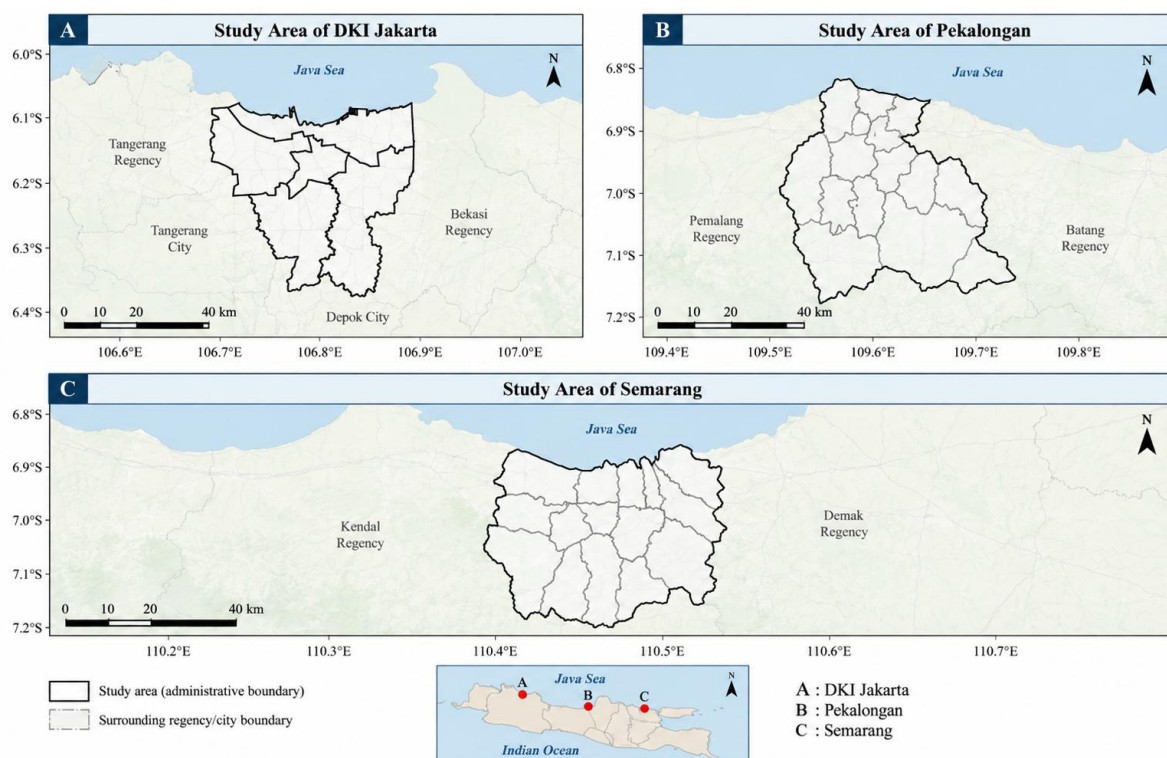


Figure 1. Study area of Jakarta (A), Pekalongan (B), and Semarang (C).

2.2. Spatial Datasets and Environmental Covariates.

The computational matrix incorporated two classes of spatial data: a geodetic target variable representing surface displacement and a multidimensional set of geo-environmental predictor variables. The target variable, representing the historical mean vertical surface deformation velocity (VEL; mm/year), was obtained through a formal research collaboration with the National Research and Innovation Agency (BRIN). The raw dataset comprised high-density Persistent Scatterer Interferometry (PSI) point clouds extracted from multi-temporal Sentinel-1A Single Look Complex (SLC) radar imagery using SARPROZ processing software. The raw tabular data included key geodetic attributes, such as unique scatterer IDs, latitude and longitude coordinates, height differences (HEIGHT), heights relative to the digital elevation model (HEIGHT WRT DEM), and multi-temporal deformation phase histories acquired from 2021 to 2022.

The VEL values generated through SARPROZ processing represent Line-of-Sight (LOS) displacement. Following standard geometric projection for Sentinel-1 Interferometric Wide (IW) mode ascending geometry, with a nominal incidence angle of approximately $\theta = 38^\circ$, the vertical displacement component was derived using a correction factor of $1/\cos(\theta) \approx 1.27$. This procedure converted LOS velocities into their corresponding vertical components. The correction accounts for the angular difference between the satellite's slant-range observation vector and the vertical displacement axis, ensuring that the VEL values incorporated into the machine-learning matrix consistently represent vertical deformation rates in mm/year.

The twelve predictor variables were harmonized to a 2022 baseline to represent environmental and socio-spatial conditions corresponding to the latter phase of the 2021–2022 InSAR observation period. For temporally dynamic remote-sensing variables, including Land Surface Temperature (LST) and the Normalized Difference Vegetation Index (NDVI), annual median composites derived from cloud-masked 2022 image collections were used. LAND_COVER was represented by a 2022 land-cover classification derived from Sentinel-2 imagery, while POPULATION was represented using the corresponding 2022 gridded population dataset. GEOLOGY_CODE, DIST_COAST, DIST_ROAD, DIST_RIVER, TWI, ELEVATION, and SLOPE were obtained from spatially static or 2022-aligned datasets and their derived spatial attributes. COAST_CHANGE was represented by a shoreline-change metric calculated from shoreline observations between 2021 and 2022 and used as the baseline predictor for 2022. Given the relatively stable characteristics of the selected geo-environmental covariates over annual timescales, this temporal alignment was considered appropriate for the purposes of the study.

A synchronized matrix of twelve geo-environmental and anthropogenic predictor variables was subsequently constructed and spatially aligned with the geodetic target coordinates to model the spatial factors associated with vertical surface deformation. Rather than relying on a single data provider, the study integrated continuous and categorical variables derived from multiple spatial datasets, global demographic repositories, and cloud-based remote-sensing platforms. As summarized in Supplementary Table S1, the predictor variables encompass geological, hydrological, topographic, demographic, land-cover, thermal, vegetation, and coastal characteristics. To ensure methodological transparency and reproducibility, Supplementary Table S1 also provides the corresponding data sources and extraction procedures for each predictor variable.

2.3. Multi-Temporal InSAR Processing and Data Cleansing (Noise Filtering).

A critical step in the data-engineering workflow was the preparation of raw geodetic observations into a high-integrity training matrix suitable for tree-based regression algorithms. PS-InSAR processing generates high-density surface deformation measurements; however, the raw point clouds may contain structural and phase-related noise artifacts. In the context of land subsidence modelling, persistent scatterer points with positive velocity values ($VEL > 0$) represent upward surface displacement. Because this study focused specifically on land subsidence, positive velocity observations were excluded from the training dataset. This approach is consistent with previous InSAR-based subsidence studies that isolate subsidence-specific signals from broader deformation datasets when the analysis is restricted to downward surface displacement [12]. Accordingly, a deterministic threshold filter was applied consistently across the three study areas to retain only observations representing downward surface deformation ($VEL \leq 0$). This preprocessing step ensured that the training datasets consistently represented the target deformation class and prevented positive displacement observations from influencing subsequent spatial regression analyses. For DKI Jakarta, the filtering procedure reduced the dataset from 308,586 raw persistent scatterer points to 120,482 subsidence points. In Semarang City, 65,670 initial measurement points were reduced to 27,333 subsidence points, while in Pekalongan Regency, the dataset decreased from 23,607 to 7,092 points. Overall, the filtering procedure substantially reduced the number of observations across all three study areas while retaining records corresponding to downward deformation. The resulting datasets formed the cleaned geodetic target matrices used for subsequent machine-learning analysis.

2.4. Proposed Machine Learning Workflow.

This study adopted a five-phase modular workflow to systematically model the relationships between multi-source geo-environmental predictors and vertical surface deformation across the three study areas. The first phase involved data acquisition and integration. Raw PS-InSAR displacement velocity data obtained from BRIN through SARPROZ processing were integrated with twelve geo-environmental predictor variables representing topographic, hydrological, geological, and anthropogenic characteristics of DKI Jakarta, Semarang City, and Pekalongan Regency. These datasets were harmonized into a unified tabular matrix through coordinate-based spatial matching.

The second phase involved data preprocessing and noise control. A deterministic threshold filter was applied to exclude persistent scatterer observations with positive velocity values ($VEL > 0$), thereby restricting the target dataset to downward surface deformation ($VEL \leq 0$). This procedure ensured consistency with the study objective of specifically modelling land subsidence and prevented observations outside the target deformation class from influencing model training. The third phase involved data partitioning. The cleaned dataset for each study area was randomly divided into an 80% training subset and a 20% independent holdout testing subset using a fixed random seed of 42 to ensure computational reproducibility. This 80/20 holdout strategy was designed to evaluate fine-scale prediction and intra-regional deformation patterns within each study area rather than model transferability across spatially independent domains. Because random partitioning may retain spatial autocorrelation between training and testing observations, this approach does not constitute fully spatially independent

validation. More rigorous evaluation of spatial generalizability using spatial block cross-validation is therefore recommended for future research.

The fourth phase involved machine-learning modelling using two tree-based ensemble architectures. The Random Forest Regressor constructs an ensemble of independent decision trees through bootstrap aggregation (bagging), using randomized samples with replacement to reduce model variance. In contrast, the XGBoost Regressor employs a sequential gradient-boosting strategy in which successive trees are fitted to residual errors generated by the preceding ensemble. XGBoost also incorporates regularization to reduce the risk of overfitting in high-dimensional datasets. A fixed random state of 42 was used where applicable to ensure reproducibility. The fifth phase involved model evaluation using the independent 20% holdout dataset. Predictive performance was assessed using three regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The analytical outputs also included predictor correlation matrices [13], feature-importance rankings, observed-versus-predicted scatter plots, and spatial maps comparing PS-InSAR observations with machine-learning predictions [10, 14]. Computational reproducibility was maintained by applying a fixed random state of 42 to stochastic procedures, including the 80/20 train–test split, Random Forest modelling, and relevant XGBoost procedures. The geospatial data-engineering and model-training workflow was implemented in Python v3.10.12. Scikit-learn v1.2.2 was used for data preprocessing, Random Forest modelling, and hyperparameter optimization using RandomizedSearchCV, while XGBoost v1.7.6 was used for gradient-boosting regression. Geospatial processing, including point-in-polygon operations and administrative-boundary validation, was performed using GeoPandas v0.13.2 and Shapely v2.0.1.

2.5. Tree-Based Ensemble Modeling Architecture.

Random Forest is an ensemble-learning algorithm that constructs multiple independent decision trees to model complex nonlinear relationships in spatial regression problems. Based on the principle of bootstrap aggregation (bagging), each decision tree is trained using a bootstrap sample drawn randomly with replacement from the original training dataset. During the node-splitting process, only a random subset of the twelve geo-environmental predictor variables is considered. This procedure reduces correlations among individual trees and, consequently, decreases overall model variance. For a spatial input vector containing the local geo-environmental attributes, the final land-subsidence velocity is predicted by averaging the outputs of all individual regression trees in the ensemble. The mathematical aggregation function is expressed as:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (1)$$

where B represents the total number of decision trees in the forest, and the corresponding prediction term represents the continuous deformation velocity predicted by the b -th regression tree. By averaging predictions across multiple independently constructed trees, Random Forest reduces the variance associated with individual models and provides robustness against

overfitting, particularly when applied to large and complex spatial datasets [9]. The overall Random Forest regression architecture used in this study is illustrated in Figure 2.

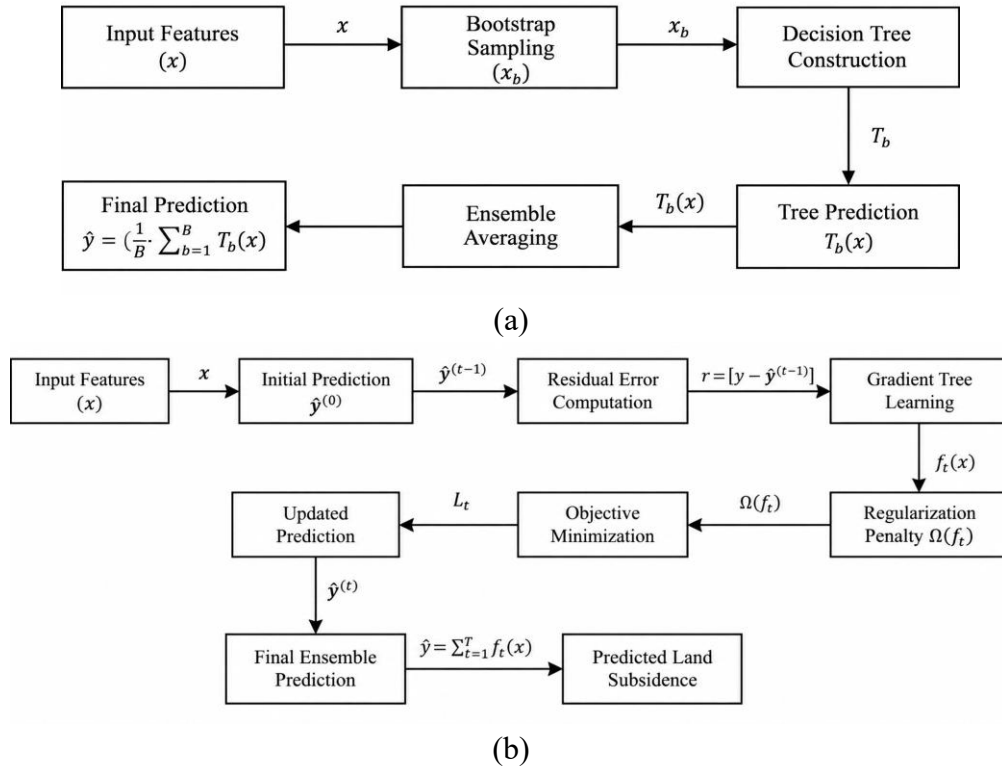


Figure 2. (a) Flow of Random Forest Regressor Architecture. (b) Flow of XGBoost Regressor Architecture.

Unlike the parallel bagging approach employed by Random Forest, the XGBoost Regressor uses a sequential gradient-boosting architecture in which decision trees are constructed iteratively to minimize prediction errors. At each iteration, a new tree is fitted to the residual information from the preceding ensemble, progressively improving the model prediction. XGBoost also incorporates an explicit regularization term into its objective function to control tree complexity and reduce the risk of overfitting, particularly when modelling high-dimensional geospatial datasets. At each iteration t , XGBoost minimizes a regularized objective function, $\mathcal{L}^{(t)}$, to evaluate the contribution of the newly added tree, f_t :

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (2)$$

With l being the differentiable loss function which defines the distance between actual InSAR velocity observation, y_i , and cumulative prediction from a previous step $\widehat{y}_i^{(t-1)}$. The second component, $\Omega(f_t)$ refers to the regularization penalty for directly penalizing structural model complexity which can take a form as defined below:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (3)$$

In this regularization equation, where T is the number of terminal nodes or leaves in the newly added tree f_t , and w_j^2 is the squared L_2 norm of leaf weights. The user-defined hyperparameters γ and λ are constants that scale the complexity penalty. XGBoost applies a second-order Taylor expansion of the loss function to efficiently optimize the regularized objective during model training. This procedure enables the algorithm to iteratively reduce prediction errors while controlling model complexity, thereby balancing predictive accuracy, computational efficiency, and resistance to overfitting. In this study, XGBoost was used as the sequential boosting counterpart to the parallel bagging architecture of Random Forest. As illustrated in Figure 2, the XGBoost workflow progresses from the initial prediction and residual-error computation through gradient-tree learning, regularization, objective minimization, and iterative prediction updating to generate the final ensemble prediction of land subsidence..

2.6. Performance Evaluation Metrics.

Three standard regression metrics were used to evaluate and compare the predictive performance of the Random Forest and XGBoost models: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Model performance was evaluated using the independent 20% holdout testing dataset. Together, these metrics provide complementary measures of average prediction error, sensitivity to large residuals, and the proportion of variance in observed land-subsidence velocities explained by the model.

2.6.1. Mean Absolute Error (MAE).

Mean Absolute Error (MAE) measures the average absolute difference between the land-subsidence velocities predicted by the machine-learning model and those observed using MT-InSAR. Because MAE treats all residuals linearly without squaring them, it provides an intuitive measure of the average prediction error expressed in the same unit as the target variable (mm/year). MAE is calculated as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

where n represents the total number of measurement points in the independent testing subset, y_i denotes the actual annual land subsidence velocity measured via MT-InSAR, and $\hat{y}_i$ is the continuous deformation rate predicted by the ensemble model for the corresponding spatial coordinates.

2.6.2. Root Mean Squared Error (RMSE).

Root Mean Squared Error (RMSE) measures the square root of the average squared difference between predicted and observed land-subsidence velocities. Unlike MAE, RMSE assigns greater weight to larger residuals because prediction errors are squared before averaging. Consequently, RMSE is particularly sensitive to localized observations where the model substantially overpredicts or underpredicts the magnitude of land subsidence. RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

where n represents the number of observations in the testing subset, y_i is the observed land-subsidence velocity, and $\hat{y}_i$ is the corresponding model-predicted value. Lower RMSE values indicate better predictive performance. A substantial difference between RMSE and MAE may indicate the presence of relatively large prediction errors or greater variability in residual magnitudes across specific coastal areas [9].

2.6.3. Coefficient of Determination (R^2).

The overall goodness of fit of each regression model was evaluated using the coefficient of determination (R^2). This metric quantifies the proportion of variance in the observed InSAR-derived land-subsidence velocities explained by the model predictions. R^2 has a maximum value of 1.0, representing perfect agreement between observed and predicted values, whereas values below zero indicate that the model performs worse than a simple prediction based on the mean of the observed data. R^2 is calculated as follows::

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

The term $\bar{y}$ refers to the average of actual annual land subsidence velocity observations within the active validation grid. Higher R^2 scores suggest the tree-based ensemble architecture has a better ability to map underlying environmental and anthropogenic drivers causing land subsidence patterns.

3. Results and Discussion

3.1. Multicollinearity and Multivariate Spatial Correlation Exploration.

Before training and evaluating the tree-based ensemble models, Pearson correlation analysis was conducted among the twelve geo-environmental and anthropogenic predictor variables for DKI Jakarta, Semarang City, and Pekalongan Regency. This exploratory analysis was performed to identify potentially strong correlations among predictors that could influence model interpretation and obscure the relative contribution of individual variables [15]. The resulting correlation coefficients are presented as lower-triangular heatmaps in Figure 3, allowing the correlation structures of the three study areas to be compared directly [11].

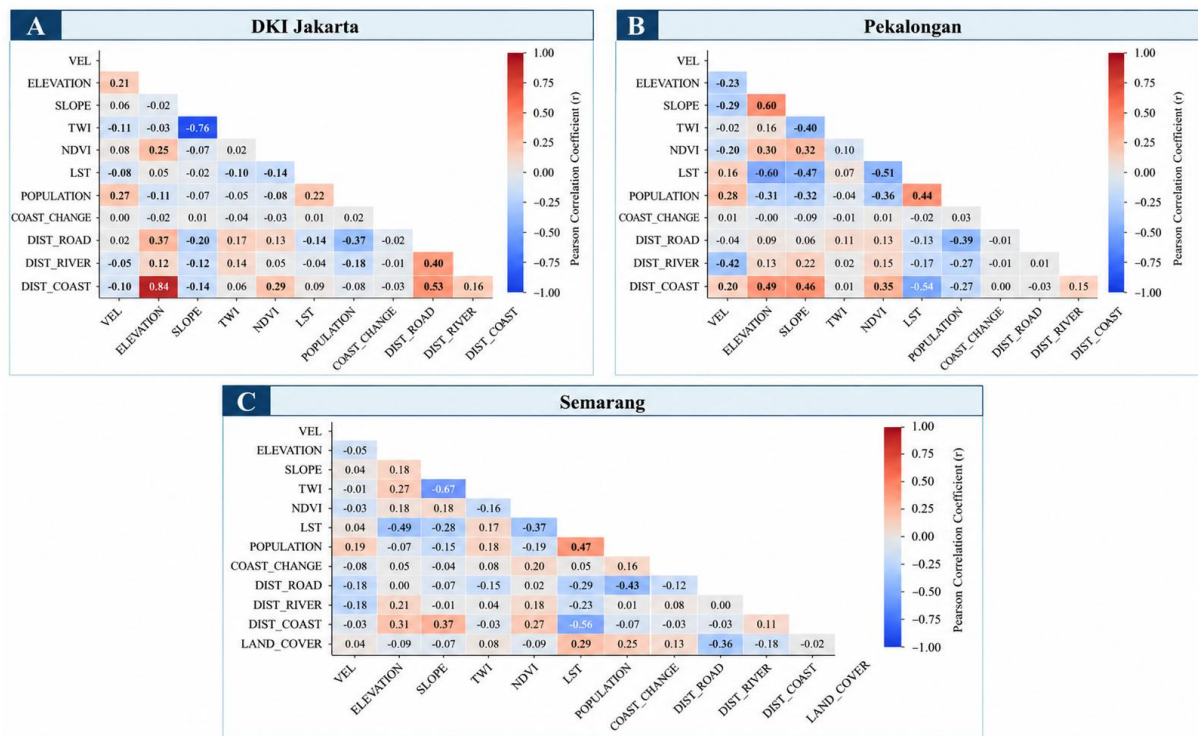


Figure 3. Pearson correlation matrices of the twelve geo-environmental and anthropogenic variables for (A) DKI Jakarta, (B) Pekalongan Regency, and (C) Semarang City.

Figure 3. Pearson correlation matrices of the twelve geo-environmental and anthropogenic predictor variables for DKI Jakarta, Semarang City, and Pekalongan Regency. The lower-triangular heatmaps illustrate the direction and magnitude of pairwise correlations among the predictor variables. DKI Jakarta exhibits a strong positive correlation between DIST_COAST and ELEVATION ($r = 0.84$). This relationship reflects Jakarta's spatial topography, where elevation generally increases with distance from the northern coastline toward the higher terrain in the south. Although this correlation indicates substantial collinearity between the two variables, both predictors were retained because they represent distinct spatial characteristics and tree-based ensemble models can accommodate correlated predictors through nonlinear partitioning. No other predictor pairs in DKI Jakarta exhibited correlations of comparable magnitude, indicating relatively limited pairwise collinearity among the remaining variables.

The correlation structures of Pekalongan Regency and Semarang City differ from that of DKI Jakarta (Figure 3). In both Central Java study areas, Land Surface Temperature (LST) was negatively correlated with DIST_COAST, with $r = -0.54$ in Pekalongan Regency and $r = -0.56$ in Semarang City. In contrast, LST was positively correlated with POPULATION, with $r = 0.44$ in Pekalongan Regency and $r = 0.47$ in Semarang City. These similar correlation patterns indicate spatial associations among coastal proximity, population density, and surface thermal conditions in the two study areas.

The observed relationships between LST and population density may reflect the concentration of warmer surface conditions within densely developed urban and industrial areas. These areas are generally characterized by greater built-up density, lower vegetation cover, and more intensive anthropogenic activity. Such spatial characteristics are relevant to land-subsidence assessment because urban development and groundwater extraction have been identified as important contributors to subsidence in Jakarta and other parts of Java [1]. Overall,

the correlation analysis indicates that, apart from the strong relationship between DIST_COAST and ELEVATION in DKI Jakarta, the predictor set does not exhibit widespread extreme pairwise correlations ($|r| > 0.90$). The twelve variables were therefore retained for subsequent machine-learning analysis.

3.2. Algorithmic Performance and Benchmarking Evaluation.

Before evaluating model performance, hyperparameter optimization was conducted to identify suitable configurations for the Random Forest and XGBoost regressors. RandomizedSearchCV was applied using predefined search spaces to ensure a consistent optimization procedure across DKI Jakarta, Semarang City, and Pekalongan Regency. The search spaces and selected hyperparameter configurations are summarized in Table 1.

Table 1. Hyperparameter Search Spaces and Optimized Configuration Results.

Algorithm & Hyperparameters	Search Space Range	Optimized configuration (Best parameter)
(Random Forest Regressor)		
n_estimators	[50, 100, 200, 300]	100
max_depth	[10, 20, 30, None]	20
min_samples_split	[2, 5, 10]	2
min_samples_leaf	[1, 2, 4]	1
max_features	['sqrt', 'log2', None]	None
(XGBoost Regressor)		
n_estimators	[100, 200, 300, 500]	200
max_depth	[3, 6, 9, 12]	9
learning_rate	[0.01, 0.05, 0.1, 0.2]	0.1
Subsample	[0.6, 0.7, 0.8, 0.9, 1.0]	0.9
colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1.0]	0.9
reg_alpha	[0, 0.1, 1, 10]	0.1
reg_lambda	[1, 5, 10, 100]	5

As shown in Table 1, the selected Random Forest configuration consisted of 100 estimators, a maximum tree depth of 20, a minimum of two samples required to split an internal node, a minimum of one sample per leaf, and the use of all available features at each split. The selected XGBoost configuration used 200 estimators, a maximum tree depth of 9, and a learning rate of 0.1. Subsample and colsample_bytree were both set to 0.9, while L1 and L2 regularization were controlled using reg_alpha = 0.1 and reg_lambda = 5, respectively. These parameters were used to constrain model complexity and reduce the risk of overfitting.

The predictive performance of the baseline and tuned models was subsequently evaluated using the independent 20% holdout dataset. The results for DKI Jakarta, Pekalongan Regency, and Semarang City are presented in Table 2. Overall, hyperparameter tuning consistently improved XGBoost performance across all three study areas, whereas the baseline Random Forest produced slightly better holdout metrics than its tuned counterpart. Although the baseline Random Forest achieved slightly better holdout metrics than the tuned configuration in all three study areas, the differences were small. The tuned Random Forest was retained for subsequent feature-importance and spatial prediction analyses to maintain consistency with the predefined optimization workflow. In particular, tuning constrained the maximum tree depth to 20, whereas the baseline model allowed unrestricted tree depth. Nevertheless, the tuned configuration did not improve the reported holdout performance; therefore, no claim is made that hyperparameter tuning enhanced Random Forest predictive accuracy.

Table 2. Model Performance Metrics Comparison across Jakarta, Pekalongan, and Semarang.

Location	Model	R ² Score (Baseline/tuned)	MAE (Baseline/Tuned)	RMSE (Baseline / Tuned)
Jakarta	XGBoost	0.8081 / 0.8482	2.35 / 1.92	4.62 / 4.11
	Random Forest	0.8710 / 0.8633	1.32 / 1.52	3.79 / 3.90
Pekalongan	XGBoost	0.7637 / 0.8017	9.39 / 8.34	14.11 / 12.92
	Random Forest	0.8017 / 0.8006	8.02 / 8.09	12.92 / 12.96
Semarang	XGBoost	0.7352 / 0.7956	9.53 / 7.77	13.79 / 12.11
	Random Forest	0.8122 / 0.8090)	6.84 / 7.00	11.61 / 11.71

In contrast, hyperparameter optimization consistently improved XGBoost performance across all three study areas. As shown in Table 2, R² increased from 0.8081 to 0.8482 in DKI Jakarta, from 0.7637 to 0.8017 in Pekalongan Regency, and from 0.7352 to 0.7956 in Semarang City. These improvements were accompanied by reductions in both MAE and RMSE, demonstrating a consistent benefit of tuning for the XGBoost models. In DKI Jakarta, the tuned Random Forest achieved an R² of 0.8633, MAE of 1.52 mm/year, and RMSE of 3.90 mm/year, outperforming the tuned XGBoost model, which achieved an R² of 0.8482, MAE of 1.92 mm/year, and RMSE of 4.11 mm/year. However, the baseline Random Forest remained the best-performing configuration overall, with an R² of 0.8710, MAE of 1.32 mm/year, and RMSE of 3.79 mm/year. Thus, although tuning substantially improved XGBoost performance, it did not surpass Random Forest in DKI Jakarta.

A similar pattern was observed in Semarang City. The tuned Random Forest achieved an R² of 0.8090, MAE of 7.00 mm/year, and RMSE of 11.71 mm/year, compared with an R² of 0.7956, MAE of 7.77 mm/year, and RMSE of 12.11 mm/year for tuned XGBoost. XGBoost nevertheless exhibited the largest improvement following optimization in this study area, with R² increasing by 0.0604 and both error metrics decreasing substantially. In Pekalongan Regency, the two tuned models showed nearly identical predictive performance. Tuned XGBoost achieved an R² of 0.8017, MAE of 8.34 mm/year, and RMSE of 12.92 mm/year, whereas tuned Random Forest achieved an R² of 0.8006, MAE of 8.09 mm/year, and RMSE of 12.96 mm/year. Thus, XGBoost produced a marginally higher R² and lower RMSE, while Random Forest produced a slightly lower MAE. These results indicate comparable predictive performance rather than a clear advantage for either tuned model.

The observed performance patterns can be interpreted in the context of the different ensemble-learning strategies employed by the two algorithms. XGBoost uses sequential boosting with explicit regularization, whereas Random Forest combines randomized decision trees through bootstrap aggregation [9]. These architectural differences provide a theoretical basis for understanding their different responses to hyperparameter optimization. However, the present analysis does not isolate the effects of individual algorithmic mechanisms; therefore, causal attribution would require dedicated ablation experiments or controlled sensitivity analyses. The relationship between observed and model-predicted deformation velocities for the three study areas is further illustrated by the validation scatter plots in Figure 4. Together with the quantitative metrics reported in Table 2, Figure 4 provides a visual assessment of the agreement between predicted and observed deformation velocities and highlights differences in prediction dispersion across the three study areas.

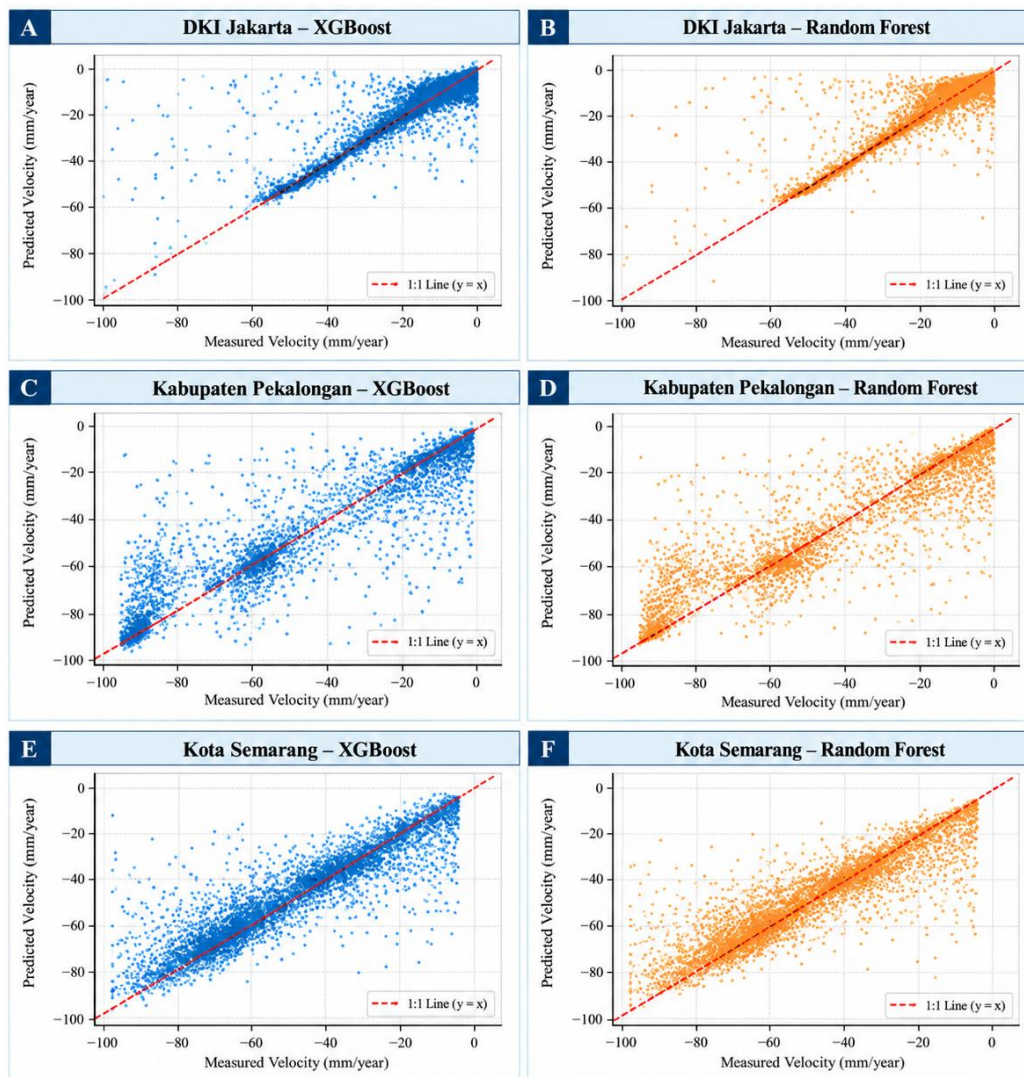


Figure 4. Validation scatter plots comparing predicted and observed deformation velocities (VEL) for XGBoost and Random Forest across the three study areas: (A–B) DKI Jakarta, (C–D) Pekalongan Regency, and (E–F) Semarang City.

3.3. Geo-Spatial Feature Importance and Factor Drivers Analysis.

The relative feature importance distributions reveal different physical and anthropogenic drivers of land subsidence in the three study areas. In DKI Jakarta, the two ensemble architectures attribute different importance hierarchies: XGBoost considers the GEOLOGY_CODE as the main predictor (34.44%), followed by DIST_COAST (17.30%) and DIST_ROAD (14.70%), while Random Forest assigns the highest importance to DIST_COAST (29.21%) and DIST_ROAD (23.65%). These differences indicate that the two ensemble architectures assigned different relative weights to the available environmental predictors when modeling surface-deformation variability [7, 8]. Such differences represent empirical statistical contributions rather than evidence of specific algorithmic causality. Both architectures indicate a consistent physical interpretation: Jakarta's subsidence is associated with geological conditions and intensive urban development, where dense urban areas coincide with compressible subsurface deposits [16, 17].

Land Surface Temperature (LST) consistently ranked among the dominant predictors across the Semarang and Pekalongan models, appearing within the top three predictors for three of the four model-city combinations and within the top four for all. In the Random Forest

configurations, LST ranked second in Semarang (15.82%) and third in Pekalongan (14.12%), while in XGBoost it ranked third in Semarang (12.17%) and fourth in Pekalongan (11.56%), where ELEVATION (17.39%) narrowly displaced it. The recurrent appearance of LST among the highest-ranked predictors suggests that urban thermal conditions are closely associated with the spatial distribution of subsidence across the Central Java coastal plain. In Semarang City, the Random Forest model weighted DIST_COAST (23.39%) and POPULATION (14.62%) as strong co-predictors alongside LST, while XGBoost distributed importance more evenly across GEOLOGY_CODE (20.70%), DIST_COAST (15.71%), and LST (12.17%). In Pekalongan Regency, ELEVATION (37.39%) and DIST_COAST (21.78%) ranked as the dominant predictors in the Random Forest model, while XGBoost distributed its weights across DIST_COAST (18.27%), GEOLOGY_CODE (17.43%), and LST (17.39%), with LST contributing (11.56%) (Figure 5).

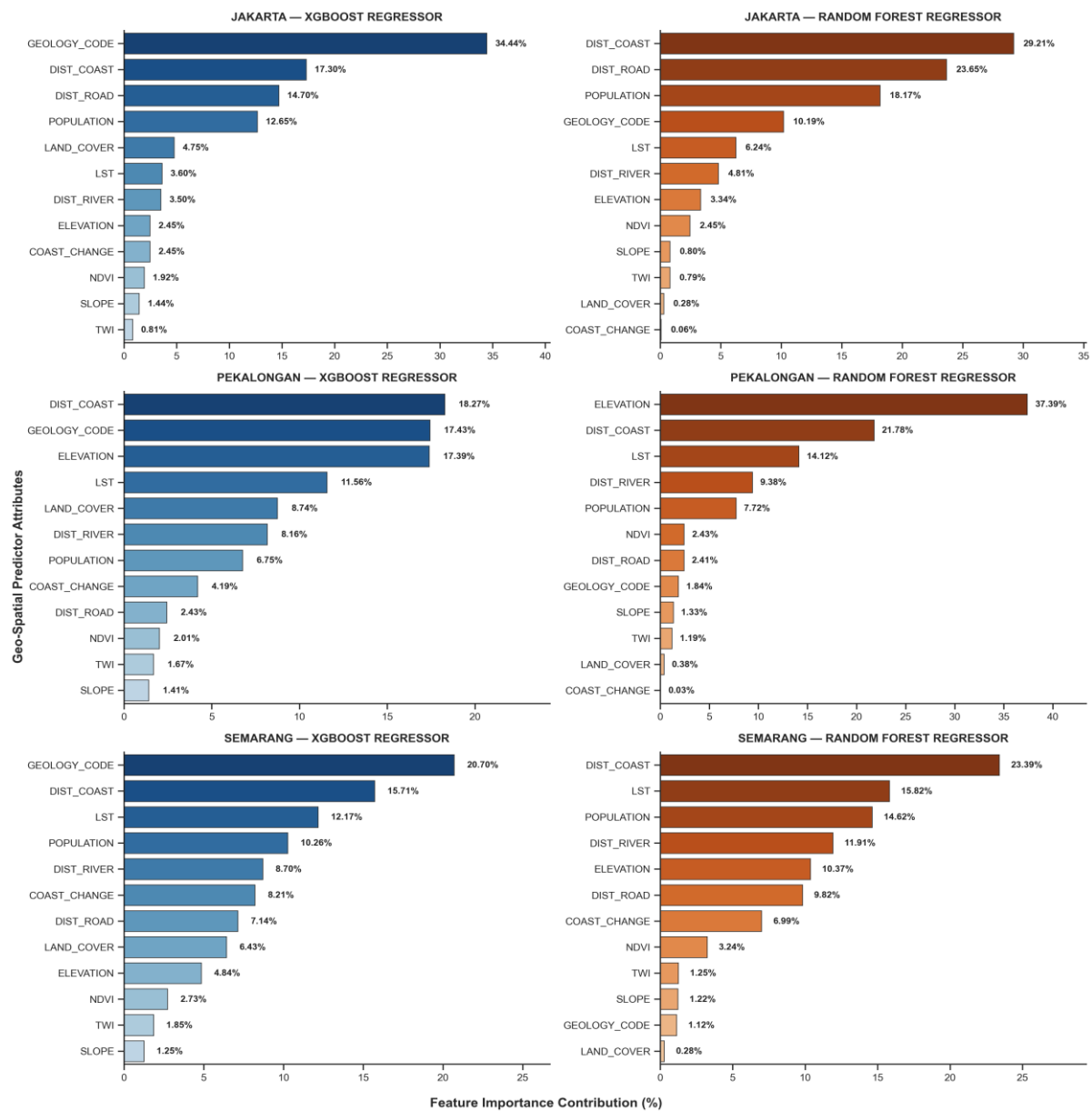


Figure 5. Relative feature importance distributions for XGBoost and Random Forest models across DKI Jakarta, Semarang City, and Kab. Pekalongan.

On the one hand, looking at it through geo-environmental lens, developed surfaces like asphalt concrete and industrial roofs have a low albedo and high heat capacity which artificially increases local LST values. On the ground, these high-temperature urban areas directly coincide with dense neighborhoods and clusters of heavy industries in the lower alluvial plains of Semarang and the flat coast of Pekalongan. Excessive groundwater exploitation is an important factor associated with land subsidence [16]. Widespread and unguided groundwater abstraction results in local reductions of pore water pressure around the semi-confined and confined aquifer systems, which lead to instantaneous volumetric compression of the overlying compressible silty clay and recent fluvial deposits [18]. As a result, feature importance rankings reflect statistical predictive associations within the ensemble models rather than direct physical causality [11]. While elevated LST values spatially coincide with accelerated subsidence zones across Semarang and Pekalongan, LST itself does not act as a direct physical mechanism for surface displacement. Instead, as discussed in previous subsidence and remote sensing literature [3, 19, 20], LST may serve as an indirect spatial proxy for densely developed urban and industrial environments, where impervious surfaces, structural concentration, and potentially higher localized groundwater demand may spatially co-occur. Therefore, the prominence of LST in the model outputs should be interpreted cautiously as an associative geo-environmental indicator rather than a direct causative driver [20].

3.4. Spatial Prediction Mapping and Inherent Model Limitations.

The final stage of the benchmarking framework consists of a spatial combination step that compares continuous spatial predictions from each trained model against geodetic PS-InSAR velocity observations. This spatialization workflow projects pixel-level model estimates back into continuous geographic coordinate space, generating a consistent hierarchy of three thematic layers for each study region: (1) Observed PS-InSAR Displacement Map, (2) Machine Learning Predicted Velocity Map, and (3) Spatial Residual (Error) Map. These spatial evaluation outputs are presented sequentially in Figures 6 a-c for DKI Jakarta, Figures 6 d-f for Semarang City, and Figures 6 g-i for Pekalongan Regency [20]. Both Random Forest and XGBoost accurately reconstructed the large-scale spatial patterns of land subsidence across the three coastal areas. Both ensemble models successfully captured high-velocity subsidence cores along the northwestern coastal corridors of Jakarta, the low-elevation alluvial industrial plains of northern Semarang, and the dynamic aquaculture-dominated coastal zone in Pekalongan. These spatial agreements demonstrate that tree-based ensemble architectures can reliably recover continuous deformation profiles by mapping target deformation vectors against the 12 multi-source geo-environmental and anthropogenic predictors. This continuous spatial reconstruction aligns with explainable artificial intelligence (XAI) findings that emphasize the spatial heterogeneity of urban deformation and the varying contributions of surface elevation and linear transportation infrastructure [20].

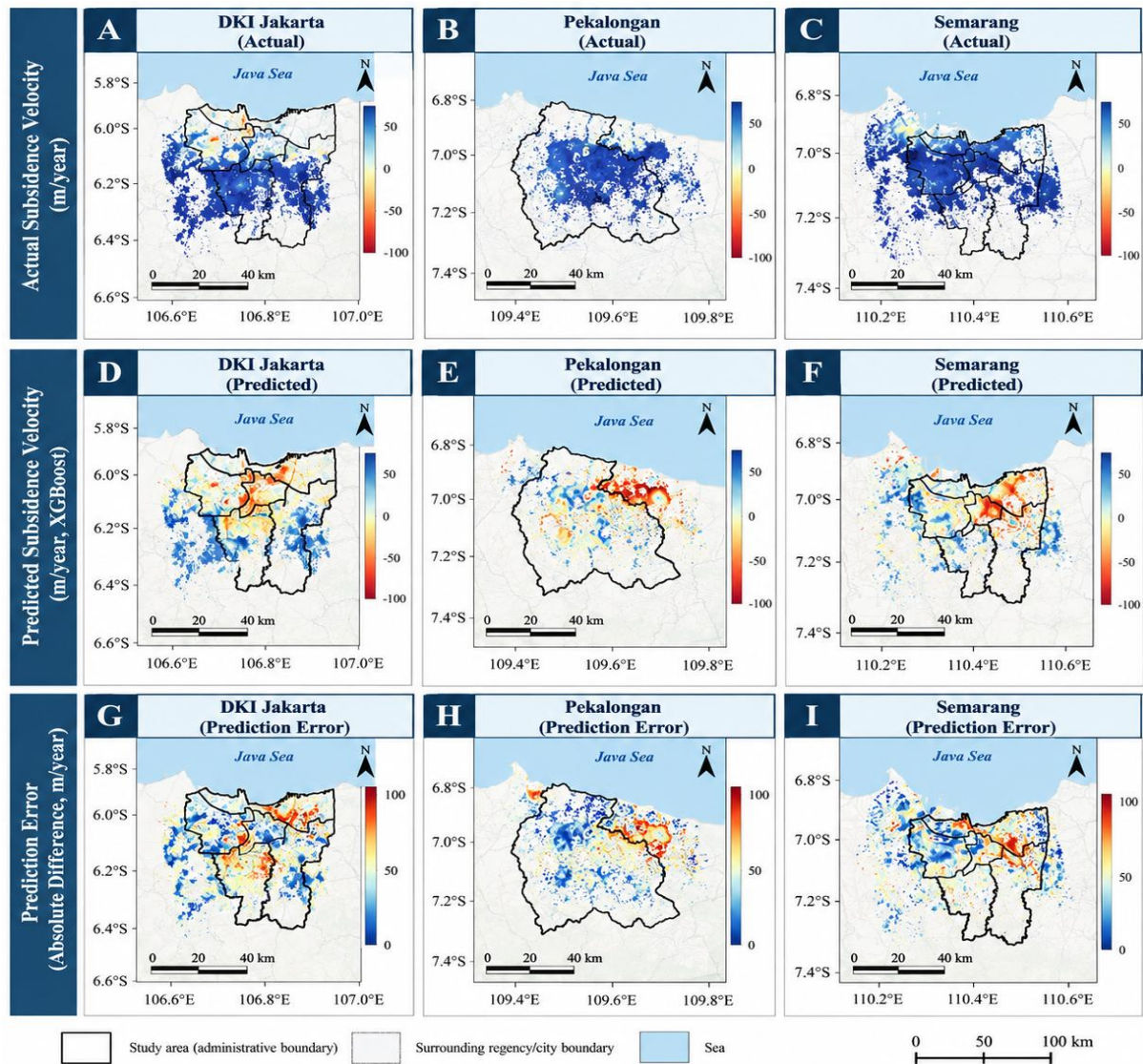


Figure 6. Actual and predicted InSAR land-subsidence velocities and prediction errors for DKI Jakarta (A, D, G), Pekalongan (B, E, H), and Semarang (C, F, I)..

To extract actionable spatial insights for risk mitigation, the continuous deformation grids were systematically intersected with administrative boundaries using Point-in-Polygon overlay. The resulting maximum subsidence hotspots at the sub-district (*kecamatan*) level are summarized in Table 3 for Jakarta, Semarang, and Pekalongan, respectively. These hotspot assessments provide spatial information that can support the identification of areas requiring closer monitoring and infrastructure risk assessment, consistent with previous studies highlighting the implications of land subsidence for infrastructure safety [21], and the relationship between groundwater extraction and land subsidence in the Jakarta Metropolitan Region [17].

Table 3. Top 5 Sub-districts (Kecamatan) in DKI Jakarta, Semarang and Pekalongan Ranked by InSAR Spatial Point Density

Sub-district (Kecamatan) Jakarta	Total persistence scatters (PS) Jakarta	Average actual VEL Jakarta	Average prediction Jakarta
Penjaringan	8.887	-11.36	-11.08
Kalideres	6.624	-9.66	-9.41
Cengkareng	5.597	-4.79	-5.79
Tanjung Priok	4.847	-4.59	-5.20
Kembangan	4.760	-9.72	-8.08
Sub-district (Kecamatan) Semarang	Total persistence scatters (PS) Semarang	Average actual VEL Semarang	Average prediction Semarang
Genuk	4.063	-70.25	-51.41
Pedurungan	3.245	-55.46	-49.74
Semarang Tengah	2.148	-20.07	-32.75
Semarang Utara	2.144	-37.33	-42.22
Semarang Timur	1.672	-40.95	-40.08
Sub-district (Kecamatan) Pekalongan	Total persistence scatters (PS) Pekalongan	Average actual VEL Pekalongan	Average prediction Pekalongan
Pekalongan Timur	2.306	-14.62	-53.80
Kedungwuni	1.181	-45.64	-40.84
Pekalongan Utara	758	-12.58	-48.49
Pekalongan Barat	569	-9.55	-45.71
Kajen	553	-52.83	-55.44

3.4.1. DKI Jakarta Hotspot Assessment.

A spatial point density filtering framework was instituted at the sub-district (Kecamatan) level to systematically isolate regional deformation macro-patterns from localized interferometric phase errors. The classification exploits the characteristic of generalization that is immune to localized spatial outliers by sorting its absolute density count of coherent Persistent Scatterers (PS) rather than only volatile maximum subsidence spikes sorted through (a). The actual displacement and prediction maps for DKI Jakarta are presented in Figure 6 (a-c), illustrating the spatial extent of alluvial compaction across the northwestern coastal corridor.

Based on the spatial validation results, Kecamatan Penjaringan is confirmed as the primary subsidence hotspot in the capital, anchored by a large network of 8,887 spatially-coherent InSAR points with significant regional subsidence activity at a baseline velocity mean of -11.36 mm/year and exhibiting significant subsidence conditions that saturates the bottom processing limit at -100.00 mm/year. By the sequential risk progression, both high-density urbanized corridors of Kecamatan Kalideres (6,624 PS points; $\mu = -9.66$ mm/year) and Kecamatan Cengkareng (5,597 PS points; $\mu = -4.79$ mm/year) represent adjacent high-risk deformation zones within the northwestern coastal corridor.

The concentration of these spatial clusters indicates that Jakarta's northwestern coastal plain is undergoing widespread regional-scale compaction. Geomechanical compaction is driven by coupling of static loads from the weight of heavy engineering and large spatial footprints of deep-aquifer groundwater depletion. The tuned spatial prediction mean from the optimized XGBoost model successfully reproduces the observed spatial subsidence pattern, yielding a spatial prediction mean of -11.08 mm/year for Penjaringan against an observed mean of -11.36 mm/year. This close agreement suggests that the model successfully captured the

dominant, large-scale subsidence pattern in this corridor. The corresponding spatial residual error distribution is illustrated in Figure 7 (a-b), confirming minimal prediction deviation across the high-density northwestern coastal corridor.

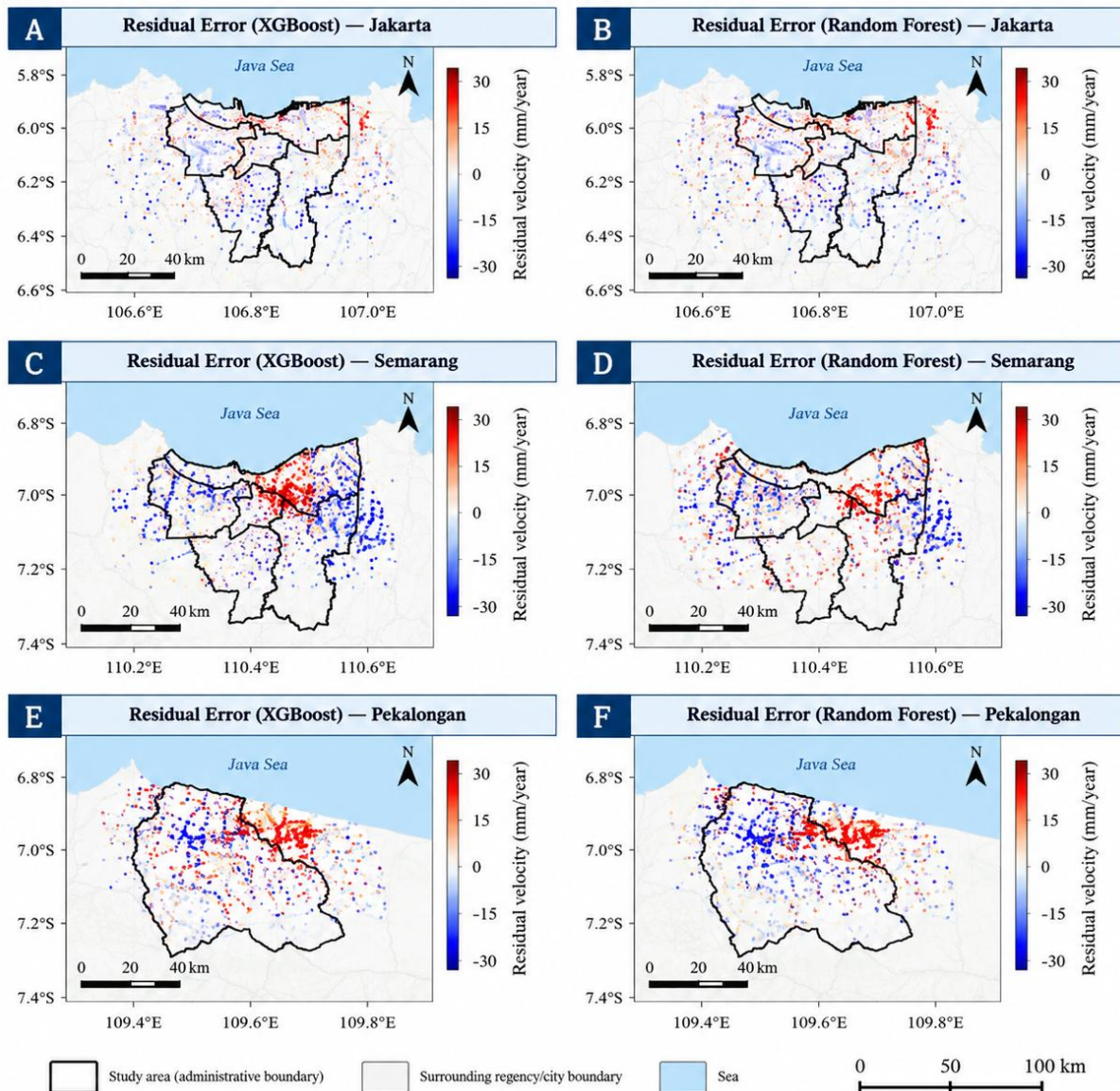


Figure 7. Spatial residual error distribution maps for Jakarta (A–B), Semarang (C–D), and Pekalongan (E–F) using XGBoost and Random Forest models.

3.4.2. City and Regency of Semarang Hotspot Assessment.

The spatial density analysis reveals two distinct subsidence mechanisms across the Central Java coastal zone. Kecamatan Genuk contains the largest concentration of Persistent Scatterer observations in the study area, covering almost the entire lowland maritime industrial plain, comprising 4,063 coherent Persistent Scatterer points. A negative baseline deformation average of -70.25 mm/year is a general estimate for Genuk, with localized maximum failures dropping sharply to -100.00 mm/year. The regional compaction, is strongly related to the density of the settlement regions and a historical consolidation of weak, highly compressible marine clay formations that was aggravated by cumulated infill through dense industrial structures and commercial groundwater abstraction. This coastal region is bordered by Kecamatan

Pedurungan (3,245 *PS* points; $\mu = -55.46$ mm/year and Semarang Tengah (2,148 *PS* points; $\mu = -20.07$ mm/year. Figure 6 (d-f) presents the actual displacement and prediction maps for Semarang City, capturing the sharp deformation contrast between the northern alluvial plain and the southern volcanic uplands.

In contrast, the density filter picks up an entirely different geomorphological hazard mechanism along Semarang polycyclic and rugged southern landscape, as it reveals high-density point distribution zones in Semarang Utara (2,144 points; $= -37.33$ mm/year) and Semarang Timur (1,672 points; $= -40.95$). The statistical convergence of high point density and elevated topographic slope vectors in these sub-districts suggests that the PS-InSAR sensors are capturing a composite deformation signal that includes not only subsurface hydrologic compaction but also active slope creep and gravitational mass wasting. This composite deformation signal may plausibly reflect broader tectonic and geomorphological activity associated with the Semarang–Ungaran upland margin; however, attributing this pattern specifically to active fault mechanisms would require dedicated structural geological investigation beyond the scope of this study.

Removing spatial leakage, the tuned XGBoost classifies effectively through these sharp topographic transitions with a mean prediction of -51.41 mm/year for Genuk and -42.22 mm/year for Semarang Utara, demonstrating strong predictive performance across contrasting geological environments. There were also prediction gaps at Kecamatan Semarang Barat (actual $\mu = -19.83$ mm/year; predicted $\mu = -37.73$ mm/year; $\Delta \approx 18$ mm/year) and Kecamatan Semarang Selatan (actual $\mu = -3.84$ mm/year; predicted $\mu = -30.96$ mm/year; $\Delta \approx 27$ mm/year). The two sub-districts lie in transitional zones between the high-velocity northern coastal plain and more geologically stable but southern uplands. Because the model was mostly trained on northern alluvial signals (as described previously), it carries a regional velocity bias that propagates into these lower-deformation transitional areas. This localized prediction discrepancy is particularly illustrated by Semarang Selatan, which has a near-stable observed velocity of -3.84 mm/year associated with a consolidated volcanic substrate but receives a model-predicted velocity of -30.96 mm/year because of high-deformation training clusters nearby. In the absence of sub-surface lithological covariates to resolve this abrupt geological boundary, the model was unable to draw a satisfactory distinction between these transitional deformation regimes and the pervasive coastal gradient, furthering the argument for geologically stratified training or geographically weighted ensemble architectures in future iterations. The spatial residual map visually captures these transitional prediction gaps, particularly the elevated error zones at Semarang Barat and Semarang Selatan.

3.4.3. Pekalongan Regency Hotspot Assessment.

For Pekalongan Regency, examining the geodetic density at spatial resolution clearly reveals a high density local bulge that transcribes the specific site of rural industrialization along the Pantura corridor. In this context, the spatial query separates Pekalongan Timur which is a high-density observation cluster with 2,306 coherent points ($\mu = -14.62$ mm/year, similar to one of the inlands manufacturing nucleus of Kecamatan Kedungwuni where a huge concentrated cluster of 1,181 persistent scatterer points has a serious regional downward velocity mean (-45.64 mm/year). This manufacturing belt with high density continues without interruption into Pekalongan Utara (758 *PS* points; $\mu = -12.58$ mm/year and Pekalongan Barat (569 *PS* points; $\mu = -9.55$ mm/year).

This spatial distribution coincides with the high feature importance scores assigned to anthropogenic and infrastructural features in Central Java. This sub-districts are characterized by high density of decentralized small-to-medium textile and batik industry. Uncontrolled high-potency groundwater abstractions from deep confined aquifers to support industrial washing and coloring procedures have led to a significant decline in pore water pressure. This hydrological imbalance causes rapid static mechanical compaction of the overlying weak fluvial sediment layers.

In this relatively homogeneous flat coastal plain, the tuned XGBoost model produced spatial predictions that closely tracked the observed deformation pattern $\mu_{\text{pred}} = -40.84$ mm/year at Kedungwuni and $\mu_{\text{pred}} = -53.80$ mm/year in Pekalongan Timur. This spatial agreement demonstrates the applicability of the optimized machine learning framework for representing anthropogenic environmental degradation in flat geomorphological profiles.

The highly spatial prediction anomaly for Pekalongan Regency is observed at Kecamatan Pekalongan Timur, where each of the optimized XGBoost models produced a mean velocity estimate of -53.80 mm/year compared to an InSAR-derived mean of only -14.62 mm/year, a residual gap of >39 mm/year. This pattern of overprediction likely reflects the interaction between local spatial heterogeneity and model behavior. The sub-district with the highest density in the dataset is Pekalongan Timur (2,306 PS points), but its real deformation speed is lower compared to other neighbouring kecamatan, like Kedungwuni (-45.64 mm/year) or Kajen (-52.83 mm/year). This spatial contrast introduces a local failure mode into the training matrix, where the model must learn how to classify a spatially concentrated patch of mid-range speed scatter points surrounded by large patches of regional high speed signatures. This spatial contrast may plausibly present challenging conditions for a globally-fitted model, potentially contributing to the observed overestimation at this moderate-velocity anchor point. Such behaviour aligns with known limitations of single global boosting architectures in geospatial domains characterized by sharp local heterogeneity within an otherwise homogeneous coastal plain, where spatial non-stationarity violates the implicit stationarity assumption of a single global model. Further, the lack of confident knowledge about key sub-surface hydrogeological covariates in particular shallow aquifer piezometric head data limits the models capacity to discern between areas of moderate extraction pressure and high-extraction industrial clusters which have similar surface proxy signatures. The prediction gap highlights the suggested acquisition of groundwater monitoring telemetry, and Geographically Weighted ensemble architectures to better capture localised deformation anomalies in future work [3]. The spatial residual map Figure 7 (e-f) highlights the concentrated overprediction anomaly at Pekalongan Timur relative to the more accurate predictions across the inland manufacturing belt.

These substantial localized prediction residuals most notably the overprediction in Pekalongan Timur (residual gap of approximately 39 mm/year) and Semarang Selatan (residual gap of approximately 27 mm/year) highlight important operational considerations when applying global tree-based ensemble models to practical subsidence-risk assessment. These discrepancies may reflect the inherent difficulty of global models in resolving sharp localized spatial heterogeneity and non-stationary deformation regimes within broad regional domains. Furthermore, the absence of direct subsurface hydrogeological variables in the current predictor matrix, such as localized piezometric head telemetry or detailed geotechnical borehole profiles, may limit the models' capacity to distinguish between subtle transitional regimes and pervasive regional deformation gradients [3].

In terms of practical application, such prediction discrepancies carry meaningful implications for coastal spatial planning and infrastructure management. Overestimating deformation rates in moderately subsiding or relatively stable zones could lead to overly conservative hazard zoning and misdirected adaptation resources. Conversely, underestimating rates in vulnerable transitional corridors could risk overlooking emerging structural stresses on critical infrastructure assets. Therefore, while the proposed ensemble models serve as effective, reproducible tools for regional-scale hazard screening and macro-level baseline assessment, fine-scale municipal planning and site-specific engineering designs should complement these regional predictions with localized InSAR monitoring, in-situ groundwater-level observations, and spatially adaptive modeling frameworks [19].

4. Conclusion

This Study present a modular geospatial data engineering and multi-model machine-learning framework based on Random Forest (RF) and Extreme Gradient Boosting (XGBoost) framework to predict land-subsidence patterns and assess the importance of geo-environmental predictors of land subsidence across three critical Pantura coastal hotspots, DKI Jakarta, Semarang City, and Pekalongan Regency . We propose a pipeline that integrates high-density PS-InSAR geodetic observations from 2021–2022, concurrent twelve geo-environmental predictor variables subjected to a statistical filtering procedure, and evaluates model performance using an independent 20% holdout testing subset. Both ensemble models demonstrated strong predictive capabilities across the surveyed coastal hotspots. Random Forest achieved superior accuracy in the more urbanized and topographically complex environments of DKI Jakarta ($R^2 = 0.8633$) and Semarang City ($R^2 = 0.8090$). In contrast, in the relatively uniform coastal plain of Pekalongan Regency , both architectures converged to comparable predictive performance, with regularized XGBoost ($R^2 = 0.8017$) reaching near-parity with the Random Forest benchmark ($R^2 = 0.8017$ baseline, 0.8006 tuned). As shown in Table 1, XGBoost performance improved considerably with hyperparameter optimization across all three study areas, most notably in Semarang City, where the regularized boosting approach provided substantial predictive gains for complex geospatial prediction tasks. Feature-importance analysis showed distinct spatial patterns across the three regions with respect to feature importance. For DKI Jakarta, GEOLOGY_CODE represented 34.44% of the XGBoost importance distribution, indicating a high predictive importance of this variable, in contrast with earlier Random Forest models, where DIST_COAST (29.21%) and DIST_ROAD (23.65%) received the highest importance as spatial proxies for infrastructure-induced gradients in loading efficiency. Land surface temperature at Semarang City was still among the dominant predictors for both models (RF: 15.82%, XGBoost: 12.17%), consistently suggesting that thermally intensive urban and industrial environments are spatially associated with areas of greater subsidence intensity across the northern alluvial coastal plain. in Pekalongan Regency, the dominant predictors for Random Forest were ELEVATION (37.39%) and DIST_COAST (21.78%), while XGBoost (XGB) included DIST_COAST (18.27%), GEOLOGY_CODE (17.43%), and ELEVATION (17.39%); consistent with the low-relief topography of the district. Given the typically flat coastal terrain of Pekalongan, the strong importance of ELEVATION in the Random Forest model may reflect the model's sensitivity to minor elevation gradients as proxies for shallow aquifer depth and compressible sediment thickness, small differences in ground height correspond to relatively larger changes in

susceptibility difference. This prominence of ELEVATION and DIST_COAST across both ensemble models indicates that geomorphological characteristics are strongly represented in the model predictions, with natural landscape susceptibility interacting with anthropogenic pressures as evidenced by deformation principally concentrating within very low-lying coastal sectors. Spatial hotspot analyses through point-in-polygon zonal aggregation identified Kecamatan Penjaringan is the primary subsidence hotspot in DKI Jakarta with 8,887 coherent PS points and a mean deformation velocity of -11.36 mm/year, with localized extremes reaching -100.00 mm/year. In Semarang City, the highest-density concentration (4,063 PS points; $\mu = -70.25$ mm/year) was located in Kecamatan Genuk where the observed deformation pattern is consistent with marine clay compressibility and industrial activity, In Pekalongan Regency, Kecamatan Kedungwuni (1,181 PS points; $\mu = -45.64$ mm/year) and Kecamatan Wonopringgo ($\mu = -57.55$ mm/year) represented the leading inland deformation concentrations associated with decentralized batik and textile-processing areas. Future research should emphasize the incorporation of shallow aquifer piezometric monitoring data to better capture subsurface hydraulic dynamics that are currently lacking in the predictor matrix. Prolonging the InSAR observation period and implementing spatial block cross-validation would further enhance the robustness of prediction evaluation in transitional deformation zones. This study presents a replicable, data-driven methodology that transforms raw satellite geodetic data into regional-scale subsidence information, providing a scalable framework for proactive spatial planning and sustainable aquifer management across the Pantura corridor.

Acknowledgements

The authors would like to express their sincere gratitude to all colleagues and research assistants for their invaluable contributions, unwavering support, and the resources provided throughout this study.

Author Contribution

Luthfy Azh-Zhaviery Puary: Conceptualization, Methodology, Visualization, Data Analysis, Writing; Rendi Handika: Data Collection, Methodology; Joko Widodo: Supervision; Panca Dewi Pamungkasari: Supervision; Endah Tri Esti Handayani: Supervision; Babag Purbantoro: Supervision; Damy Matheus Heezen: Supervision.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.53623/gisa.v6i2.1233>.

References

- [1] Ohenhen, L.O.; et al. (2026). Land subsidence on Java Island and its contributions to relative sea level change. *Science Advances*. <https://doi.org/10.1126/sciadv.aec0172>.
- [2] Yavuz Ozalp, A.; Akinci, H.; Zeybek, M. (2023). Comparative analysis of tree-based ensemble learning algorithms for landslide susceptibility mapping: A case study in Rize, Turkey. *Water*, 15(14). <https://doi.org/10.3390/w15142661>.
- [3] Hakim, W.L.; Fadhillah, M.F.; Lee, K.J.; Lee, S.J.; Chae, S.H.; Lee, C.W. (2023). Land subsidence and groundwater storage assessment using ICOPS, GRACE, and susceptibility mapping in

- Pekalongan, Indonesia. *IEEE Transactions on Geoscience and Remote Sensing*, 61. <https://doi.org/10.1109/TGRS.2023.3324043>.
- [4] Yuwono, B.D.; Sabri, L.M.; Wijaya, A.P.; Awaluddin, M. (2024). Assessment of flood risk induced by land subsidence using machine learning. *Indonesian Journal of Geography*, 56(3), 376–386. <https://doi.org/10.22146/ijg.94726>.
- [5] Setyaningrum, N.; et al. (2026). SBAS-InSAR monitoring and time-lag correlation analysis of peatland subsidence in Central Kalimantan, Indonesia. *Canadian Journal of Remote Sensing*, 52(1). <https://doi.org/10.1080/07038992.2026.2663665>.
- [6] Alesheikh, A.A.; Chatrsimab, Z.; Rezaie, F.; Lee, S.; Jafari, A.; Panahi, M. (2024). Land subsidence susceptibility mapping based on InSAR and a hybrid machine learning approach. *Egyptian Journal of Remote Sensing and Space Science*, 27(2), 255–267. <https://doi.org/10.1016/j.ejrs.2024.03.004>.
- [7] Hosseinzadeh, E.; Anamaghi, S.; Behboudian, M.; Kalantari, Z. (2024). Evaluating machine learning-based approaches in land subsidence susceptibility mapping. *Land*, 13(3). <https://doi.org/10.3390/land13030322>.
- [8] Eghrari, Z.; Delavar, M.R.; Zare, M.; Beitollahi, A.; Nazari, B. (2023). Land subsidence susceptibility mapping using machine learning algorithms. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4/W1-2022, 129–136. <https://doi.org/10.5194/isprs-annals-X-4-W1-2022-129-2023>.
- [9] Huynh, N.D.Q.; Nguyen, G.T.; Ha, T.K.; Le, D.T.; Tran, V.A. (2025). Machine learning approaches for predicting land subsidence in Ca Mau: XGBoost, Random Forest, and MAF. *Inzynieria Mineralna*, 1(2), 703–714. <https://doi.org/10.29227/IM-2025-02-58>.
- [10] Azarm, Z.; Mehrabi, H.; Nadi, S. (2025). Enhanced land subsidence interpolation through a hybrid deep convolutional neural network and InSAR time series. *Geoscientific Model Development*, 18(19), 6903–6919. <https://doi.org/10.5194/gmd-18-6903-2025>.
- [11] Yaragunda, V.R.; Vaka, D.S.; Oikonomou, E. (2025). Land subsidence susceptibility modelling in Attica, Greece: A machine learning approach using InSAR and geospatial data. *Earth*, 6(3). <https://doi.org/10.3390/earth6030061>.
- [12] Tasan, M.; Ghorbaninasab, Z.; Haji-Aghajany, S.; Ghiasvand, A. (2023). Leveraging GNSS tropospheric products for machine learning-based land subsidence prediction. [*Journal name not stated in source*]. <https://doi.org/10.1007/s12145-023-01143-z>.
- [13] Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Razavi, S.; Choi, S.M. (2024). Enhancing flood-prone area mapping: fine-tuning the K-nearest neighbors (KNN) algorithm for spatial modelling. *International Journal of Digital Earth*, 17(1). <https://doi.org/10.1080/17538947.2024.2311325>.
- [14] Sankhyan, S.; Sharma, S.; Pahal, S.; Uday, K.V.; Dutt, V. (2026). SFEL: A machine learning framework for forecasting radar backscatter based ground deformation. *Scientific Reports*, 16(1). <https://doi.org/10.1038/s41598-026-37619-3>.
- [15] Sabdu Busoroh, L.; Panuntun, H. (2025). Monitoring of vertical ground surface movement in the coastal area of Cilacap Regency using the Small Baseline Subset Interferometry SAR (SBAS InSAR) method. *Elipsoida: Jurnal Geodesi dan Geomatika*. <https://doi.org/10.14710/elipsoida.2025.28688>.
- [16] Handika, R.; Widodo, J.; Pravitasari, A.E. (2024). Combined land subsidence analysis in Jakarta based on Ps-InSAR and MICMAC methods (Kombinasi analisis penurunan muka tanah di Jakarta berdasarkan metode Ps-InSAR dan MICMAC). *Jurnal Teknologi Lingkungan*, 25(1), 137–145. <https://doi.org/10.55981/jtl.2024.1125>.
- [17] Hakim, W.L.; Fadhillah, M.F.; Won, J.S.; Park, Y.C.; Lee, C.W. (2025). Advanced time-series InSAR analysis to estimate surface deformation and utilization of hybrid deep learning for susceptibility mapping in the Jakarta metropolitan region. *GIScience & Remote Sensing*, 62(1). <https://doi.org/10.1080/15481603.2025.2465349>.

- [18] Widodo, J.; et al. (2025). Technical and policy analysis: Time series of land subsidence for the evaluation of the Jakarta groundwater-free zone. *Urban Science*, 9(3). <https://doi.org/10.3390/urbansci9030067>.
- [19] Shi, M.; Wang, X.; Gu, C.; Gao, M.; Zhou, C.; Gong, H. (2026). Insights into spatial heterogeneity of land subsidence susceptibility using InSAR and explainable machine learning. *Remote Sensing*, 18(9). <https://doi.org/10.3390/rs18091298>.
- [20] Widodo, J.; et al. (2024). Dynamic geo-visualization of urban land subsidence and land cover data using PS-InSAR and Google Earth Engine (GEE) for spatial planning assessment. *Urban Science*, 8(4). <https://doi.org/10.3390/urbansci8040234>.
- [21] Liu, X.; et al. (2025). Analysis of ground subsidence evolution characteristics and attribution along the Beijing–Xiong'an intercity railway with time-series InSAR and explainable machine-learning technique. *Land*, 14(2). <https://doi.org/10.3390/land14020364>.



© 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).