Green Intelligent Systems and Applications

Research Article
Volume 5(1), 2025, 106-122
https://doi.org/10.53623/gisa.v5i1.606

Land Subsidence Analysis Using Machine Learning Algorithm Random Forest Method in DKI Jakarta

Camelia Nur Hidayah[1], Panca Dewi Pamungkasari[1], Sari Ningsih[1]*, Muhammad Fauzan Azhiman[1], Joko Widodo[2], Elfady Satya Widayaka[3]

[1]Universitas Nasional Faculty of Communication and Information Technology, Jakarta, Indonesia
[2]Research and Innovation Agency (BRIN), Jakarta 10340, Indonesia
[3]Senior Software engineer PT Motiv-research, Japan

*Correspondence: sari.ningsih@civitas.unas.ac.id

SUBMITTED: 19 February 2025; REVISED: 24 Marh 2025; ACCEPTED: 26 March 2025

ABSTRACT: Land subsidence is an environmental phenomenon that causes the earth's surface to decline gradually or suddenly. Land subsidence occurred in DKI Jakarta due to various factors such as excessive groundwater exploitation, infrastructure loads, and geological conditions. The purpose of this study was to analyze land subsidence in DKI Jakarta and the distribution of existing land subsidence. The results were compared with previous findings using PS-InSAR. Land subsidence was predicted using the Random Forest algorithm. Random Forest, as a type of machine learning, was able to reduce noise and minimize the impact of overfitting through ensemble techniques. Researchers used four metrics, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2, and Kling-Gupta Efficiency (KGE), to assess the accuracy of the algorithm. The analysis results of land subsidence in DKI Jakarta using Random Forest aligned with the PS-InSAR method. It was observed that areas experiencing land subsidence were predominantly in North and West Jakarta compared to other regions. Furthermore, the prediction of land subsidence using the 2017–2021 dataset indicated a decrease of up to -60 mm/year.

KEYWORDS: DKI Jakarta; land subsidence; PS-InSAR; random forest

1. Introduction

Land subsidence was a major hazard to land surface stability in DKI Jakarta. It occurred in several areas of DKI Jakarta, especially in Pantai Indah Kapuk, Marunda, Ancol (North Jakarta), and Kembangan (West Jakarta). On the other hand, slight land subsidence was observed in the Kalibaru area, Central Jakarta. Land subsidence was an environmental phenomenon that caused the earth's surface to decrease gradually or suddenly [1]. Land subsidence occurred in DKI Jakarta due to various factors such as excessive groundwater exploitation, infrastructure loads, and geological conditions [2]. For land subsidence monitoring, the widely used Persistent Scatterer Interferometric Synthetic Aperture Radar (PS-InSAR) technology was employed. PS-InSAR detected land surface changes through multi-temporal radar data analysis [2]. While accurate in historical analysis, PS-InSAR had limitations in predicting and could not model future land subsidence patterns. To overcome

these limitations, machine learning approaches such as Random Forest were used to predict subsidence patterns based on multivariate data.

Random Forest, as a type of machine learning, was able to reduce noise and minimize the impact of overfitting through ensemble techniques. Previous research compared the Random Forest algorithm with LightGBM and XGBoost. The results showed that Random Forest performed better than LightGBM and XGBoost, with values of (R^2 = 0.84), (KGE = 0.89), (RMSE = 2.19), and (MAE = 1.42) [3]. In Handika et al.'s research, the Random Forest model demonstrated better performance, achieving a value of 0.73 compared to SVM. However, Random Forest had shortcomings—poor parameter tuning increased bias, so researchers addressed this issue with hyperparameter tuning [6]. Researchers used four matrices to assess the accuracy of the algorithm. RMSE calculated the square root of the mean square error, providing an overview of prediction error by assigning more weight to larger errors. The Coefficient of Determination (R^2) measured how well the model explained the variability of the data. The value of R^2 ranged between 0 and 1, with values close to 1 indicating a good model. KGE was designed to overcome weaknesses in traditional evaluation (R^2), considering correlation, relative bias, and variance ratio. The use of these hyperparameters adjusted the decision tree (n_estimators, max_features, and min_samples_leaf) by using worker timeouts (joblib) to ensure workers had enough time, and threading was implemented to optimize CPU usage. This research had the following objectives: analyzing land subsidence in DKI Jakarta, outlining the distribution of existing land subsidence, the cumulative land subsidence for the period 2017–2021 on a year-to-year basis, and reclamation areas with alluvial soil characteristics in DKI Jakarta. The three results of the analysis using Random Forest were compared with previous results using PS-InSAR. A new development was introduced—a prediction for 2022 with validation data from previous studies..

2. Materials and Methods

2.1.Random forest.

Random Forest was an ensemble-based algorithm designed to handle large-scale and complex data by combining multiple independent decision trees to improve prediction accuracy [3]. In research [3] on the study of land subsidence in the Bangkok vicinity, Random Forest was able to model the non-linear relationship between driving factors and land subsidence. The study revealed that subsidence was increasing by more than -9.0 mm/year over the next few decades. The results showed that Random Forest achieved values of R^2 = 0.84, KGE = 0.89, RMSE = 2.19, and MAE = 1.42. This algorithm operated on the principle of bootstrap aggregating (Bagging), where each original dataset was randomly sampled with replacement to form several subsets of data [2]. Random Forest was a method derived from the original bootstrap aggregation (Bagging) algorithm proposed by Breiman in 1996, which aimed at creating a set of trees. In Bagging, a dataset was sampled as bags that matched the size of the original dataset. The predictions from these trees were combined across the bags by averaging (regression) or polling (classification) to produce the final prediction [3].

Random Forest applied random feature selection to each split in the decision tree. This process, known as feature randomness, ensured that only a subset of features was randomly selected each time the tree made a split decision. By doing so, the correlation between the trees in the forest was reduced, making the model more stable and less dependent on specific

features. This approach helped mitigate overfitting compared to single decision tree methods. In regression with Random Forest, the final prediction was calculated as the average of all tree predictions, which was formulated as:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i \quad (1)$$

Where $\hat{y}$ is final prediction, N is number of trees in random forest, and $\hat{y}_i$ is the prediction of the tree to-i.

Formula (1) calculated regression in Random Forest by predicting the final value y^\hat{y}y^ as the average of all predictions made by Random Forest trees. In this research, the final value y^\hat{y}y^ represented the prediction of land subsidence. In the context of subsidence modeling, Random Forest was used to map vulnerability areas based on factors that influenced subsidence distribution. In research [8], Random Forest achieved an accuracy of more than 80% in predicting vulnerable areas, while research [3] utilized the Random Forest algorithm to model non-linear relationships between driving factors and land subsidence.

Random Forest demonstrated robustness against overfitting, had the ability to handle missing values, and showed good model performance on large datasets. However, parameter tuning posed a major challenge in the application of this algorithm. If the parameter selection was not optimal, the model could suffer from high bias or excessive complexity. Therefore, in this study, hyperparameter adjustment was performed by reducing the number of trees (n_estimators) and limiting the maximum number of features (max_features) for division [3]. The construction of the Random Forest model was built by setting the number of trees (n_estimators) and maximum features (max_features) for each node. Once the data was cleaned, the dataset was split, with 80% used for training and 20% for testing. The model was then trained using the training data, and the predicted results were compared to the actual values using the following evaluation metrics:

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (4)$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (5)$$

Where n is total number of data points, y_i is actual value, $\hat{y}_i$ is predicted value of the model, r is correlation between predicted and actual values, α is ratio between predicted and actual standard deviation, and β is Ratio of average predicted to actual values.

The above formula was used to evaluate the Random Forest prediction metrics and assess the accuracy of the model. The MAE metric (2) measured the average absolute error between the actual value (yiy_iyi) and the predicted value (y^i\hat{y}_iy^i). If the predicted subsidence value was small, it indicated that the model produced accurate predictions. The RMSE metric (3) assigned greater weight to large errors because it squared the differences, making it useful for detecting outliers or significant errors. The R^2 metric (4) measured how well the model

explained the variance in the target data. Its values ranged between 0 and 1, with 1 indicating a highly accurate model and 0 indicating that the model could not explain the variability in the data. The KGE metric (5) provided an overall assessment of model performance by considering correlation, bias, and variability. The KGE value ranged from $-\infty$ to 1, where a value of 1 indicated good prediction performance. After metric evaluation, if the model showed high bias or suboptimal performance, hyperparameter tuning was performed using RandomSearchCV and GridSearchCV. These two methods were used to find the best combination of parameters to improve the model by minimizing prediction error. If the evaluation results after tuning showed an increase in performance and model stability on test data, the model was considered ready for further analysis.

2.2.Land subsidence.

Land subsidence was an environmental issue of global concern. It could be caused by a combination of factors such as the natural compaction of sediments/rocks, loading, excessive groundwater extraction, and tectonic activity [7]. Previous research provided information that the scientist Gambolati found land subsidence was caused by fluid withdrawal. Additionally, research conducted in 290 major cities worldwide by Bagheri-Gavkoch showed that land subsidence was mostly caused by human activities (76.92%), with half (50.75%) attributed to water extraction [2].

2.3.Research location.

The location of this research was the DKI Jakarta City area. DKI Jakarta was the capital city of Indonesia, situated on the coast of Java Island, with a total area of 662.33 km². Jakarta was a lowland area above sea level and the only city in Indonesia with a status equivalent to a province. The study area covered West Jakarta, East Jakarta, South Jakarta, North Jakarta, and Central Jakarta. Jakarta's physiography was classified within the Coastal Plain zone.

2.4.Research focus.

This research focused on comparing the results of the machine learning-based Random Forest algorithm with PS-InSAR. The objective was to analyze and compare the performance of these two algorithms in assessing land subsidence in DKI Jakarta. Table 1 was divided into two sections comparing research algorithms. The advantages of each algorithm were highlighted with a white mark, while the disadvantages were indicated with an orange mark. Researchers identified the strengths and weaknesses of the methods under study by referring to previous research on land subsidence that utilized various machine learning algorithms. This comparison was based on journal references listed in the bibliography, each marked with a corresponding number. After comparing the three algorithms, the researchers proposed the Random Forest method for land subsidence analysis due to its advantages in handling noise, preventing overfitting, ensuring data accuracy, improving model performance, and enhancing model sensitivity. However, the Random Forest model had shortcomings, particularly in parameter tuning, which could lead to bias in data processing. To prevent tuning parameters from causing bias, tree reduction was performed using RandomSearchCV and GridSearchCV hyperparameter tuning by adjusting n_estimatorsn_estimatorsn_estimators,

max_featuresmax_featuresmax_features, and min_samples_leafmin_samples_leafmin_samples_leaf [2, 10].

Table 1. Algorithm comparison.

PARAMETER	RF	SVM	BLR
Noise	The ensemble technique can reduce noise [3, 5].	Moderate noise tolerance with RBF Kernel [6]	Susceptible to noise [4]
Overfitting	Ensemble techniques can reduce the impact of overfitting individual trees[3, 5].	Overfitting on small data if parameters are not well tuned [4, 6].	Minimal risk of overfitting, model linearity limits complexity [4].
Parameter Tunning	Poor parameter tuning can increase bias [5].	Tuning must be done carefully [6].	Excellent for linear data [4]
Data Accuracy	Excellent data accuracy on non-linear, linear is quite good [3, 5].	Less than optimal on linear data [6]	Efficient for small datasets [4]
Mode Performance	Able to handle large and complex data [3, 5].	High model performance for non-linear data with optimal tuning [3, 6].	Performance drops on complex datasets[4]
Model Sensitivity	Low to Hypeparameter [3, 5].	Very sensitive to C and Gamma parameters [6]	Less sensitive to C parameters and intensive tunics[4]

2.5. Data source.

The sample data used in this study was obtained from a journal written by Joko Widodo et al., entitled "Aperture Radar Method of Jakarta City Region Using TerraSAR-X Spaceborne Data." The dataset contained 91,987 records, which were divided into training and testing data. The available data included information on ID, Latitude (Lat), Longitude (Lon), Height, Height Wrt Dem, Sigma Height, Velocity (Vel), Sigma Velocity (Sigma Vel), Seasonal, Cumulative Displacement (Cumul.Disp), Coherence (Coher), Svet, Lvet, Initial (In), Final (Fin), Standard Deviation (Stdev), and Temporal Data. Based on these attributes, the independent variables consisted of Lat, Lon, Height, Height Wrt Dem, Sigma Height, Sigma Vel, Seasonal, Svet, Lvet, and Temporal Data, while the dependent variable was Cumul.Disp.

2.6.Research stage.

In the research method, several stages were carried out sequentially, as shown in Figure 1. In this study, the dataset consisted of two variables based on land subsidence parameters. The dataset contained 91,987 records, which were divided for model training and testing. The next stage was data preprocessing, which was necessary to ensure that the data was of high quality, consistent, and stable for processing by machine learning algorithms. The preprocessing stage included checking for data consistency, missing values, outliers, and duplicate records. After preprocessing, Exploratory Data Analysis (EDA) was conducted to examine the relationships between variables. EDA helped determine the feature and target variables related to land subsidence prediction. The target variable was Cumul.Disp, which was referenced as y = df[target]. The feature variables were defined as X = df[features], which included Height, Height Wrt Dem, Vel, Sigma Height, Sigma Vel, Coher, Stdev, Lvet, Svet, Seasonal, and Temporal_Mean. The research focus was on Cumul.Disp, which represented the cumulative land subsidence predicted using the associated features.

The dataset was divided using the random sampling method with a proportion of 80:20, where 80% of the data (73,589 records) was used for training, and 20% (18,398 records) was

used for testing the model's performance. At the modeling stage, researchers implemented a machine learning algorithm, namely Random Forest. The Random Forest model was built using Scikit-learn with random_state=42 to ensure reproducibility and n_estimators=100 to maintain model stability, even though it resulted in a longer computation time. After building the model, an evaluation matrix was applied to assess its performance. The evaluation metrics included RMSE, R^2, MAE, and KGE. If the R^2 value was close to 1 and MAE/RMSE was low, the model was considered to have good performance. Additionally, if KGE was close to 1, it indicated optimal model efficiency. If the initial metric evaluation revealed high bias or suboptimal performance, hyperparameter tuning was performed using RandomSearchCV and GridSearchCV. These methods focused on reducing the number of trees while utilizing joblib to manage computation time efficiently. The objective was to find the best combination of parameters to improve the model by minimizing prediction error. After hyperparameter tuning, metric evaluation was conducted again, followed by visualization of the land subsidence prediction results. Finally, the results were compared between the Random Forest model and PS-InSAR, and explanations were provided regarding the findings.

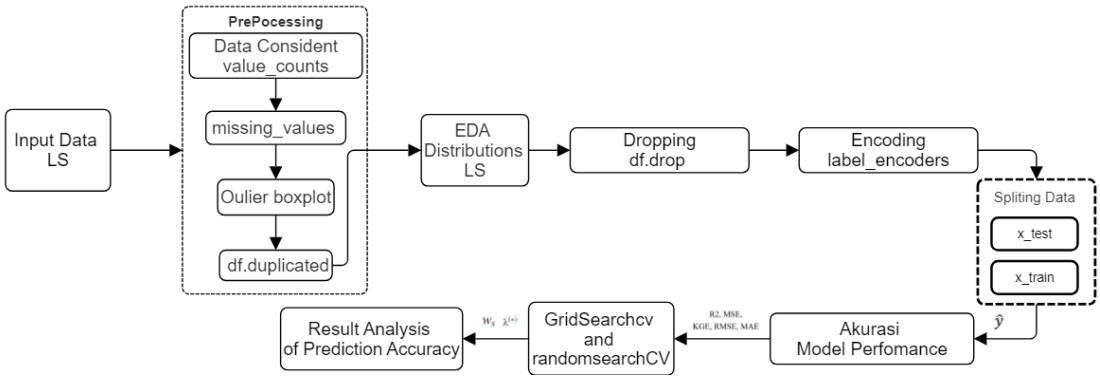

Figure 1. Research stages.

3. Results and Discussion

3.1. Data prepocessing.

In this preprocessing stage, the process of data cleansing was carried out to ensure that the dataset was free from missing values, duplicates, and outliers. Although the data from the original source was available, it needed to be thoroughly checked before use. The purpose of preprocessing was to improve data quality and enhance the accuracy and reliability of the analysis results. Therefore, preprocessing was performed to ensure that the data was clean, consistent, and ready for analysis (Table 2).

Table 2. Stages of preprocessing.

	Prepocessing Stage
Duplicate	There are the same columns in the temporal column, namely HH and HH1, both of these columns have temporal values removed one of them for accurate analysis results.
Missing Value	There are no missing values even rechecking the missing values that are not visible.
Outlier	Outliers that have been cleaned using the Interquartile Range method which takes the difference between the third quartile and the first quartile. Data below and above the upper limit are considered outliers. To keep the data distribution representative, outlier values were replaced with the media value of the attribute in question. This approach was chosen because the media is more robust to outliers than the mean.

	However, some outliers remain visible in the visualization as this method does not completely remove extreme values but replaces them with the median of the feature in question to maintain the integrity of the data distribution.
Consident Data	Consistency checks were performed on the numerical data. Data consistency was verified by ensuring the absence of double values, unrealistic numerical ranges, and inconsistencies in units of measurement.

3.2.Ground subsidence analysis.

Exploratory Data Analysis (EDA) was conducted as an initial analysis of the dataset to better understand its key characteristics, patterns, and relationships before proceeding with further modeling. This EDA helped identify the datasets that were analyzed further, as explained below:

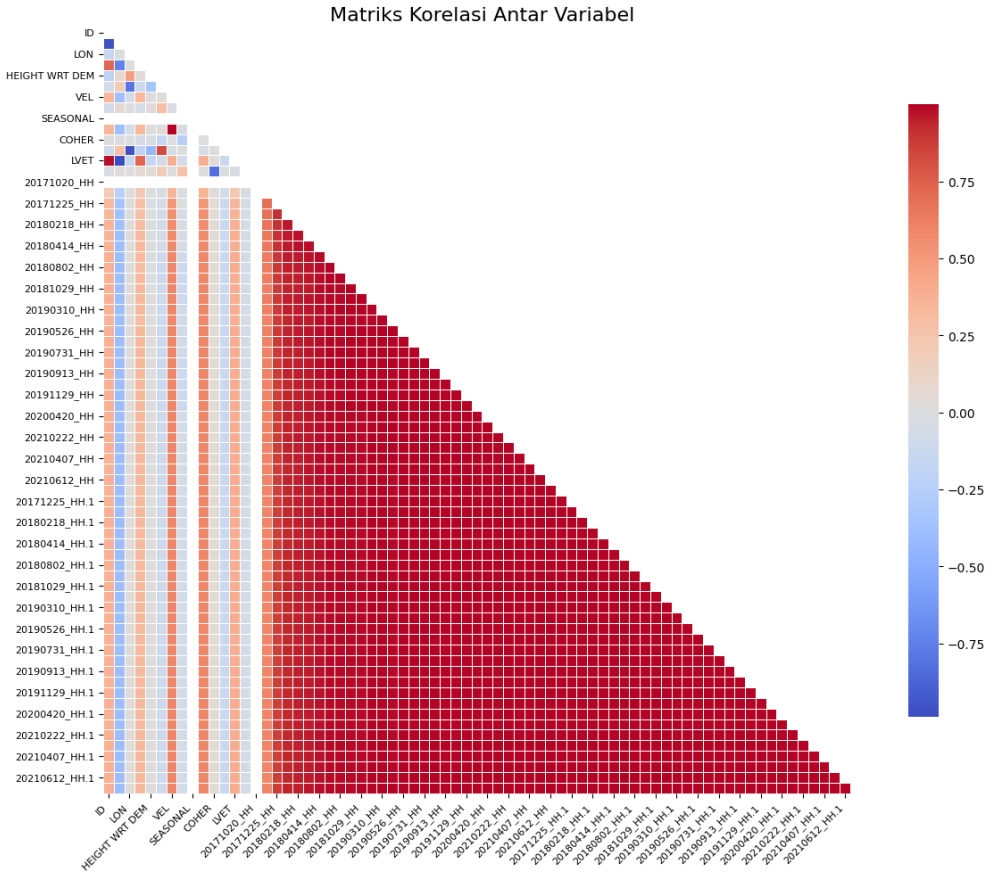

Figure 2. Relationship between variables.

Figure 2 shows that the relationship between temporal variables was very strong, as indicated by a solid red color. This suggested that the temporal measurement values had a consistent pattern, making temporal variables effective for time trend analysis. The variables VEL, HEIGHT, and CUMUL.DISP exhibited significant positive correlations with some temporal variables, indicating a strong relationship between soil characteristics and temporal measurements. In contrast, categorical variables such as SEASONAL, LVET, and SVET had low correlations with other variables, suggesting a more independent relationship.

Table 3. Random forest model evaluation.

Model Evaluation		
No	Training Data	Testing Data
R^2	1.000	1.000
MSE	0.0001	0.0002
RMSE	0.0080	0.0132
KGE	1.000	1.000

In Table 3, the model demonstrated very high performance, with R^2=1.000 for both training and testing data. Although this result indicated an excellent fit, a perfectly high R^2 value suggested overfitting, meaning the model fit itself too closely to the training data and lost its ability to generalize to new data. To reduce the risk of overfitting, additional validation was conducted to ensure accuracy in a broader scenario. Since the prediction results were extremely high, reaching 1.000, further parameter tuning was performed using RandomSearchCV and GridSearchCV. This process aimed to find the optimal combination of parameters, ensuring that the model was not only accurate on training data but also capable of making reliable predictions on new data. RandomSearchCV was used to perform a broad search across a wide range of parameters, while GridSearchCV fine-tuned the parameters based on the best results from RandomSearchCV.

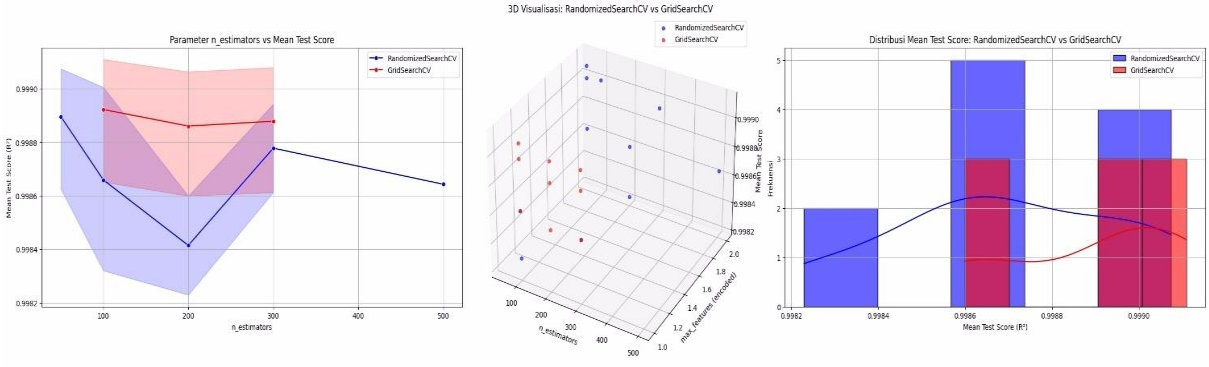

Figure 3. Parameter tuning.

Figure 3 is based on the results of the parameter search using GridSearchCV and RandomizedSearchCV. It was found that GridSearchCV provided more optimal results compared to RandomizedSearchCV. The best parameters identified were n_estimators = 200, max_features = 'sqrt', and min_samples_leaf = 2, which produced the highest Mean Test Score and were more consistent than other parameter combinations. When visualizing the mean test score, it was observed that GridSearchCV had a more concentrated distribution of scores at high values, while RandomizedSearchCV exhibited greater variability, including some lower results.

Additionally, the relationship between n_estimators and the Mean Test Score showed that the model performance reached an optimal value at n_estimators around 200–300. Beyond this point, adding more estimators did not lead to significant improvement. Furthermore, the 3D scatter plot analysis revealed that the parameter combinations tested by GridSearchCV were more concentrated at the optimal point, whereas RandomizedSearchCV demonstrated a wider spread. This finding indicated that GridSearchCV was more effective in identifying the best parameters, while RandomizedSearchCV was more exploratory but less stable in its results.

The Mean Test Score distribution showed a wider variety of values in RandomizedSearchCV, which resulted in greater variation, whereas GridSearchCV produced a more concentrated distribution around the highest values, indicating higher stability. Based on this analysis, the best model selected was the one from GridSearchCV, with the optimal parameters identified. This model was then used for predictions on new data. Overall, the parameter tuning results demonstrated that proper parameter selection had a significant impact on model performance. GridSearchCV proved to be more reliable in producing models with

higher accuracy and better stability compared to RandomizedSearchCV. Following this parameter tuning, the model was re-evaluated, as shown in Table 4.

Table 4. Modeling evaluation after parameter tuning.

	Before Tuning	After Tuning
R^2 (Training)	1.000	0.9999
R^2 (Testing)	1.000	0.9995
MSE(Training)	0.0001	0.0189
MSE (Testing)	0.0002	0.1098
RMSE(Training)	0.0080	0.1376
RMSE(Testing)	0.0132	0.3314
KGE(Training)	1.000	0.9992
KGE(Testing)	1.000	0.9986

Table 4 presents the model evaluation results after parameter tuning, showing a slight change in performance. The R^2 value, which previously reached 1.000 for both training and testing data, decreased slightly to 0.9999 for training and 0.9995 for testing, still indicating very high accuracy. Additionally, the Mean Squared Error (MSE) increased, particularly in training data, from 0.0001 to 0.0189, and in testing data, from 0.0002 to 0.1098. A similar trend was observed for the Root Mean Squared Error (RMSE), which rose from 0.0080 to 0.1376 in training and from 0.0132 to 0.3314 in testing. However, the Kling-Gupta Efficiency (KGE) value remained high, with a slight decline from 1.000 to 0.9992 for training and from 1.000 to 0.9986 for testing, indicating that the model still maintained a strong balance between correlation, bias, and prediction variability. These changes demonstrate that parameter tuning helped mitigate overfitting by making the model less perfect in fitting the training data. Although the error increased, the model became more realistic and gained better generalization ability for new data. Overall, parameter tuning enhanced model balance while still maintaining very high performance.

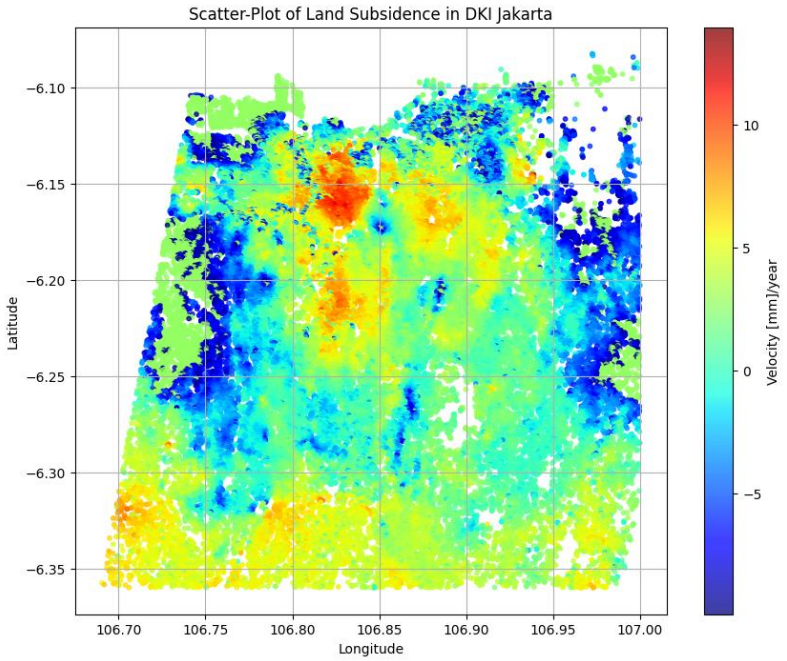

Figure 4. Scatter plot of land subsidence.

The scatter plot visualizes land subsidence and surface changes across DKI Jakarta. Areas with significant subsidence are represented in dark blue to bluish-green, indicating negative velocity values. In contrast, green and light yellow regions signify stable areas with minimal or no subsidence. The most affected areas are North Jakarta and West Jakarta, where substantial subsidence is observed. Meanwhile, Central Jakarta and South Jakarta remain relatively stable. Additionally, certain areas exhibit positive velocity values, indicating a slight increase in elevation.

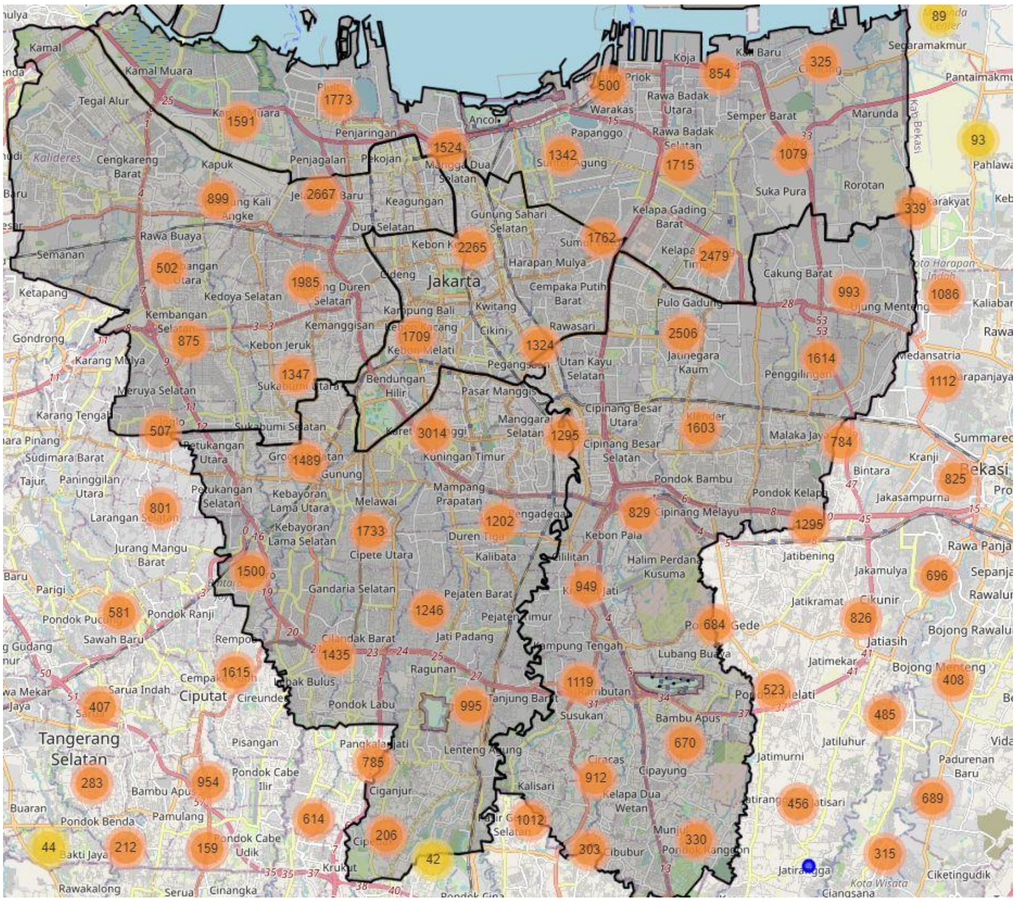

Figure 5. Land subsidence map.

Figure 5 is an area with subsidence in DKI Jakarta where the data is filtered based on geographic coordinates covering the DKI Jakarta area with a latitude range between -6.36 to -6.08 and longitude between 106.48 to 107.00. The data used includes subsidence velocity (VEL) and cumulative displacement visualized in an interactive map using folium. Each map point is assigned a color and size based on the subsidence velocity and cumulative displacement. Red indicates subsidence velocity of more than 5 mm/year while blue indicates subsidence velocity of less than 5 mm/year. The marker size is proportional to the cumulative displacement value using a radius scale = max (5, min (CUMUL.DISP./5.30)) to limit the size to remain visible on the map.

The orange and red areas show a significant level of decline in North Jakarta around Penjaringan, Tanjung Priok, Ancol, Cilincing, Kelapa Gading, Sunter, Pantai Indah Kapuk, and other coastal areas show a significant decline as well as the surrounding areas compared to the South. West Jakarta and Central Jakarta with areas such as Kebon Jeruk, Palmerah, Tanjung Duren, Mangga Besar show a significant decline as well but not as dense as north Jakarta. The central Jakarta area has some decline around Gondangdia and Cikini. The southern and eastern

regions have a lower decline such as Cililitan, Ujung Menteng, Kuningan and around Senayan. These coastal areas show a more significant decline than the South. Central Jakarta itself has excessive water use, while north Jakarta has dominant geological factors due to compaction. Regarding the east Jakarta area, South Jakarta shows land surface stability with a slight decline compared to north Jakarta and central Jakarta. The north Jakarta area is the top priority for mitigating the risk of land subsidence.

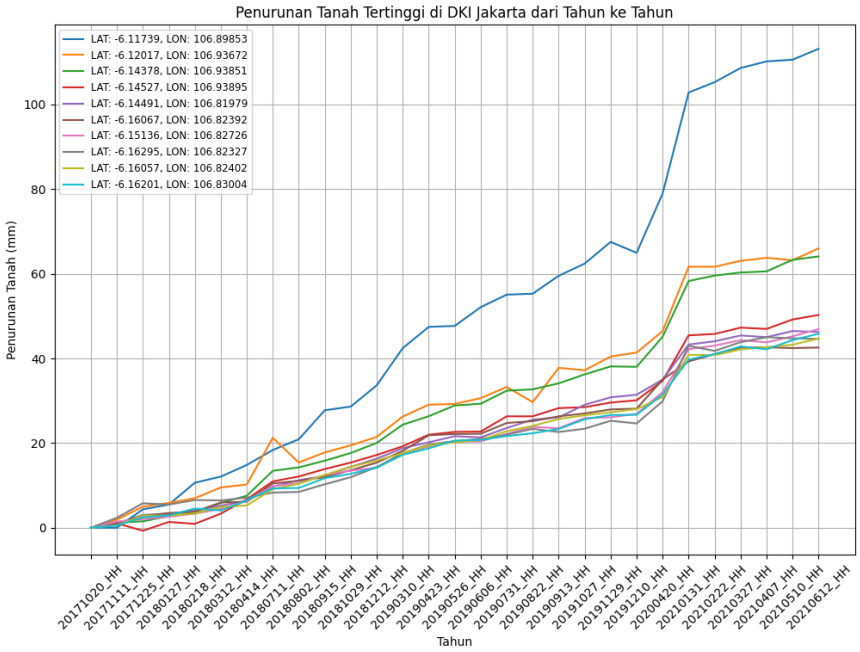

Figure 6. Height reduction area.

Figure 6 illustrates the pattern of accumulated land subsidence at the study site. The first area experienced a total cumulative subsidence of 680.55 mm from 2017 to 2021, with an average subsidence rate of 2.70 mm/year and data coherence of 0.98. Similarly, the second area recorded a total cumulative subsidence of 678.98 mm, with an average rate of 2.70 mm/year and a coherence value of 0.98. Both locations exhibited the most significant subsidence, showing an exponential increase after 2019. However, in the second area, land subsidence stabilized until 2021, though the decline was less severe compared to the first location.

Figure 7 shows that the area with significant subsidence is expanding, especially in North Jakarta and some areas in coastal Jakarta. In 2017, the color distribution shows that subsidence was still small and focused on North Jakarta but unevenly distributed. In 2018, there was an increase in the intensity of the blue color, indicating that the increase in subsidence was becoming more pronounced in parts of North Jakarta and widespread in other parts. In 2019 there was a wider distribution of subsidence with increasing intensity, especially in North Jakarta and West Jakarta, indicating more extreme subsidence above 20mm/year. In 2020, there was a slight increase in the area that experienced a significant decline, some points experienced a decline of more than -30mm / year. And in 2021 shows a more extreme decline, especially in coastal areas with several points experiencing a decline of more than -40 mm / year in North Jakarta and West Jakarta more.

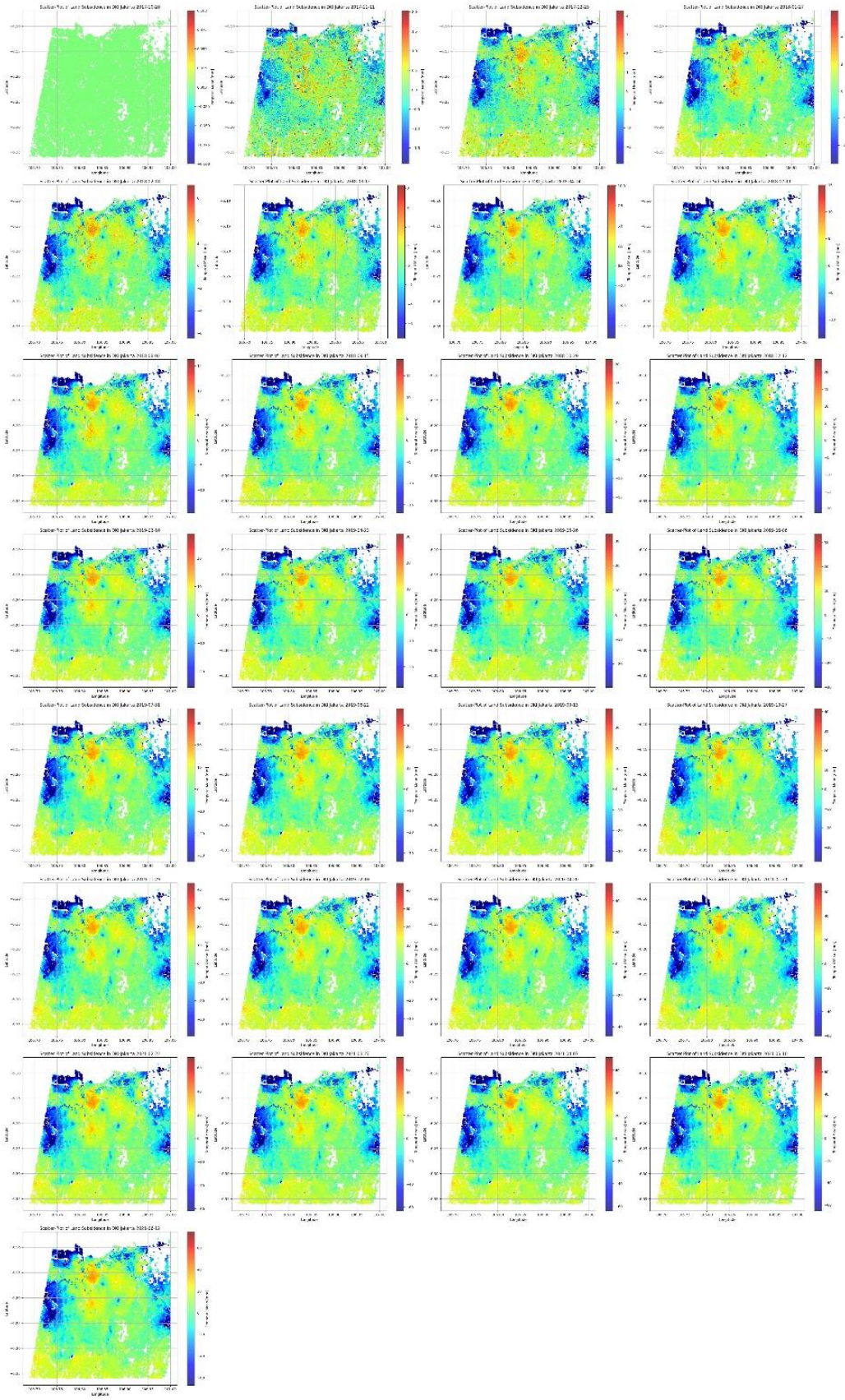

Figure 7. Scatter plot for 2017-2021.

Figure 8 shows that the red color indicated areas with lower elevations, which were associated with reclamation areas, while the blue color represented areas with higher elevations, consisting of alluvial soils or naturally elevated land. North Jakarta predominantly

displayed red coloration, but some blue areas were still present, indicating a division between reclaimed land and alluvial terrain. South and Central Jakarta exhibited a combination of red and blue. However, low elevation did not always indicate reclamation, as Central and South Jakarta were not reclaimed but still had low-lying areas. North Jakarta showed a denser concentration of reclamation.

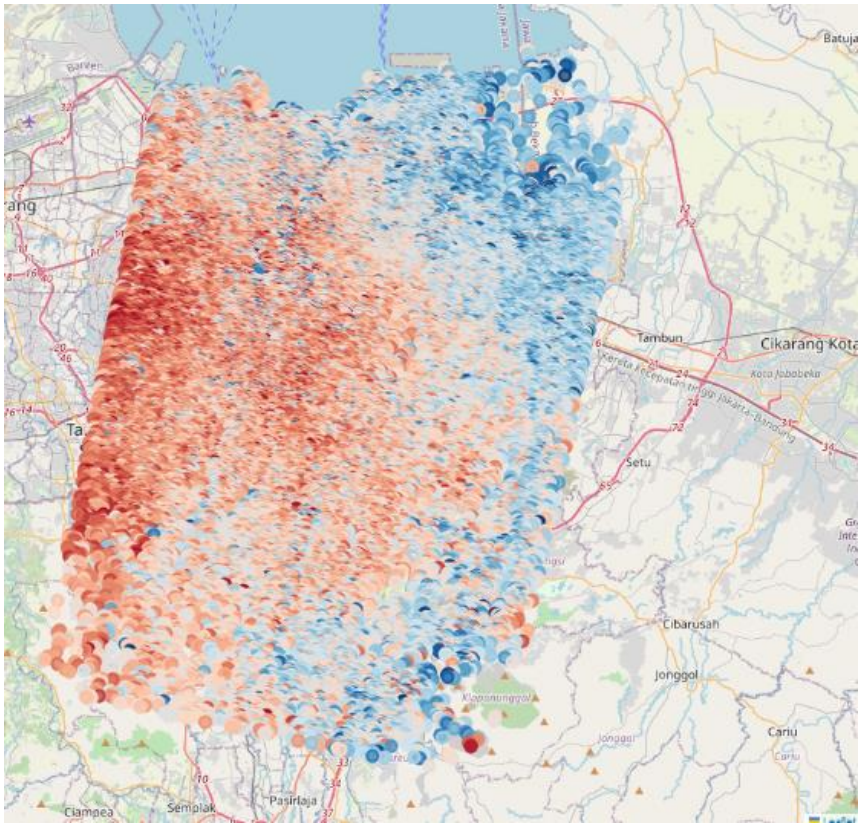

Figure 8. Map of reclaimed and alluvial areas.

3.3.Result comparison.

In this research, land subsidence predictions using the Random Forest method were compared with the PS-InSAR method. Table 5 below presents a comparison of the results obtained from both methods.

Table 5. PS-InSAR and random forest results comparison table.

	PS-InSAR	**Random Forest**
Region Highest decline	North Jakarta, West Jakarta, and Central Jakarta	North Jakarta, West Jakarta, and Central Jakarta
Lowest Decline Region	South Jakarta, East Jakarta	South Jakarta, East Jakarta
Area Reklamation	North Jakarta	North Jakarta
Alluvial Region	South Jakarta	South Jakarta

In Table 5, the PS-InSAR method indicates that land subsidence occurred in several areas of DKI Jakarta, particularly in North Jakarta (Pantai Indah Kapuk, Marunda, Ancol) and West Jakarta (Kembangan). Additionally, slight land subsidence was observed in the Kalibaru area, Central Jakarta. With the Random Forest method, significant subsidence was also detected in North Jakarta, particularly in Penjaringan, Tanjung Priok, Ancol, Cilincing, Kelapa Gading,

Sunter, Pantai Indah Kapuk, and coastal areas. In West and Central Jakarta, areas such as Kebon Jeruk, Palmerah, Tanjung Duren, and Mangga Besar also experienced considerable subsidence, though not as severe as in North Jakarta. In Central Jakarta, subsidence was observed in Gondangdia and Cikini. Meanwhile, the southern and eastern regions remained more stable. The analysis of both methods suggests that North Jakarta experienced the most significant subsidence, primarily due to geological conditions. The PS-InSAR method highlights that subsidence is more pronounced in reclamation areas such as Pantai Indah Kapuk, compared to regions with alluvial soil. Similarly, the Random Forest method supports the finding that North Jakarta is subsiding at a faster rate due to the instability of reclaimed land. Areas with alluvial soil exhibit slower subsidence compared to reclaimed areas, as they possess greater stability.

3.4.Predicted land subsidence 2022.

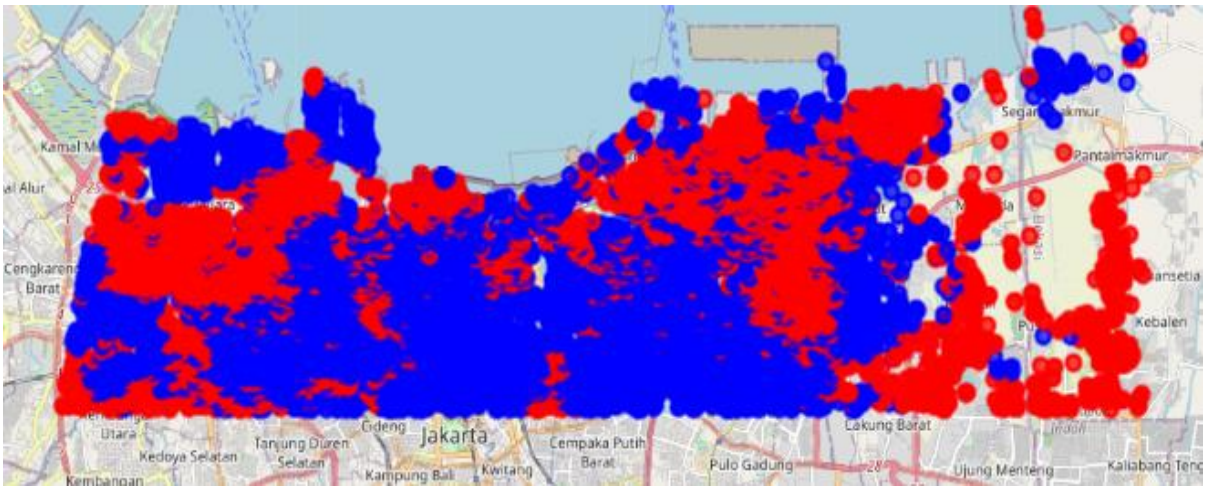

Figure 9. 2022 Prediction map.

Figure 9 presents a visualization of the land subsidence distribution map, where the red color indicates areas experiencing significant subsidence. The affected areas include Pantai Indah Kapuk, Tanjung Priok, and Cilincing, which continue to experience considerable subsidence. A small portion of the area around Pluit and the reclamation zone has shown a smaller decline compared to previous observations. Meanwhile, areas near the Central Jakarta border, such as Sunter and parts of Kemayoran, display more varied colors, suggesting that land changes in these regions are more balanced.

The land subsidence prediction results in Figure 10 indicate that areas in solid blue exhibit a decrease of up to -60 mm/year. A significant increase in land subsidence is observed in several parts of North Jakarta. Conversely, some areas marked in solid red indicate an increase in land surface elevation of more than 60 mm/year. In general, North Jakarta experiences dominant geological factors contributing to subsidence, primarily due to compaction. According to the study by Rendi Handika et al., titled *"Combined Land Subsidence Analysis in Jakarta Based on PS-InSAR and MICMAC Methods,"* land subsidence in DKI Jakarta is estimated at approximately -57.1 mm/year.

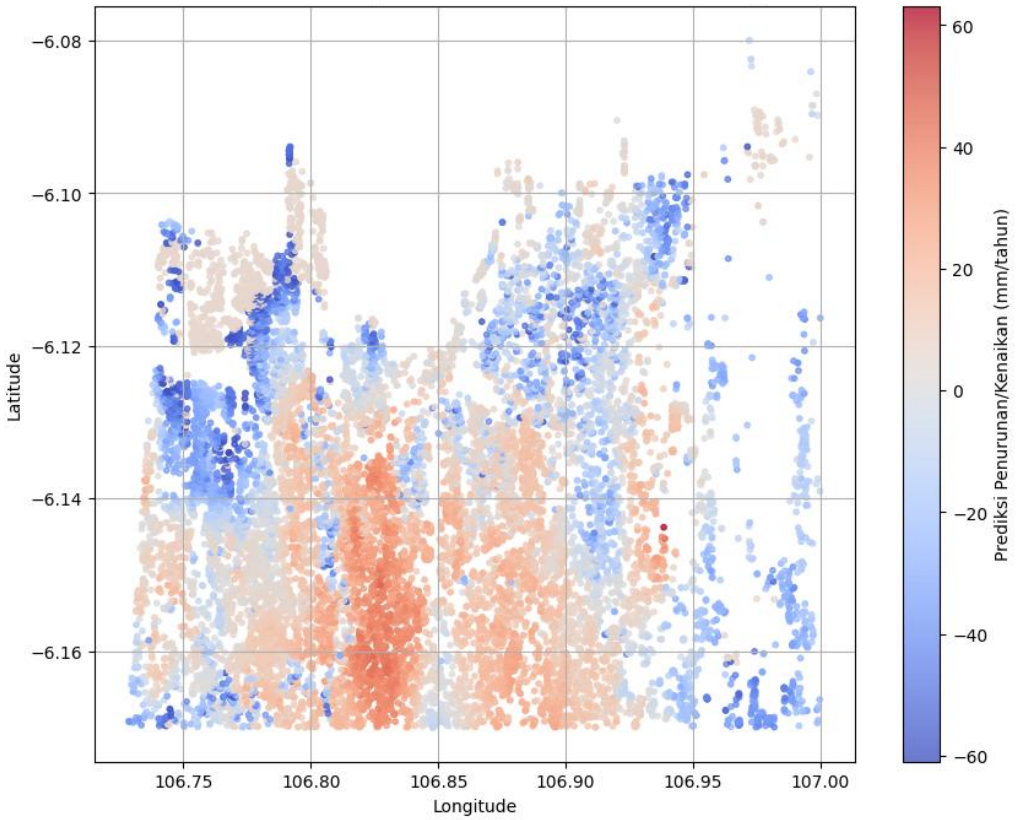

Figure 10. Scatter plot of 2022 predictions.

Table 6. Result score error.

Score Error Results: Predicted 2022 Decline	
R^2	2.90 mm/tahun
MAE	08.41 mm/tahun
RMSE	-1.79

The validation evaluation in Table 6, using R^2, shows that the average prediction deviates by approximately 2.90 mm/year from the actual data. The MAE result is 8.41 mm/year, indicating a relatively small prediction error, as the impact of extreme values has been minimized. However, the RMSE value is negative, suggesting that the prediction model requires further adjustments to improve accuracy and align more closely with the MICMAC data.

Table 7. 2022 Prediction modeling evaluation.

No	Training Data	Testing Data
R^2	0.9999	1.000
MSE	0.0349	0.0349
RMSE	0.1868	0.0997
KGE	0.9996	0.9993

Overall, the model evaluation in Table 7, using the 2022 prediction data, shows an R^2 value of 1.0000, indicating that the model can explain almost all the variance in the data. The high Kling-Gupta Efficiency (KGE) value further confirms a strong correlation between the model and the observed data. Additionally, the RMSE and MSE values demonstrate that the prediction error in the training data is very small. The slightly lower error in the testing data compared to the training data suggests that the model maintains high accuracy and generalizability.

4. Conclusions

The analysis of land subsidence in DKI Jakarta using the Random Forest method aligns with the results obtained from the PS-InSAR method. Both approaches indicate that land subsidence is most prominent in North and West Jakarta. Furthermore, the Random Forest method supports the finding that North Jakarta is experiencing accelerated land subsidence due to the unstable nature of reclaimed land. In contrast, areas with alluvial soil characteristics experience slower subsidence, as they provide greater stability. Beyond analysis, the land subsidence prediction using the 2017–2021 dataset indicates a decline of up to -60 mm/year. Validation against findings from a study by Rendi Handika et al., titled "Combination of Land Subsidence Analysis in Jakarta Based on PS-InSAR and MICMAC Methods," shows that land subsidence in DKI Jakarta is approximately -57.1 mm/year. The evaluation using R^2 reveals that the average prediction differs by about 2.90 mm/year from actual data. The MAE result of 8.41 mm/year suggests a relatively small prediction error, as the influence of extreme values has been minimized. However, the RMSE value is negative, indicating that the prediction model requires further refinement to improve accuracy in alignment with MICMAC data. For future research, a more in-depth investigation of critical areas, particularly North Jakarta, is recommended. This could be achieved by comparing prediction results with direct observations from satellite imagery or in-situ measurements to enhance model validation and reliability.

Acknowledgments

We sincerely thank our colleagues and research assistants for their invaluable contributions, support, and resources that made this research possible.

Competing Interest

The authors declare that they have no competing interests.

Authors Contributions

Camelia Nur Hidayah: Conceptualization, Methodology, Data Collection, Writing – Original Draft Preparation, Visualization, Data Analysis; Panca Dewi Pamungkasari: Supervision; Sari Ningsih: Supervision; Joko Widodo: Supervision; Elfady Satya Widayaka ; Supervision Muhammad Fauzan Azhiman: writing

References

[1] Widodo, J.; Arief, R.; Dinanta, G.; Setyaningrum, N.; Setiyoko, A.; Putra, A.; Oktaviani, A.; Wisyanto; Santoso, E.; Hidayat, N.; Awaluddin; Kurniawan, F.; Pradono, M.H.; Razi, P.; Izumi, Y.; Sri Sumantyo, J. (2022). Time Series Land Subsidence Analysis Based on Persistent Scattered Interferometric Synthetic Aperture Radar Method of Jakarta City Region Using Terra SAR X Spaceborne Data. *International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, *2022*, 106–113. http://doi.org/10.1109/ICRAMET56917.2022.9991211.

[2] Tien Bui, D.; Shahabi, H.; Shirzadi, A.; Chapi, K.; Pradhan, B.; Chen, W.; Khosravi, K.; Panahi, M.; Bin Ahmad, B.; Saro, L. (2018). Land Subsidence Susceptibility Mapping in South Korea Using Machine Learning Algorithms. *Sensors*, *18*, 2464. http://doi.org/10.3390/s18082464.

[3] Ahmed, S.; Hiraga, Y.; Kazama, S. (2024). Land Subsidence in Bangkok Vicinity: Causes and Long-Term Trend Analysis Using InSAR and Machine Learning. *Science of the Total Environment*, *946*, 174285. http://doi.org/10.1016/j.scitotenv.2024.174285.

[4] Hosseinzadeh, E.; Anamaghi, S.; Behboudian, M.; Kalantari, Z. (2024). Evaluating Machine Learning-Based Approaches in Land Subsidence Susceptibility Mapping. *Land*, *13*, 322. http://doi.org/10.3390/land13030322.

[5] Liu, B.; Mazumder, R. (2024). Randomization Can Reduce Both Bias and Variance: A Case Study in Random Forests. http://doi.org/10.48550/arXiv.2402.12668.

[6] Handika, R.; Widodo, J.; Pravitasari, A.E. (2024). Combined Land Subsidence Analysis in Jakarta Based on Ps-InSAR and MICMAC Methods. *Jurnal Teknologi Lingkungan*, *25*, 137–145. http://doi.org/10.55981/jtl.2024.1125.

[7] Maulina, S. (2021). Prediksi Penurunan Muka Tanah Terhadap Pola Konsumsi Air Bersih di DKI Jakarta Tahun 2021-2025 Menggunakan Machine Learning. Undergraduate Thesis, Universitas Bakrie, Indonesia.

[8] Hakim, W.L.; Achmad, A.R.; Lee, C.-W. (2020). Land Subsidence Susceptibility Mapping in Jakarta Using Functional and Meta-Ensemble Machine Learning Algorithm Based on Time-Series InSAR Data. *Remote Sensing*, *12*, 3627. http://doi.org/10.3390/rs12213627.

[9] Khoirunisa, R.; Yuwono, B.D.; Wijaya, A.P. (2015). Analisis Penurunan Muka Tanah Kota Semarang Tahun 2015 Menggunakan Perangkat Lunak Gamit 10.5. *Jurnal Geodesi Undip*, *4*, 341-350. http://doi.org/10.14710/jgundip.2015.9961.

[10] Probst, P.; Wright, M.N.; Boulesteix, A.-L. (2019). Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining and Knowledge Discovery*, *9*, e1301. http://doi.org/10.1002/widm.1301.

[11] Wanto, K. (2016). Analisis Intervensi Data Deret Waktu Untuk Peramalan Pendapatan Domestik Bruto Indonesia. Undergraduate Thesis, Universitas Negeri Jakarta, Indonesia.

[12] Apit, S.; Sunarno, D.; Djuniadi. (2024). Perbandingan Performa Model Machine Learning dalam Prediksi Suhu di Semarang. *Jurnal Ilmu Teknik Elektro dan Teknologi Informasi*, *12*, 2770-2775. http://doi.org/10.23960/jitet.v12i3.4884.

[13] Nazar, R. (2024). Implementasi Pemrograman Python Menggunakan Google Colab. *Jurnal Informatika dan Komputer*, *15*, 50-56.

[14] Lestari, E.; Astuti, I. (2022). Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah dan Cosine Similarity untuk Rekomendasi Rumah pada Provinsi Jawa Barat. *Jurnal Ilmiah FIFO*, *14*, 131. http://doi.org/10.22441/fifo.2022.v14i2.003.