ISSN 2354-9114 (online), ISSN 0024-9521 (print)
Indonesian Journal of Geography Vol 56, No 3 (2024): 375-385
DOI: 10.22146/ijg.94726 website: https://jurnal.ugm.ac.id/ijg

INDONESIAN JOURNAL OF GEOGRAPHY

RESEARCH ARTICLE

Assessment of Flood Risk Induced by Land Subsidence Using Machine Learning

B. D. Yuwono* [1], **L.M. Sabri**[1]. **A.P. Wijaya**[1], **and M. Awaluddin**[1]

[1]Department of Geodetic Engineering, Faculty of Engineering, Diponegoro University, Indonesia

Submit: 2024-03-10.
Received: 2024-04-02
Accepted: 2024-07-25
Published: 2024-10-03

Key words: flood risk, machine learning, dataset, hyperparameter.

Abstract Semarang City is facing significant environmental challenges, with land subsidence being a critical issue that intensifies flood inundation and worsening flood damage. As urban areas expand and climate change impacts become more pronounced, understanding and mitigating flood risks are crucial for sustainable urban development and disaster management. Therefore, this study aimed to assess flood risk induced by land subsidence using machine learning to improve flood management. Five different machine learning models (MLMs) were used to assess flood risk, which included Decision Tree (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). Additionally, fourteen different indices and 2884 sample points were used to train and test the models, with hyperparameter optimization ensuring fairness in comparisons. To address uncertainty in the sample dataset, flood hot spots were used to validate the rationality of flood risk zoning maps. The study investigated driving factors of different flood risk levels, focusing on flood areas to determine flood risk mechanisms in the highest-risk areas. The results showed that KNN performed the best and provided the most reasonable flood risk value among the models. Meanwhile, curve number (CN), distance to the river (DTRiver), and Building Density (BD) were identified as the top three significant factors of flood risk, ranked using the average score decrease in KNN model. Finally, this study expanded the application of machine learning for flood risk assessment and also deepened understanding of the potential mechanisms of flood risk, and provided perceptions about better flood risk management.

Correspondent email: bdyuwono92@gmail.com

1. Introduction

Semarang is a city located on the north of Java Island and faces significant challenges due to its vulnerability to flooding. The population of this city is 1.7 million, with several individuals living in lowland areas that are prone to flooding. The impacts of flooding are very severe, including property damage, loss of life, and disruption of essential services. As a major economic center in Central Java, Semarang plays a significant role in commercial, industrial, and agricultural sectors. However, this economic importance is threatened by recurring flooding leading to infrastructure damage, loss of crops, and transportation disruption. The city is currently experiencing significant urbanization, with the occurrence of new residential and commercial developments in flood-prone areas. Therefore, to ensure new development is planned and constructed to be flood-resilient, there is a need for a comprehensive understanding of flood risk (Yuwono et al., 2021).

Land subsidence refers to the sinking of land, which can be caused by various factors such as groundwater pumping, natural geological processes, and human activities such as urbanization. When land subsides, its height decreases, thereby increasing flood risk. In the context of this study, Semarang experiences land subsidence, especially in the northeast area, which results in the expansion of flood inundation (Yuwono et al., 2024).

According to Chen et al. (2021), there are four methods for assessing flood risk which include, historical disaster mathematical statistics (HDMS), scenario simulation analysis (SSA), multi-criteria decision analysis (MCDA), and machine learning models (MLMs). HDMS includes analyzing the frequency and severity of past floods, as well as identifying patterns and trends in historical flood data. This information assists disaster management professionals in estimating the possibility and potential impact of future floods in affected areas. In this study, HDMS also considers factors such as climate change, land use (LU), and infrastructure development, which can affect flood frequency and severity. Moreover, this method requires a significant amount of historical data and may not generalize properly to rapidly changing environments. To identify vulnerable areas, SSA with 2D hydraulic/hydrodynamic model was adopted. The model requires extensive hydrological data, high-resolution Digital Elevation Models (DEMs), and geometric data, which make the models computationally expensive and resource-intensive.

MCDA has been widely applied in many regions (Ha-Mim et al., 2022) using a systematic method. This method combines analytical hierarchy process, multi-criteria decision-making, and indexing methods with geographic information systems (GIS) to create comprehensive flood risk maps. Furthermore, the method includes weighting indices to measure risk, which often relies on the knowledge of experts. On the other hand, MLMs use intelligent algorithms to automatically learn flood risk characteristics, providing a new perspective on reliable flood risk assessment (Deroliya et al., 2022). Both MLMs and MCDA are flexible methods for evaluating complex systems.

However, MLMs produce more objective outcomes than MCDA, as the models rely on statistical analysis and data rather than subjective expert opinions. The models can handle missing or incomplete data, resulting in a more efficient and cost-effective solution for flood risk mapping compared to SSA, which is resource-intensive and computationally demanding (Chen et al., 2021).

MLMs have become increasingly popular and are even more effective and appropriate for assessing flood risk in recent years. Despite this popularity and effectiveness, there is still a need to further explore and develop the model (Horvitz and Mulligan, 2015). Previous studies have only focused on risk zoning, defining high-risk areas, and analyzing methods. There has been little investigation into flood risk analysis at various levels or detailed characterization of land subsidence areas.

Support Vector Machine (SVM) (Salvati et al., 2023) is effective in mapping flood susceptibility by using support vector regression with hyperparameter optimization, achieving high accuracy. Meanwhile, Random Forest (RF) (Costache et al., 2022) (Saber et al., 2023) is a popular choice due to its accuracy in flood risk assessment and high performance in predicting flood-prone areas. Logistic Regression (LR) is another valuable tool (Costache et al., 2022) for flood susceptibility zonation mapping, as it provides accurate results for identifying flood risk areas. Decision Trees (DT) are known for interpretability and effectiveness in flood risk assessment, which makes these models a widely used method in flood mapping endeavors (Costache et al., 2024; Prakash et al., 2023). The discussion on the improved spatial K-Nearest Neighbor (KNN) algorithm focuses on incorporating remote sensing (RS) and GIS data to predict flood inundation risk. This comprehensive framework includes the acquisition of spatial data, development of predictive models, implementation of risk mapping, and evaluation modules (Liu et al., 2021).

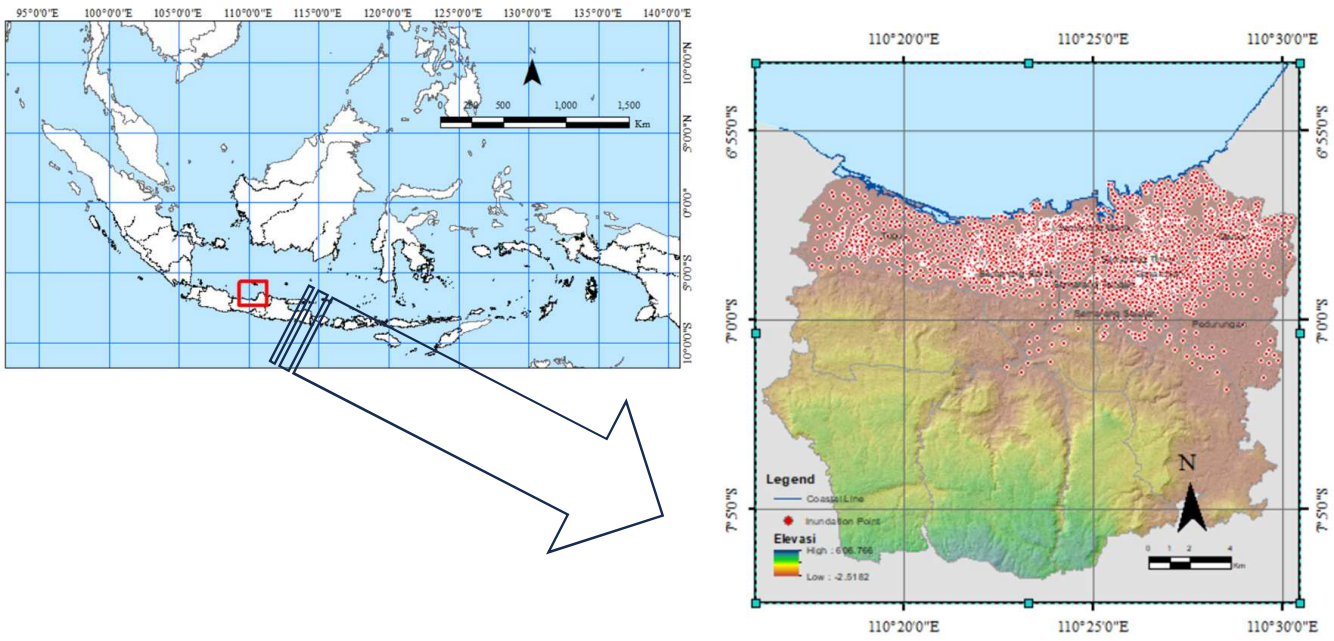
A total of five different MLMs, consisting of DT, RF, SVM, KNN, and LR are applied to evaluate flood risk. The accuracy of the model is determined by assessing the prediction accuracy and the Receiver Operating Characteristic (ROC) curve, which are evaluated using a separate testing dataset. The resulting risk maps are compared to recorded inundation hot spots. Consequently, this study aimed to determine the factors leading to flood risk by examining the fundamental mechanisms of flood risk. To achieve this aim, optimal model was used in areas with varying levels of risk, with particular attention on locations with the highest risk.

The novelty of this exploration is in its comprehensive method of assessing flood risk in the context of land subsidence, specifically in the exceptional setting of Semarang. The study aimed to identify and evaluate ML methods adapted for assessing flood risk due to land subsidence, an area that may not be extensively covered in existing literature. Moreover, by distinguishing and analyzing risk factors at varying levels of severity, the exploration provides a nuanced understanding of how different factors contribute to flood risk. The study conducts a detailed spatial analysis of flood risk patterns in the most dangerous areas across various locations, offering understanding that can inform aimed interventions. However, the main aim of the exploration is to bridge the gap between theoretical study and practical application by providing actionable and implementable recommendations for flood risk management. These elements collectively contribute to the originality of the exploration and potential impact on improving flood risk management strategies in Semarang.

Semarang is located at 6o 58'S and 110o 25'E in the north of Java Island, and has a population of approximately 1.81 million people, with an annual growth rate of 1.57% per year (Yuwono et al., 2019). The northern area of Semarang is characterized by infrastructure facilities such as airports and bus stations, densely populated areas, ponds, and agricultural land, while the southern area is dominated by settlements. In addition, the geological structure of Semarang consists of three lithologies such as volcanic rocks, marine sediments, and alluvial deposits (Kuehn et al., 2010), as shown in Figure 1.

In this study, Kuehn et al. (2010) observed that these geological features have contributed to the exceptional

Figure 1. Study Area



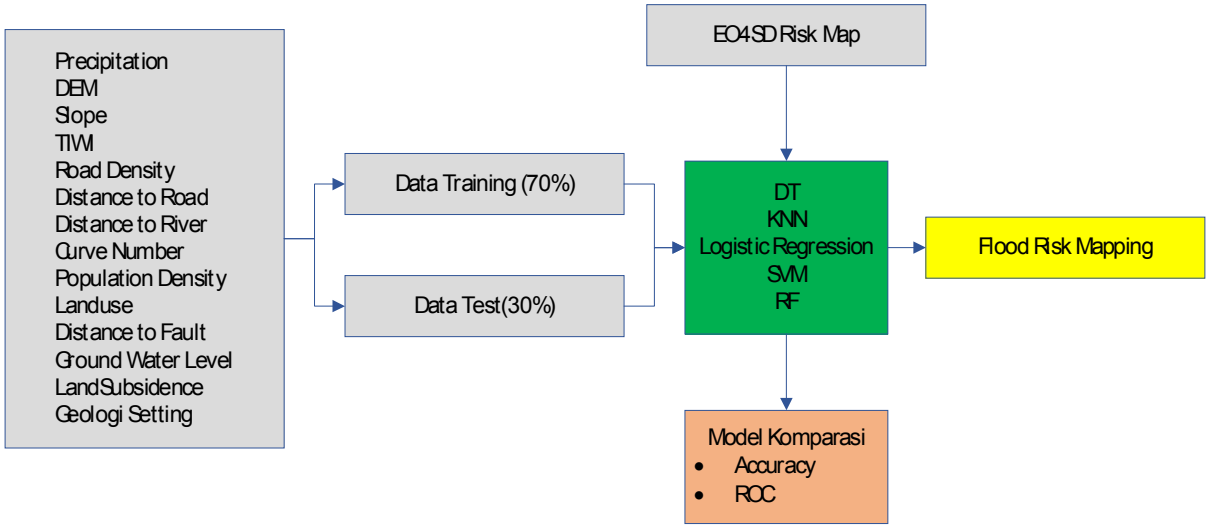


Figure 2. Research Flowchart

topography and landforms of Semarang, such as hills, valleys, and coastal areas. These features have also influenced the natural resources and ecosystems of the city, including forests, wetlands, and coastal mangroves. Moreover, studying the geological and environmental aspects of Semarang is crucial to understanding the development and sustainability of the city.

2. Methods

The assessment of flood risk using MLMs consisted of three stages, as shown in Figure 2. The first stage included creating a sample dataset for model training and testing by combining fourteen indices from three different aspects with the flood risk inventory map. In addition to this stage, the flood risk inventory map was considered as the aim variable to be modeled, while the indices served as the independent variables or predictors that were used to model the aim variable. The flood risk inventory map was based on a finding of EO4SD (2017), where EO4SD-Urban stands for "Earth Observation for Sustainable Development - Urban." Moreover, this method was a program that used Earth Observation (EO) data and satellite imagery to support sustainable urban development initiatives. The program focused on leveraging the capabilities of EO technology to address various urban challenges such as urban planning, infrastructure development, environmental management, and disaster risk reduction.

In the second stage, five machine-learning models were selected for the assessment of flood hazards. To ensure a fair comparison between the models, the training and assessment procedures of all models were considered. Furthermore, the accuracy of the models was evaluated using recorded inundation statistics to verify the predictions of the models. At the final stage, additional analyses were performed using the flood risk map produced by the best model. This process included exploring the characteristics of risk indices for different levels of risk and identifying patterns of flood risk in the highest-risk areas across different locations. Following this discussion, the study aimed to provide effective guidance for managing flood risk.

Machine Learning

2.0.1. Decision Tree

DT was used for flood modeling because it was efficient and effective, with a simple procedure for easy interpretation (Chen et al., 2021). Even though DT took time to process data, it handled uncertainty for a significant level in the data set (Tehrany et al., 2019). Moreover, DT was very flexible in handling data of various scales and had high efficiency in managing complex relationships. DT algorithm classified the influencing factors hierarchically and equivalently according to the level of vulnerability and created decision rules based on a tree structure that was built on a significant level from a set of independent parameters (Tehrany et al., 2019).

2.0.2. Random Forest

RF was a type of machine learning algorithm that combined several DT to create an ensemble classifier. Each DT in the ensemble was built using a random subset of the data and a random subset of the features. In addition, RF used a voting mechanism to combine the predictions of each DT, leading to a total prediction that was often more accurate and strong compared to a single DT. RF method had proven to be highly effective in solving classification and prediction problems, including those related to flood risk assessment. Following this discussion, the method was also widely used in flood risk assessment and showed excellent predictive accuracy and generalizability (Wang et al., 2015). Two important factors in RF included the number of trees and variables in each division. In the context of this study, Out of Bag (OOB) error represented the prediction error rate, while the average decrease in Gini impurity index denoted the importance of variables in the model (Arabameri et al., 2019). Gini impurity index was expressed in the following Equation.

$$G(Xi) = \sum_{j-1}^{J} P(X_i = L_j)\left(1 - P(X_i = L_j)\right) = 1 - \sum_{j-1}^{J} P(X_1 = L_j) \quad (1)$$

$$OOB_{error} = \left(1 - \frac{1}{N \sum_{i\varepsilon oob}^{\delta i}}\right) \times 100\% \quad (2)$$

as follows:

$$G(Xi) = \sum_{j-1}^{J} P(X_i = L_j)\left(1 - P(X_i = L_j)\right) = 1 - \sum_{j-1}^{J} P(X_1 = L_j) \quad (1)$$

$$OOB_{error} = \left(1 - \frac{1}{N \sum_{i\varepsilon oob}^{\delta i}}\right) \times 100\% \quad (2)$$

Where:

$G(Xi)$ = purity index

$P(X_i = L_j)$ = category estimation

$X_i = L_j$ = probability

OOB_{error} = error prediction level OOB

N = number of observations

δi = truth indicator variable

2.0.3. Logistic Regression

LR is a statistical method that analyzes the relationship between multiple factors and the probability of an event occurring. Different from other methods, LR did not require the variables to have a normal or causal distribution and could work with both continuous and discrete variables or a combination of both. However, this process showed that the method was a versatile tool that was used in various applications. LR model provided an understanding of the strength of the relationship between the dependent variable and multiple independent variables (Ghosh & Dey, 2021). Moreover, the probability of the impact was expressed in Equation 3 as follows.

$$P = \frac{1}{(1+e^{-z})} \quad (3)$$

Where P was the probability of an impact between 0 and 1 on the sigmoid curve and z represented the linear combination shown in Equation 4 as follows:

$$z = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \quad (4)$$

Where was the intercept of the model, b1 represented the coefficient of LR model and x1 was the condition variable. As the value of the logistic coefficient increased, the chance of an impact increased. In this context, a positive coefficient value in the model showed that the increase in the value of the corresponding independent variable was associated with an increase in the value of the dependent variable, showing a direct relationship between the two variables. In essence, the presence or increase of these variables was expected to improve the impact, while the absence or decrease of the variables led to a decrease in the impact. A negative coefficient value showed an inverse relationship between the variables, showing that an increase in the independent variable was associated with a decrease in the dependent variable (Tehrany et al., 2017).

2.0.4. K-Nearest Neighborhood

The algorithm relied on feature similarity to create new predictions of data points. Predicted data was assigned a value depending on which points the data best matched in the training set. In general, KNN algorithm for flood modeling was as follows (Gauhar et al., 2021):

1. Determined the value of i.
2. Calculated the distance between the training data points and the points to be classified.
3. Sorted the training data by descending value distance.
4. Made predictions with the majority of nearest neighbors.

2.0.5. Support Vector Machine

SVM is a supervised MLMs that uses statistical learning theory and the principles of structural risk minimization. This method mapped native inputs into a high-dimensional feature space, finding the maximum separation margin between classes and constructing a hyper-classification field in the middle of the maximum margin (Tien Bui et al., 2018) with the help of a training dataset. Additionally, the strong ability of SVM to partition non-linear data made it a highly popular and efficient MLMs for assessing flood risk (Tehrany et al., 2019).

Flood Risk Assessment

Flooded areas were detected by examining the flood inventory of areas affected during the 2012-2020 floods, as well as documents obtained from the local government (The Regional Agency for Disaster Countermeasure of Semarang) that showed areas that had flooded historically. Following the discussion, the history of flood events was identified by a ground survey conducted in 2021. This survey was confirmed by Regional Board for Disaster Management (BPBD) in Semarang and also through social media, which showed the severity of the flooding. The process of selecting index variables should be systematic and inclusive, and the criteria for selecting index variables should be autonomous. Moreover, this study selected fourteen indices, which were categorized into three aspects, namely disaster triggers, disaster-prone environment, and disaster-affected population (Chen et al., 2021).

Disaster triggers consisted of four variables which included precipitation (P), land subsidence rate (Ls), geology (GL), and groundwater level (GWL) as reported by (Nadiri et al., 2021). Furthermore, precipitation triggered floods because when the precipitation rate exceeded the capacity of soil to absorb the water, the excess water flowed over the ground as surface runoff, which accumulated in lowlands and caused flooding. This process was worsened by factors such as topography, soil type, and LU. In addition to the discussion in this exploration, heavy and prolonged precipitation also increased the water levels in rivers, lakes, and other bodies of water, leading to flooding. Land subsidence was selected because it was often associated with the expansion of inundation of flooding (Abidin et al., 2015; Yuwono et al., 2021; Zainuri et al., 2022). Moreover, when the ground sank or settled due to natural or human-induced factors, such as over-extraction of groundwater, the elevation of the land surface was lowered. This process caused changes in the direction and flow of water, leading to the formation of new drainage patterns and the alteration of existing ones. In some cases, land subsidence also led to the formation of sinkholes, which rapidly drained water and caused localized flooding. The subsidence caused the elevation of river beds and canals to be lower, which increased the risk of flooding during heavy rainfall events. Therefore, land subsidence was closely associated with flood risks and worsened the impacts in affected areas. Topographic Wetness Index (TWI) showed the distribution of soil moisture in a watershed, while CN represented the amount of surface runoff from rainfall events. In addition, CN method was developed by Natural Resources Conservation Service (NRCS) in the United States (Soulis, 2021). Table 1 showed the fourteen indices considered as independent variables.

The hazard-prone environment consisted of eight indices, each representing a different characteristic of the environment

Table 1. Dataset for Fourteen Indices

Indices	Feature	Source	Data	Resolution	Type
1	Precipitation (P) 2017	Geospatial	Prihanto et al. (2017)	0.1 deg x 0.1 deg	Raster
2	Digital Elevation Model (DEMs)	Information Agency	DemNAS	8.2 x 8.2	Raster
3	Slope (S)	Geospatial Information Agency	DemNAS		
4	Topographic Wetness Index (TWI)	Cloud Processing	Topographic Indeks	200m x 200m	Raster
5	Road Density (RD) (m/ Km2)				Vector
6	Distance to Road (D2R)	Analysis	Geospatial Data shp	200m x 200m	Vector
7	Distance to River (DTRiver)	Analysis	Geospatial Data shp	200m x 200m	Vector
8	Curve Number (CN)	Analysis			Vector
9	Population Density (PD)	Analysis	Central Bureau of Statistics	200m x 200m	Vector
10	Land Use (LU)	Cloud Processing	Lansat Image	200m x 200m	Vector
11	Distance to Fault(D2F)	Analysis	Geology Map	20m x 22m	Vector
12	Ground Water Level (GWL)	Yuwono et al. (2013)	Geology Agency – Center for Groundwater and Environmental Geology	200m x 200m	Vector
13	Land Subsidence(LS)	Yuwono et al (2024)	Sentinel 1A	20 x 22 m	Vector
14	Geological Setting (GS)	Geology Agency	Geology Map	200m x 200m	Vector

(Chen et al., 2021). The first index was precipitation, and the second index was DEM, providing information about the elevation of the land surface. Furthermore, regions that showed a lower elevation in DEM were typically at a higher risk of experiencing flood hazards (Wang et al., 2015). The third index was slope (S), which showed the steepness of the terrain, while the fourth index was TWI. The fifth index was RD, showing how many roads were present in the surrounding area. RD served as a significant parameter in assessing the drainage capacity of a region since roads facilitated floodwater drainage. The sixth was Distance to Road (D2R), measuring how far an area was from a road. The seventh index was the distance to the river (DTRiver) which measured the proximity of an area to a river and the potential for flooding. Meanwhile, the eighth index was CN, used to assess the susceptibility of an area to soil moisture and waterlogging.

The disaster-prone area was determined by PD (Chen et al., 2021), and Indices were used to measure the intensity of the population and assets in the area (Li et al., 2020). Regions with higher population densities were expected to have greater vulnerability to flooding and suffer more severe consequences as a result. This effect happened because there were more people and assets at risk of damage or displacement. Therefore, it was important to consider population density when assessing the potential impact of flooding in a particular area.

LU significantly contributed to flood hazards (Miladan et al., 2019), as urbanization and industrial activities often included extensive groundwater extraction, leading to subsidence as the ground compacted and sank. Additionally, construction and paving over natural landscapes reduced the ability of the land to absorb water, increasing runoff and flooding. Agricultural practices, such as irrigation, also lowered groundwater levels, worsening subsidence. Moreover, deforestation and land development disrupted natural water flow and soil stability, further increasing flood risks.

The distance to the fault (D2F) significantly influenced flood hazards induced by land subsidence (Ebrahimy et al., 2020). Faults facilitated subsidence through tectonic activities such as faulting and ground movements. In addition, areas closer to active faults were more susceptible to subsidence, as fault movements caused rapid changes in land elevation. This subsidence altered drainage patterns, increased flood vulnerability, and compromised flood control infrastructure effectiveness. Moreover, faults intersected with groundwater aquifers, accelerating subsidence processes when water extraction occurred.

GS performed a crucial role in influencing flood risks induced by land subsidence (Abidin et al., 2015). Regions with loose, unconsolidated sediments such as clay, silt, and sand were more prone to subsidence when groundwater was extracted, leading to increased flood risk. Furthermore, extensive groundwater aquifers in certain geological settings worsened subsidence when over-exploited, while active tectonic areas experienced subsidence due to faulting, altering topography, and increasing flooding. Low permeability soils retained water longer, prolonging floods in subsided areas, but coastal and deltaic regions faced compounded risks from both subsidence and sea-level rise.

Machine learning methods were applied to create risk maps from spatial distribution models (SDMs) in a Python programming environment. There was a necessity to transfer all variables onto a grid with a resolution of 200 meters. Moreover, this process included converting the original data into a gridded format in which each variable value corresponded to a specific location on the grid, allowing for a more precise and accurate analysis of the data.

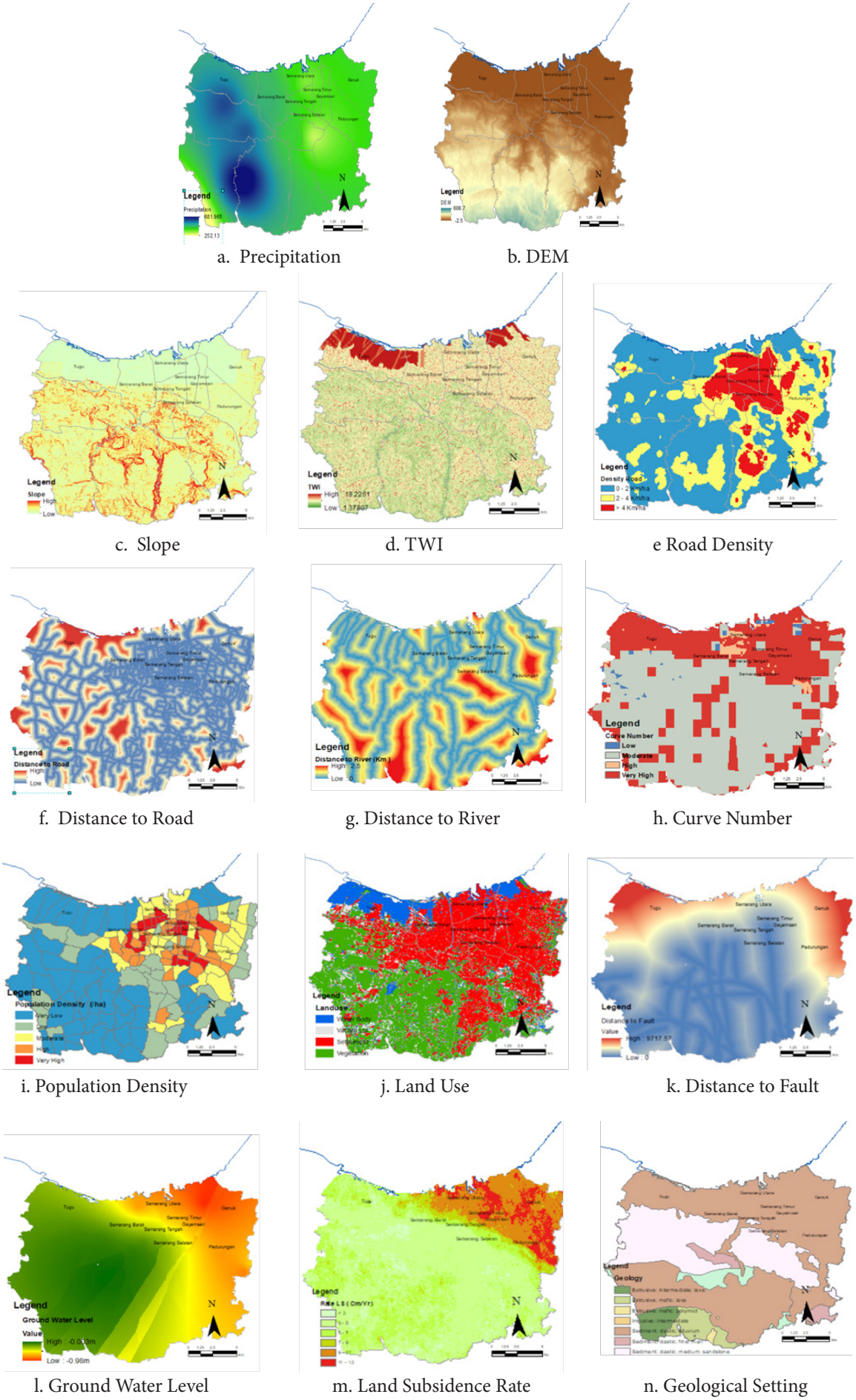


Figure 3. *Spatial Distribution of Fourteen Indices included (a)Precipitation, (b) DEM, (c) Slope, (d) TWI, (e) Road Density, (f) Distance to Road, (g) Distance to River, (h) Curve Number, (i) Population Density, (j) Land Use, (j) Distance to Fault, (l) Ground Water Level, (m) Land Subsidence Rate, and (n) Geological Setting.*

3. Result and Discussion

The dataset used in this study was created by using flood risk inventory maps that were published in EO4SD Flood Risk Map. This map was a project by European Space Agency (ESA) that aimed to develop and provide satellite-based information services to support disaster risk reduction and management activities related to floods. Additionally, the project used EO data to generate flood maps and risk assessment models. The spatial distributions of fourteen indices were shown in Figure 3.

Model Comparison

A hyperparameter was a parameter whose value was set before the learning process started. The parameter was not directly learned from the data but rather specified by the practitioner, usually through trial and error, and affected the performance of the learning algorithm. Following the discussion, examples of hyperparameters in machine learning included learning rate, regularization parameter, and number of hidden layers in a neural network. Hyperparameter function in machine learning was used to tune the parameter of a learning algorithm. However, this hyperparameter was a major part of the model selection process and was typically performed using a validation set. The function took in a set of hyperparameters as input and returned a scalar value that represented the performance of the learning algorithm on the validation set. The objective was to find the set of hyperparameters that produced the best performance on the validation set. Moreover, this process was known as hyperparameter tuning or hyperparameter optimization. The main hyperparameters of DT, KNN, LR, SVM, and RF were shown in Table 2. The flood risk maps of the five models, shown in Figure 4 were compared to the actual inundation points to determine the accuracy of the flood features captured by these models.

Six different models were selected with the best hyperparameters for data training and data testing which consists of 2230 points for training and 558 points for testing. The accuracy of these models on the testing dataset was recorded and shown in Table 3 to conveniently compare all the models' ROC curves. Additionally, for assessing multi-class categorization models using macro average (Figure 4), each class was handled equally irrespective of size because it calculated the average metric of each class without considering class imbalance. The metric such as accuracy, precision, recall, or F1-score was computed individually for each class to determine the macro-average. This method was helpful when there was an uneven distribution of classes and it was crucial to give each subject the same weight.

ROC curve was a graphical representation of the performance of a binary classifier system, plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC curve determined the optimal threshold for a given model, and the area under the curve (AUC) was a measure of how well the model distinguished between positive and negative classes. In addition, Figure 5 showed ROC curves for the different models being evaluated, with each curve representing the performance of a particular model, where x-axis represented FPR, and y-axis represented TPR. Each point on the curve corresponded to a specific threshold setting, and the diagonal line represented the performance of a random classifier. As a curve became farther from the diagonal line, the model's performance became better. Moreover, Table 3 showed AUC values for the different models being evaluated, along with other metrics such as accuracy, precision, recall, and F1 score. AUC values ranged from 0 to 1, with a value of 0.5 showing a random classifier and a value of 1 representing a perfect classifier. Given the scenario, as AUC value increased, the performance of the model increased. Both ROC and AUC provided a comprehensive evaluation of the performance of different models on the testing dataset. By analyzing ROC curves and AUC values, explorers determined which model was the most effective in distinguishing between positive and negative classes.

Table 2. Hyperparameter Tuning

Method	Hyperparameter Tuning
DT	'dt__criterion': 'gini', 'dt__max_depth': 5, 'dt__max_features': 'auto', 'dt__max_leaf_nodes': 8, 'dt__splitter': 'best'}
KNN	'knn__n_neighbors': 1, 'knn__p': 1
LR	'lr__C': 3, 'lr__multi_class': 'multinomial', 'lr__penalty': 'l2', 'lr__solver': 'lbfgs'
SVR	'kernel: Sigmoid' 'svm__C': 10, 'svm__gamma': 'scale',
RF	'rf__criterion': 'gini', 'rf__max_depth': 4, 'rf__max_features': 'auto', 'rf__max_leaf_nodes': 6, 'rf__n_estimators': 100}

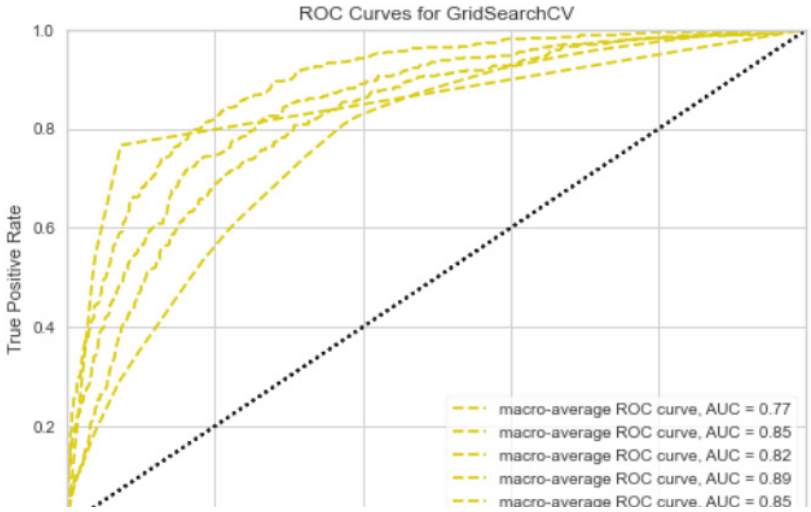


Figure 4. Macro Average Flood Risk in Five ML Models

Table 3. Accuracy and AUC Values of Five ML Models

Model	DT	KNN	LR	SVM	RF
Accuracy	45.52	76.21	51.25	64.77	60.36
ROC	0.77	0.85	0.82	0.89	0.85

The findings showed that KNN model performed the best, followed by SVM, RF, LR, and DT. Specifically, KNN achieved a prediction accuracy of 76.21% and AUC value of 0.85. Observation showed that the performance differences between the models were not significant, and the sample dataset used might have some uncertainty. Therefore, it would be premature to conclude that KNN was the best model for flood risk assessment.

KNN model was considered in this study to generate the flood risk map because it performed best at accurately predicting flood-prone areas. KNN algorithm excelled in classification tasks by finding the similarity between data points and assigning labels based on the majority class of neighboring points.

After analyzing the flood risk map, evidence was shown that urban areas bore the highest risk of flooding, particularly in North Semarang, Central Semarang, and Genuk districts. This finding showed the vulnerability of densely populated urban zones to flooding events. Moreover, urbanization led to increased impervious surfaces such as roads and buildings, which worsened flood risks by limiting natural drainage pathways and increasing surface runoff.

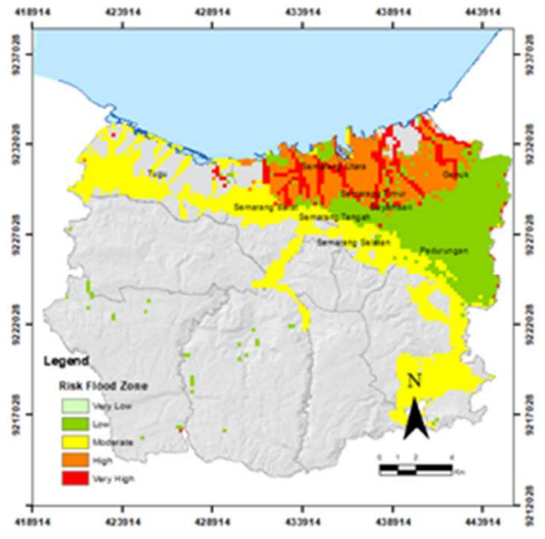
North Semarang, Central Semarang, and Genuk districts possibly faced heightened flood risk due to various factors such as inadequate drainage infrastructure, improper LU planning, and proximity to water bodies prone to overflow during heavy rainfall or storm events. Additionally, factors such as topography, soil type, and historical flood data might have contributed to the elevated risk levels in these regions.

Understanding the specific vulnerabilities of these urban districts was crucial for implementing effective flood mitigation measures and urban planning strategies. This process could include investment in better drainage systems, green infrastructure to improve water absorption, LU zoning regulations to limit development in flood-prone areas, and community awareness programs to promote preparedness and resilience against flooding events. However, by addressing these issues proactively, authorities could work towards reducing the impact of floods on both infrastructure and communities, eventually improving the total resilience of the affected regions.

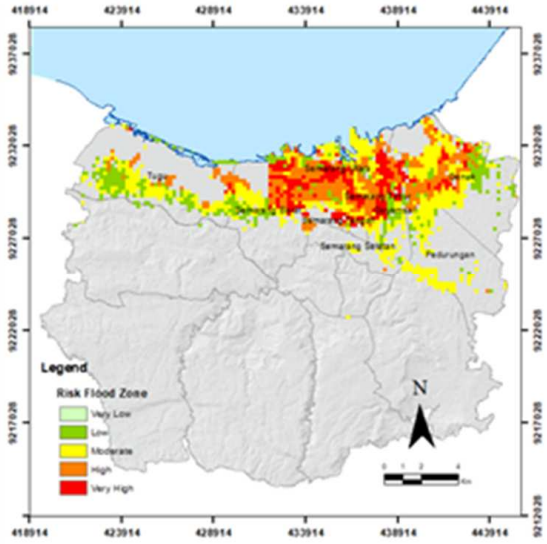
High-Risk Zone Analysis

The city of Semarang faced a range of flood risks that were connected to different factors across its various neighborhoods and districts. North Semarang area and neighborhoods located around the East_Banjir Kanal in North Semarang were particularly prone to flooding due to the urban texture, high population density, and presence of commercial centers of these cities. Conversely, the southern part of the Semarang had a low flood hazard rating because of its higher elevation and slope, causing less flood risk. The suburbs of the city such as Tugu district in the west-north, had a low vulnerability to flooding, leading to low levels of flood risk. Relating to the discussion, understanding the specific characteristics of each area was crucial for developing and implementing effective flood management strategies. Urban-coastal regions were observed to have the highest flood risk, largely due to severe disaster-inducing factors. These areas were particularly vulnerable because of their proximity to the coast and high population density. Consequently, implementing rigorous and scientifically backed measures for flood control in coastal cities was crucial. Such measures mitigated the potential damage and improved the resilience of these urban areas against future flooding events.

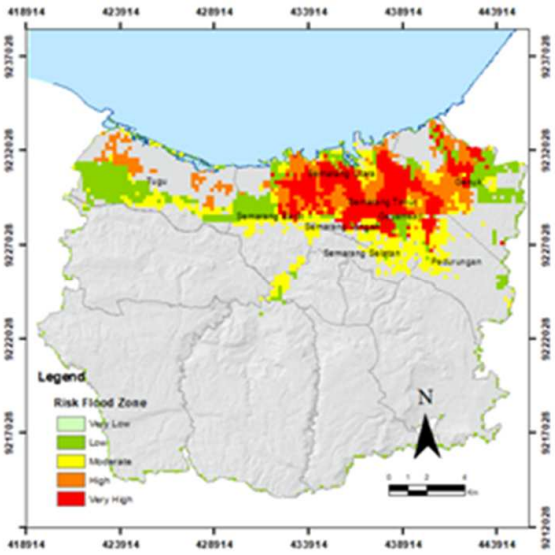
An average score decrease was a useful tool for understanding and optimizing MLMs, especially when handling dimensional datasets with many potential features (Breiman, 2001). This decrease was used to quantify the ranking of values, which were CN, DTRiver, BD, TWI, DR, D2R, slope, GWL, Geology, and Land LS, as shown in Figure 6.



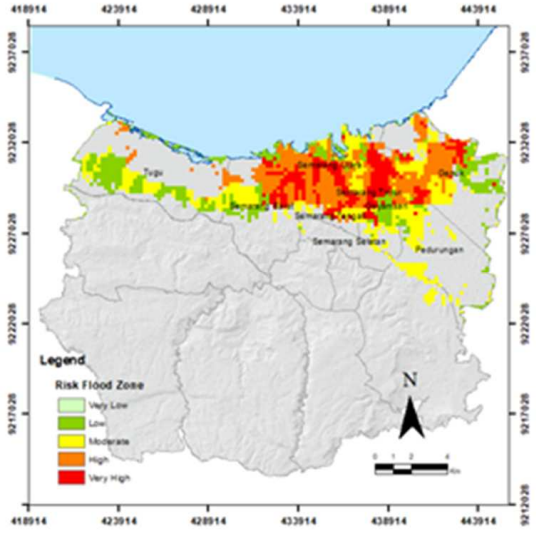
(a) DT



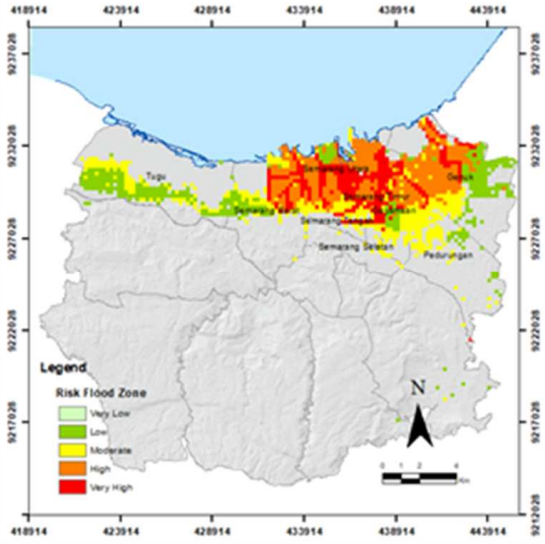
(b) KNN



(c) LR



(d) SVM



(e) RF

Figure 5. *Flood Risk Map using Five Models (DT, KNN, LR, SVM, and RF)*

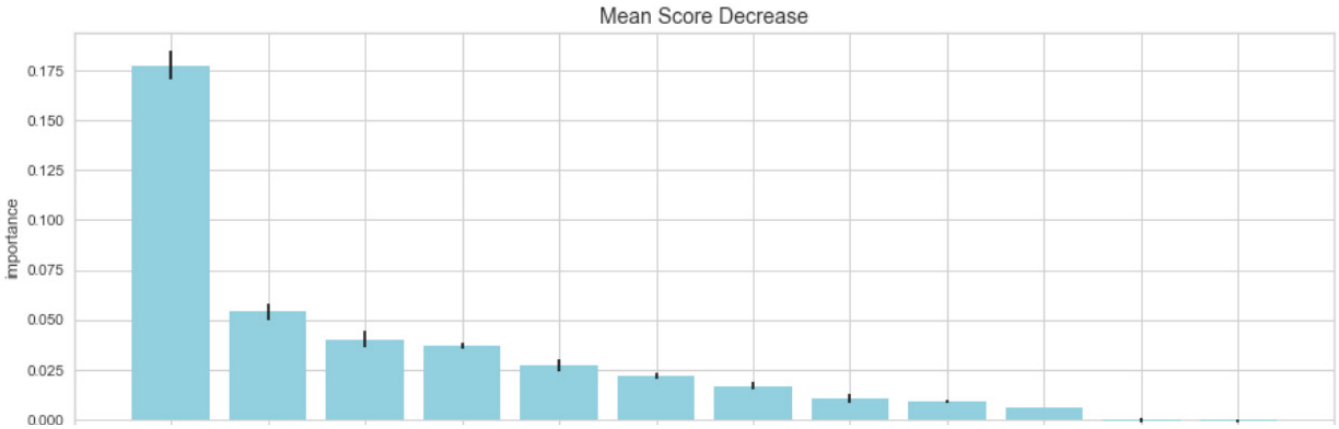


Figure 6. *Importance of Indices Feature to Flood Risk Map*

4. Conclusion

In conclusion, the five MLMs used in this study were carefully selected to ensure a comprehensive analysis of flood risk in the studied areas. The twelve indices used in the exploration provided a strong framework for flood risk assessment induced by land subsidence, and the flood risk inventory maps provided a comprehensive overview of the affected areas. Based on this study, KNN model had the highest performance, followed by SVM, RF, LR, and DT, respectively. Additionally, the model algorithm was a widely applied supervised learning method, known for its simplicity and effectiveness, as corroborated by (Le et al., 2021). In the context of flood risk mapping, which often included large datasets with various environmental and geographical factors, KNN was exceptional for its ability to handle high-dimensional data efficiently. The method achieved this result by focusing on the most relevant features for handling noisy or missing data (Tuerhong et al., 2021). Moreover, the model provided interpretable results by basing its predictions on the proximity of data points. This interpretability aided in understanding the factors contributing to flood risk and in identifying areas more possibly to be affected by flooding in Estahban Iran, as shown by (Razavi-Termeh et al., 2024). Additionally, radar interferometric methods and MLMs (KNN, RF, and CART) to predict and map land subsidence in a semiarid region of Iran with all three models had acceptable performance (Gharechaee et al., 2023).

After analyzing the potential mechanisms of flood risk, several critical factors were identified as susceptible to flooding, such as topography, LU, precipitation patterns, CN, flood vulnerability, land subsidence, and factors that contributed to land subsidence, such as geological setting and depletion of the elevation of groundwater. Moreover, CN was essential in this context, as it combined land use, soil type, and moisture conditions to estimate potential runoff, providing a simplified yet accurate representation of watershed's hydrological response. The results showed that CN was the most important variable in flood risk modeling. This result was because CN was a critical component in flood modeling due to its hydrological relevance, simplicity, integration with existing models, ability to incorporate detailed land and soil information, scalability, and historical validation. These factors collectively improved the accuracy and effectiveness of flood prediction and management efforts. Relating to this discussion, the finding was similar to the study by (Naemitabar et al., 2020).

By using KNN machine learning algorithm to assess flood risk triggered by land subsidence, new understanding was provided concerning the variables assumed to be factors causing the expansion of floods, particularly in areas experiencing land subsidence. In addition, high flood-risk conditions were found in areas with significant land subsidence rates, specifically in North Semarang, East Semarang, Gayamsari, and Genuk. The study findings supported flood hazard mapping presented by (Yuwono et al., 2024)

The study aimed to handle the uncertainty of the dataset and optimize the hyperparameters of the models, which had not been done in previous studies. The findings showed the importance of considering uncertainty and hyperparameter optimization in the modeling process and provided a more comprehensive understanding of the potential of MLMs. Moreover, effective flood management strategies were proposed, such as improving DEMs' accuracy and road conditions, constructing new urban areas at higher elevations, and strengthening underground drainage systems and road constructions in urban areas. Additionally, the exploration showed the requirement for a more comprehensive evaluation system that considered social and economic dimensions of coping capacities and resilience in future studies. The study in general offered a valuable understanding of flood risk mechanisms and proposed effective strategies to manage the risk.

Acknowledgment

The authors are grateful to grant funds RKAT FT UNDIP 2024 for funding the study.

References

Abidin, H. Z., Andreas, H., Gumilar, I., & Brinkman, J. J. (2015). Study on the risk and impacts of land subsidence in Jakarta. *Proceedings of the International Association of Hydrological Sciences*, *372*, 115–120. https://doi.org/10.5194/piahs-372-115-2015

Arabameri, A., Pradhan, B., & Rezaei, K. (2019). Gully erosion zonation mapping using integrated geographically weighted regression with certainty factor and random forest models in GIS. *Journal of Environmental Management*, *232*, 928–942. https://doi.org/10.1016/j.jenvman.2018.11.110

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Chen, J., Huang, G., & Chen, W. (2021). Towards better flood risk management: Assessing flood risk and investigating the potential mechanism based on machine learning models. *Journal of Environmental Management*, *293*, 112810. https://doi.org/10.1016/j.jenvman.2021.112810

Costache, R.-D., Arabameri, A., Costache, I., Crăciun, A., Islam, A. R. Md. T., Abba, S. I., Sahana, M., Pandey, M., Tin, T. T., & Thai Binh, P. (2022). Flood Hazard Potential Evaluation Using Decision Tree State-of-The-Art Models. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4028267

Deroliya, P., Ghosh, M., Mohanty, M. P., Ghosh, S., Rao, K. H. V. D., & Karmakar, S. (2022). A novel flood risk mapping approach with machine learning considering geomorphic and socio-economic vulnerability dimensions. *Science of The Total Environment*, *851*, 158002. https://doi.org/10.1016/j.scitotenv.2022.158002

Ebrahimy, H., Feizizadeh, B., Salmani, S., Azadi, H., 2020. A comparative study of land subsidence susceptibility mapping of Tasuj plane, Iran, using boosted regression tree, random forest and classification and regression tree methods. Environ. Earth Sci. 79, 223. https://doi.org/10.1007/s12665-020-08953-0

EO4SD. (2017). *Earth Observation for Sustainable Development*.

Gauhar, N., Das, S., & Moury, K. S. (2021). Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm. *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 357–361. https://doi.org/10.1109/ICREST51555.2021.9331199

Gharechaee, H., Nazari Samani, A., Sigaroodi, S.K., A. Hubbart, J., Moehammad Moein Sadeghi, S., 2023. Land Subsidence Susceptibility Mapping Using Interferometric Synthetic Aperture Radar (InSAR) and Machine Learning Models in a Semiarid Region of Iran. Land 12, 843–843. https://doi.org/10.3390/land12040843

Ghosh, A., & Dey, P. (2021). Flood Severity assessment of the coastal tract situated between Muriganga and Saptamukhi estuaries of Sundarban delta of India using Frequency Ratio (FR), Fuzzy Logic (FL), Logistic Regression (LR) and Random Forest (RF) models. *Regional Studies in Marine Science*, *42*, 101624. https://doi.org/10.1016/j.rsma.2021.101624

Ha-Mim, N. M., Rahman, Md. A., Hossain, Md. Z., Fariha, J. N., & Rahaman, K. R. (2022). Employing multi-criteria decision analysis and geospatial techniques to assess flood risks: A study of Barguna district in Bangladesh. *International Journal of Disaster Risk Reduction*, *77*, 103081. https://doi.org/10.1016/j.ijdrr.2022.103081

Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. *Science*, *349*(6245), 253–255. https://doi.org/10.1126/science.aac4520

Kuehn, F., Albiol, D., Cooksley, G., Duro, J., Granda, J., Haas, S., Hoffmann-Rothe, A., & Murdohardono, D. (2010). Detection of land subsidence in Semarang, Indonesia, using stable points network (SPN) technique. *Environmental Earth Sciences*, *60*(5), 909–921. https://doi.org/10.1007/s12665-009-0227-x.

Le, L., Xie, Y., & Raghavan, V. V. (2021). KNN Loss and Deep KNN. Fundamenta Informaticae, 182(2), 95–110. https://doi.org/10.3233/FI-2021-2068.

Li, G., Zhao, H., Liu, C., Wang, J., & Yang, F. (2022). City Flood Disaster Scenario Simulation Based on 1D–2D Coupled Rain–Flood Model. *Water*, *14*(21), 3548. https://doi.org/10.3390/w14213548

Li, S., Wang, Z., Lai, C., & Lin, G. (2020). Quantitative assessment of the relative impacts of climate change and human activity on flood susceptibility based on a cloud model. *Journal of Hydrology*, *588*, 125051. https://doi.org/10.1016/j.jhydrol.2020.125051

Miladan, N., Ariani, F., Pertiwi, S. N. I., Setiawan, R., & Handayani, K. N. (2019). Land Use Vulnerability towards the Flood Risk in Surakarta City. *MATEC Web of Conferences*, *280*, 01011. https://doi.org/10.1051/matecconf/201928001011.

Naemitabar, M., Zangeneh Asadi, M. A., Amirahmadi, A., & Goli Mokhtari, L. (2020). Evaluating and Zoning Flood Susceptibility Using Curve Number (CN) Logistic and Hydrological Regression Model (Case Study of Kalateh Qanbar Drainage Basin, Nishabur). https://doi.org/10.20944/preprints202012.0650.v1

Nadiri, A.A. et al. (2021) 'Mapping risk to land subsidence: developing a two-level modeling strategy by combining multi-criteria decision-making and artificial intelligence techniques', Water (Switzerland), 13(19). doi:10.3390/w13192622.

Prakash, A. J., Begum, S., Vilímek, V., Mudi, S., & Das, P. (2023). *Development of an Automated Method for Flood Inundation Monitoring, Flood Hazard and Soil Erosion Susceptibility Assessment Using Machine Learning and AHP-MCE Techniques*. https://doi.org/10.21203/rs.3.rs-3083674/v1

Soulis, K.X., 2021. Soil Conservation Service Curve Number (SCS-CN) Method: Current Applications, Remaining Challenges, and Future Perspectives. Water 13, 192. https://doi.org/10.3390/w13020192

Shafapour Tehrany, M., Shabani, F., Neamah Jebur, M., Hong, H., Chen, W., & Xie, X. (2017). GIS-based spatial prediction of flood prone areas using standalone frequency ratio, logistic regression, weight of evidence and their ensemble techniques. *Geomatics, Natural Hazards and Risk*, *8*(2), 1538–1561. https://doi.org/10.1080/19475705.2017.1362038.

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), 2311325. https://doi.org/10.1080/17538947.2024.2311325.

Salvati, A., Nia, A. M., Salajegheh, A., Ghaderi, K., Asl, D. T., Al-Ansari, N., Solaimani, F., & Clague, J. J. (2023). Flood susceptibility mapping using support vector regression and HYPER-PARAMETER optimization. *Journal of Flood Risk Management*, *16*(4), e12920. https://doi.org/10.1111/jfr3.12920

Tehrany, M. S., Jones, S., & Shabani, F. (2019). Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *CATENA*, *175*, 174–192. https://doi.org/10.1016/j.catena.2018.12.011

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*(8), 2464. https://doi.org/10.3390/s18082464

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, *527*, 1130–1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Yuwono, B., Awaluddin, M., & Najib. (2019). Land Subsidence monitoring 2016 - 2018 analysis using GNSS CORS UDIP and DinSAR in Semarang. *KnE Engineering*. https://doi.org/10.18502/keg.v4i3.5832

Yuwono, B.D., 2013, Analisa Penyebab dan Dampak Penurunan Muka Tanah ; Studi Kasus Kota Semarang, Tesis, Sekolah Paska Sarjana. Institut Teknologi Bandung.

Yuwono, B. D., Andreas, H., & Abidin, H. Z. (2021). *Assessing the Impact of Flood Induced by Sea Level Rise and Land Subsidence in Semarang City*. https://www.researchgate.net/publication/352934437

Yuwono, B.D., Abidin, H.Z., Poerbandono, Andreas, H., Pratama, A.S.P., Gradiyanto, F., 2024. Mapping of flood hazard induced by land subsidence in Semarang City, Indonesia, using hydraulic and spatial models. Nat. Hazards. https://doi.org/10.1007/s11069-023-06398-9

Zainuri, M., Helmi, M., Novita, M. G. A., Pancasakti Kusumaningrum, H., & Koch, M. (2022). An Improve Performance of Geospatial Model to Access the Tidal Flood Impact on Land Use by Evaluating Sea Level Rise and Land Subsidence Parameters. *Journal of Ecological Engineering*, *23*(2), 1–11. https://doi.org/10.12911/22998993/144785

Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019