Land Subsidence Susceptibility Mapping in Kakinada, India using Machine Learning Techniques

Nadimpalli, V. K.[1*] and Mahammood, V.[2]
[1]Department of Geo Engineering, Andhra University, Visakhapatnam, India
E-mail: vamsikrishna.rs@andhrauniversity.edu.in*
[2]Department of Civil Engineering, Andhra University, Visakhapatnam, India
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v21i8.4369

Abstract
Land subsidence poses significant risks to the infrastructure and ecosystems in coastal urban areas, making it a critical concern for regions like Kakinada, India, a rapidly developing coastal city. To address this problem in Kakinada, this research integrated Interferometry-based land subsidence inventory data with seven key conditioning factors: elevation, lithology, geomorphology, land use, Normalised Difference Vegetation Index (NDVI), drainage density, and distance to roads. Five machine learning models were employed, namely Extreme Gradient Boosting (XG Boost), Random Forest, Gradient Boosting Machine (GBM), Logistic Regression, and Multilayer Perceptron, for susceptibility mapping. Sixty per cent of the data was used for model training, 20% for hyperparameter tuning, and 20% for testing model accuracy. The models' ability to make accurate predictions was assessed through the area under the receiver operating characteristic curve (AUC-ROC) analysis. Results demonstrated that the GBM model achieved superior performance, with 82% accuracy and an AUC score of 0.88, while Random Forest showed comparable effectiveness. The analysis identified 50 square kilometres at high risk for future subsidence. The findings of this study will help urban planners identify the most vulnerable regions and implement strategies for sustainable development.

Keywords: InSAR, Kakinada, Land Subsidence, Machine Learning, Susceptibility Mapping

1. Introduction

Land subsidence, a major geohazard, is posing increasing risks to urban regions globally, leading to economic and environmental impacts. It occurs because of natural factors, such as tectonic activity, soil compaction, and sediment consolidation, as well as anthropogenic causes, including excessive groundwater extraction, mining, urbanisation, and heavy infrastructure loads [1]. It was estimated that nearly 6.3 million square kilometres of land surface have been experiencing subsidence [2]. It has been observed in numerous cities worldwide, including Beijing, Jakarta, London, and New Delhi [3][4] and [5].

In countries such as India, the demand for land has increased due to population growth and urban expansion. This is particularly prevalent in coastal areas, where natural ecosystems such as mudflats and mangrove forests are converted into residential and industrial spaces, thereby accelerating subsidence-related issues by altering the soil structure and increasing ground load. In coastal areas, land subsidence further increases the flood risk and land loss, making urban planning and hazard assessment even more critical [6]. A combination of conditioning factors influences the susceptibility to subsidence in any region, each having its own influence on subsidence.

Geological factors, such as lithology, play a significant role, as unconsolidated sediments like clay and silt are more susceptible to compaction than rock formations. Hydrological factors, such as groundwater decline, are often cited as a major driver of subsidence, as the loss of pore pressure leads to sediment compaction. Elevation and slope influence surface runoff and soil moisture retention, thereby affecting consolidation rates. Anthropogenic factors like land use/land cover (LULC) significantly impact susceptibility, with urban and industrial areas typically experiencing higher subsidence due to the combined effects of heavy infrastructure loads. Other factors commonly considered in susceptibility modelling include soil properties, proximity to river networks, and the distance from tectonic faults. Although the latest technologies, such as GPS [7] and Interferometric SAR methods [8] and [9], have simplified the detection of subsidence, accurately

modelling susceptibility remains challenging due to the nonlinear interactions between geological, hydrological, and human-induced variables. Researchers have utilised statistical and probabilistic tools to map the susceptibility of various geohazards, including subsidence, landslides, and floods. These include the Analytic Hierarchy Process (AHP) [10], Multi-Criteria Decision-Making (MCDM) [11], Evidential Belief Function (EBF) models [12], Frequency Ratio (FR), and Fuzzy Logic (FL) methods [13]. The major limitation of these methods is the assumption of linear relationships between variables that often simplifies subsidence dynamics, particularly in heterogeneous coastal environments where groundwater withdrawal, surface loading, and sediment characteristics are involved. Machine learning (ML) techniques for land subsidence susceptibility mapping have gained significant momentum due to their capability to analyse non-linear relationships among the involved variables.

Multiple studies have shown the reliability of machine-learning algorithms in predicting areas at risk of land subsidence. A comprehensive study employed a random forest algorithm to create a high-resolution global land subsidence map, trained with InSAR and GNSS-based coastal subsidence data [14]. Land subsidence in South Korea's Jeong-am area was mapped and assessed by [15]. They employed four machine-learning algorithms, and results indicated that the BLR model performed best, as validated by the AUROC curve and other statistical indices. In another study, researchers mapped subsidence susceptibility in Iran's Semnan Plain using a Support Vector Machine (SVM) and weight of evidence (WOE) models, using 12 conditioning factors, such as groundwater table, topography, and land use. Their results indicated that the SVM model (AUC = 0.748) outperformed the WOE model (AUC = 0.726) in predicting susceptible areas, highlighting groundwater overexploitation as a primary driver of subsidence [16]. In contrast, [17] focussed on Jakarta, employing Logistic Regression, Multilayer Perceptron, AdaBoost, and LogitBoost to map susceptibility using InSAR-derived ground deformation data. Among these, AdaBoost produced the most accurate susceptibility maps. In Fars province, researchers compared Logistic Regression (LR), RF, Boosting Regression Tree (BRT), and SVM models, finding that RF and SVM models achieved very high accuracy [18]. The effectiveness of boosting algorithms is further supported by a study in Ca Mau Province, Vietnam, showing that AdaBoost, Gradient Boosting, and XG Boost all achieved high accuracy, with XGBoost reaching an AUC greater than 0.88 [19].

Recent advances highlight hybrid and ensemble learning architectures. Metaheuristic-optimised neural networks (e.g., Water Cycle Algorithm-ANN with AUC 0.974) significantly outperform standalone models in Iran's Rokn Abad Plain [20]. Similarly, hybrid decision stump classifiers coupled with Alternating Decision Trees (DSC-ADTree) achieved 98.3% AUC in recent case studies by resolving class imbalances in subsidence training data [21]. These findings highlight the capability of machine learning in subsidence prediction across diverse geographical contexts. However, there remains a need for comprehensive comparative analyses to determine which algorithms perform best under different urban conditions and with data availability constraints. Despite global advancements in land subsidence research, a significant knowledge gap remains regarding susceptibility mapping in many developing coastal cities, particularly in India. In the case of Kakinada, a rapidly urbanizing coastal city, some studies have already identified ongoing land subsidence phenomena. However, studies that integrate conditioning factors and subsidence for predictive modelling have not been conducted. The region's characteristics, such as its flat topography, unconsolidated alluvial and deltaic sediments, and increasing anthropogenic stress, make it particularly vulnerable to subsidence. Therefore, there is a crucial need for machine learning based studies that can more effectively model the complex interactions among factors influencing land subsidence.

This research aims to develop a land subsidence susceptibility map for Kakinada by integrating Interferometry-derived deformation data with geo-environmental and anthropogenic conditioning factors. This study seeks to employ and compare five machine learning algorithms, Logistic Regression, Random Forest, XGBoost, Gradient Boosting, and Multi-Layer Perceptron, in accurately modeling the susceptibility to land subsidence. By evaluating their performance through AUC-ROC analysis, the study aims to identify the most suitable algorithm for this coastal region. The resulting susceptibility map will assist urban planners and policymakers in identifying vulnerable areas and developing strategies for sustainable land use and infrastructure development. Logistic Regression was selected as a well-established baseline model, commonly used in geohazard studies for binary classification tasks due to its simplicity and interpretability. Multi-Layer Perceptron, an artificial neural network, was selected for its ability to identify non-linear relationships such as the relationship between subsidence and conditional factors. Random Forest was included for its ensemble learning method that handles high-dimensional data effectively and is also reliable in

capturing nonlinearity and reducing overfitting. It has been proven in many susceptibility assessment studies. XGBoost and Gradient Boosting were chosen for their ability to improve predictions by learning from the errors of earlier trees and their capability to capture subtle patterns. Their ability to generalize well across complex terrains makes them highly suitable for spatial susceptibility mapping.

2. Study Area

The study area was Kakinada City in the state of Andhra Pradesh, India, between 82° 8′ 52″ N and 82° 16′ 55″ N latitude and 16° 57′ 41″–17° 4′ 16″ E longitude, spanning over an area of 160 sq. km. The region has a maximum elevation of only 12m, with the majority of the areas under 5m having an almost flat topography. The region is primarily composed of recent alluvial sediments, including silt, clay, and fine sand deposited by the Godavari River over time. These unconsolidated sediments are characteristic of deltaic environments and are particularly susceptible to compaction under natural and anthropogenic pressures. In coastal fringes, wetlands comprising mudflats and mangroves are present. Land use in Kakinada is a mix of urban, industrial, agricultural, and aquaculture zones. Approximately one-third of the city's area is settlements, while the peripheral regions are dominated by paddy cultivation and aquaculture activities. Rapid urbanization, especially after the bifurcation of Andhra Pradesh in 2014, has led to significant expansion of industrial areas, particularly related to oil and natural gas due to the city's proximity to the Krishna-Godavari basin. These land use changes intensify the loading on underlying soft soils, especially those rich in clay and silt, further making Kakinada vulnerable to ground settlement. Recent studies have confirmed that the region has subsided at a rate of 10 mm/yr [22] and [23]. The region's strategic importance as an industrial and port city, combined with its low-lying topography (0–5 m elevation), creates acute infrastructure vulnerability where even a small amount of subsidence could cause saltwater intrusion and storm surge impacts. This highlights the need to identify areas that may be susceptible to subsidence in the future.

3. Data Collection and Preparation

3.1 Land Subsidence Inventory

A land subsidence database is essential for training any machine learning model for susceptibility prediction. Due to the lack of prior field-based subsidence observations for the study area, the inventory was extracted from land deformation maps prepared using the PS-InSAR technique for the period 2020-2024. Given its proven reliability, this inventory can be used as a reliable dataset for this study. The points that exhibited displacements greater than 10 mm/year were considered subsidence points, and those with displacements of ± 2 mm/year were selected as non-subsidence points. Due to the presence of a high density of PS points, random selection was performed to choose a total of 1000 points, with 500 each of subsidence and non-subsidence points, which were distributed well in the region, as displayed in Figure 1.

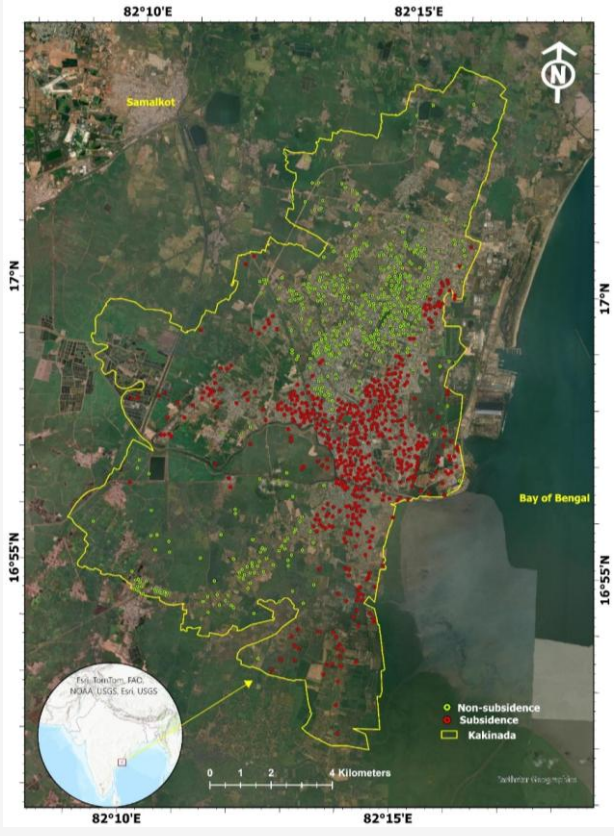

Figure 1: Kakinada City, Andhra Pradesh, India and subsidence inventory points

3.2 Land Subsidence Conditioning Factors

Choosing conditional factors is crucial for susceptibility mapping because they influence the extent and magnitude of subsidence. These factors include topographic, geological, hydrological, and anthropogenic conditions. Topographic parameters have been widely recognized as crucial influencing factors in land subsidence susceptibility mapping. Elevation, curvature, aspect, slope and topographic wetness index are commonly employed. These factors directly influence subsidence by affecting the stability of the terrain. Hydrological and hydrogeological factors like groundwater drawdown, distance to stream (DtS), drainage density, hydraulic conductivity, and specific yield are also identified as

influencing factors in some studies. Lithology, soil thickness, distance to fault (DtF), lineament density, and distance to lineament have been incorporated in some studies. Human activities also significantly impact subsidence patterns. Land cover patterns, distance to road (DtR), and normalised difference vegetation index (NDVI) were also considered in susceptibility models [24][25][26] and [27]. Overall, groundwater-related parameters emerge as the most influential factors across global studies, followed by topographic, geological, and land-use characteristics. However, the selection of influential factors for land subsidence susceptibility mapping depends on regional characteristics and the availability of data. The maps of condition factors employed in this study are shown in the Figure 2.

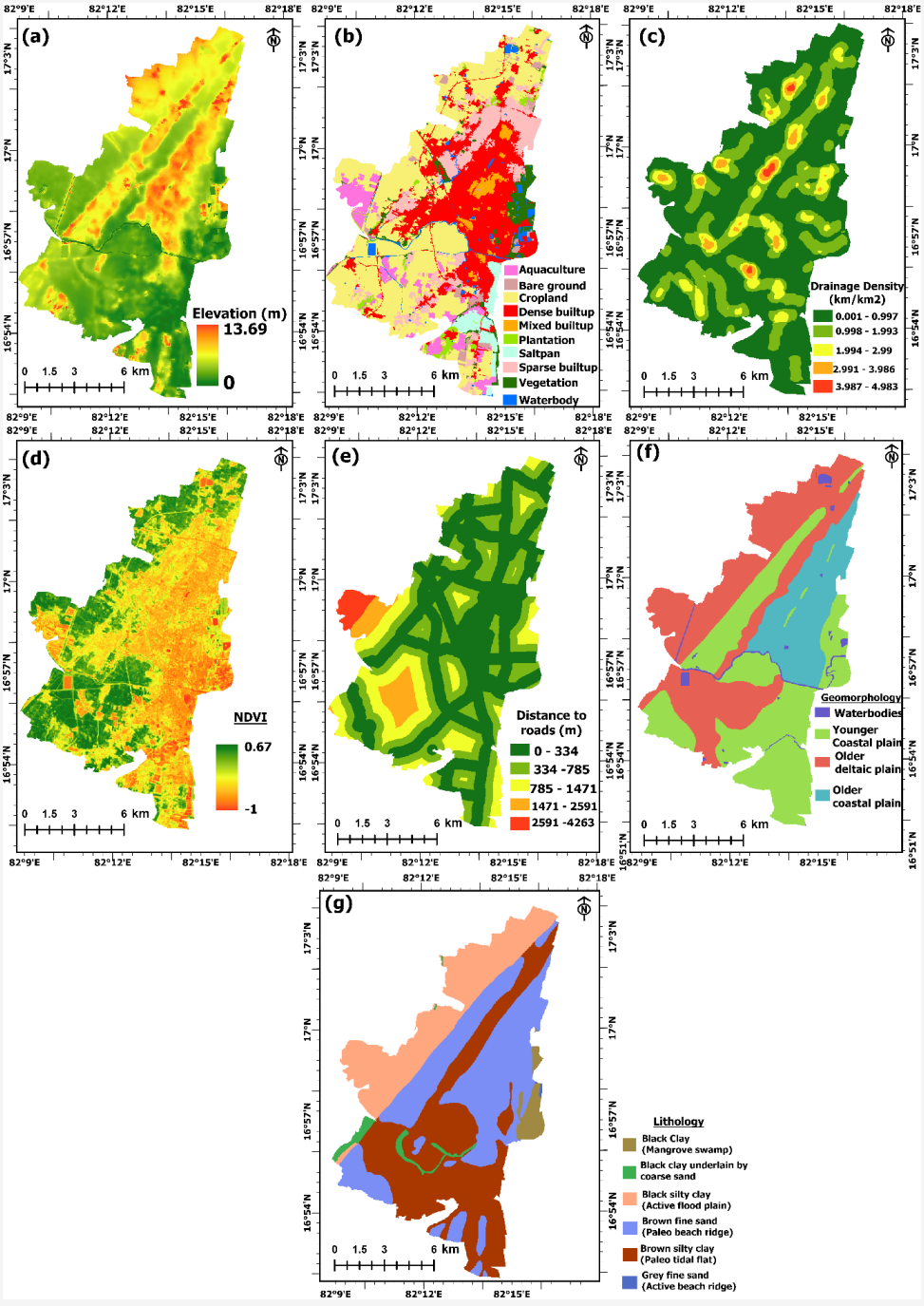

Figure 2: Conditioning factors: (a) elevation (b) land use (c) drainage density (d) NDVI (e) distance to roads (f) geomorphology and (g)lithology

Through careful examination of the study region's characteristics, particularly its low elevation difference, and considering data availability, seven influential factors were identified. The Elevation layer was extracted from FABDEM with a 30 m resolution. It is an improved version of the Copernicus Global 30m DEM (GLO-30), where vegetation and building artefacts are removed [28]. The region exhibits minimal topographic variation, with most areas having elevations below 5 meters above sea level. The Normalised Difference Vegetation Index (NDVI) provides information about vegetation that plays an important role in ground stability, particularly in coastal deltas. The NDVI was derived from Sentinel-2 imagery at a 10 m resolution for the study area. A land use map was prepared using high-resolution data from Planet Labs using object-based image analysis. Geomorphology and Lithology data were taken from 1:50000 maps provided by the Geological Survey of India.

Drainage density was prepared from the FABDEM elevation layer, and finally, distance to roads was extracted from the road data provided by OpenStreetMap (OSM) in the ArcGIS environment.

4. Methodology

The methodology of the study is illustrated in the Figure 3. It includes the following: (i) collection and preparation of land subsidence conditional factors and resampling them to uniform grids for spatial consistency. (ii) Collection and separation of the land subsidence inventory, including splitting into training, validation, and test data. (iii) Model Training (iv) Hyperparameter tuning of the models using the validation data and selection of optimal parameters. (v) Generation of subsidence susceptibility maps with models. (vi) Finally, all models were tested using test set data and multiple performance metrics like sensitivity, specificity, and AU-ROC score. Initially, thematic layers corresponding to the selected land subsidence conditioning factors were collected from various sources. All layers were resampled to 30-meter spatial resolution. Utilising GIS spatial analysis tools, conditional factor values were extracted for all points in the training dataset. The resultant dataset included sample ID, corresponding conditioning factor values, and a binary class label indicating whether the sample was subsiding or not. The dataset was then split into training (60%), validation (20%), and testing (20%) sets.

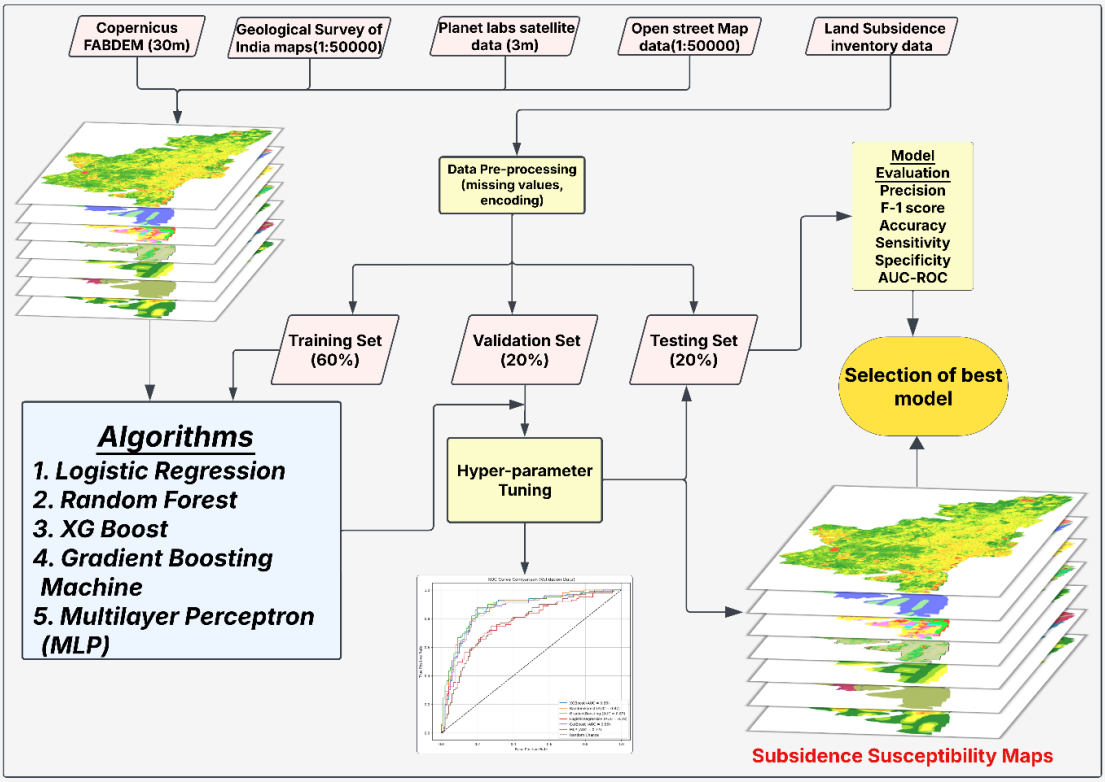

Figure 3: Subsidence susceptibility study workflow

The training data was used to develop the models, the validation data for tuning hyperparameters and selecting optimal configurations, and the test set was exclusively for final evaluation. All models were trained using the training dataset and optimised using grid search-based hyperparameter tuning guided by performance on the validation set. Subsidence susceptibility maps were created using the top-performing configuration of each model by applying the trained models across the entire study area. Each grid cell was assigned a probability value between 0 and 1, indicating the likelihood of being prone to land subsidence. These values were then classified into susceptibility zones (e.g., low, moderate, high) based on a quantile classification method. Finally, the predictive performance of each model was assessed using the independent test dataset. The model with the highest AU-ROC score and balanced sensitivity-specificity trade-off was selected as the most effective for creating the final susceptibility map.

4.1 Machine Learning Algorithms

4.1.1 Logistic regression

Logistic regression is a supervised learning algorithm commonly employed for binary classification problems. It estimates the likelihood of an outcome belonging to a particular class using the logistic (sigmoid) function [29]. Equation is used to estimates the probability that y belongs to a particular category given inputs $X = (x_1, x_2, ..., x_k)$:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i X_i)}}$$

Equation 1

Where β_0 is the intercept, β_i is the coefficient, and X_i are input features.

4.1.2 Multi-layer perceptron (MLP)

Multilayer Perceptron (MLP) is a kind of artificial neural network made up of fully connected neurons that use nonlinear activation functions and are arranged in layers. It comprises one input layer, one or more hidden layers, and an output layer. MLPs are trained using supervised learning via backpropagation [30]. For a feedforward MLP with L layers, the output of the l^{th} layer can be represented by Equation 2:

$$a^{(l)} = \sigma(W^{(l)} a^{(l-1)} + b^{(l)})$$

Equation 2

Where $W^{(l)}$ is the weight matrix for layer l, $b^{(l)}$ is the bias vector, $a^{(l-1)}$ is the activation from the previous layer, and σ is the activation function. The final prediction is given by the output of the last layer $a^{(l)}$. MLPs are capable of modelling complex non-linear relationships and have been effectively used to various classification and regression problems, including subsidence susceptibility mapping.

4.1.3 Random Forest

Random forest is an ensemble learning technique that constructs multiple decision trees and combines their results to improve accuracy and reduce overfitting. This algorithm creates a set of uncorrelated trees by introducing randomness in both the observation selection (through bootstrapping) and feature selection at each split [31] and [32]. For a given input, each decision tree in the forest produces a prediction, and the final result is determined by majority voting in classification and averaging in regression tasks. The algorithm can be expressed in Equation 3:

$$F(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$$

Equation 3

Where $F(x)$ is the final prediction, B is the number of trees, and $f_b(x)$ is the prediction of the b^{th} tree.

4.1.4 Gradient boosting machine (GBM)

Gradient boosting constructs an ensemble of decision trees in a stage-wise manner to create a strong predictive model. Unlike random forests, which build trees separately, gradient boosting builds trees sequentially, with each tree correcting the errors of the previous ensemble [33]. The GBM algorithm is expressed in Equation 3:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

Equation 4

Where $h_m(x)$ is the *m-th* weak learner and γ_m is its weight. The model is built sequentially by adding new trees that minimize a loss function L. At each step, the algorithm fits $h_m(x)$ to the negative gradient of the loss function, which is defined in Equation 5:

$$h_m(x) \approx -\frac{\partial L\left(y, F_{m-1}(x)\right)}{\partial F_{m-1}(x)}$$

Equation 5

This approach allows gradient boosting to optimize any differentiable loss function, making it highly versatile for different types of problems.

4.1.5 XGBoost

XGBoost is an advanced form of gradient boosting. It extends the traditional gradient boosting framework with regularisation and a more sophisticated tree-building algorithm, which significantly improves computational efficiency and predictive performance [34]. The XGBoost merges a differentiable loss function with regularisation term as presented in Equation 6:

$$L = \sum_{i=}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} W(f_k)$$

Equation 6

Where l is the loss function, Ω is the regularisation term, and f_k represents the k^{th} tree. The regularisation term is defined in Equation 7:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|W\|^2$$

Equation 7

Where T is the number of leaves, w represents the vector of leaf weights, γ controls the number of leaves, and λ is the L2 regularization parameter on weights. Second-order Taylor expansions approximate the loss function expressed in Equation 8:

$$\hat{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

Equation 8

Where g_i and h_i are first and second derivatives of loss.

5. Results

5.1 Land Subsidence Susceptibility Maps

The data consisted of 2000 PSI points derived from the land deformation dataset, which were then labelled as 1(subsidence) and 0 (non-subsidence). Conditional factor values were extracted using GIS for all points in the inventory. The data were then divided into training, validation, and test sets, as they are essential for evaluating the model's performance. After fine-tuning hyperparameters to optimise model performance, we applied these refined models to generate land subsidence susceptibility maps. Table 1 summarises the optimal hyperparameters identified for all five models after the tuning process. The maps were classified into five categories of risk: very low, low, moderate, high, and very high. The subsidence susceptibility maps generated using all five models are shown in Figure 4. A comparative analysis of the generated maps revealed distinct patterns among different models. The susceptibility maps produced by RF, XGB, and GBM showed a high degree of similarity, identifying the same regions as highly susceptible to land subsidence. By contrast, the maps generated using MLP and Logistic Regression demonstrated a different spatial distribution of susceptibility, with noticeable deviations from the other models. The results of this class-wise area distribution are presented in Table 2. These variations suggest that tree-based ensemble models (RF, XGB, and GBM) similarly capture subsidence-prone regions, likely due to their ability to capture complex feature interactions.

Table 1: Optimal parameters for algorithms selected using the validation data

Algorithm	Parameters
Logistic Regression	C: 1
MLP	alpha: 0.1; hidden layer sizes: (50, 50); learning rate: 0.001.
Random Forest	Max_depth: 10; min samples leaf: 1; min samples split: 2; N_estimators: 200.
Gradient Boost Machine	Learning_rate: 0.1; max_depth: 5; n_estimators: 100; subsample: 1.0.
XG Boost	learning_rate: 0.05; max_depth: 5; n_estimators: 200; subsample: 1.0.

Table 2: Risk susceptibility classification results (percentage area)

Algorithm	Very Low	Low	Medium	High	Very High
Logistic Regression	10.23	19.68	27.72	25.75	16.59
Multilayer Perceptron	18.64	22.26	20.62	19.55	18.90
Random Forest	14.05	21.54	25.90	23.43	15.04
Gradient Boost Model	20.04	20.31	19.20	20.22	20.23
XG Boost	31.05	13.66	11.10	13.35	30.83

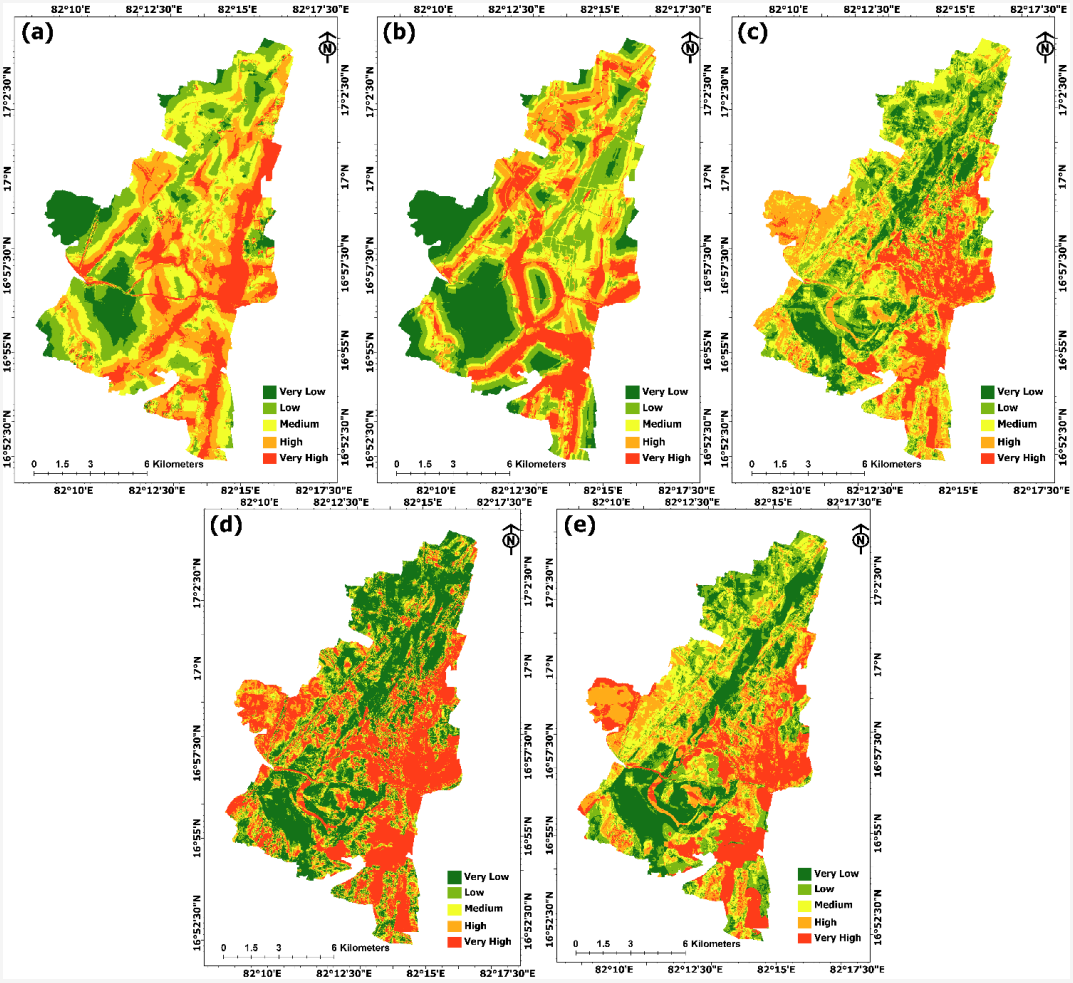

Figure 4: The land subsidence susceptibility maps for Kakinada: (a) logistic regression (b) multilayer perceptron (c) random forest (d) xgboost, and (e) gradient boosting machine

5.2 Model Validation

The accuracy and reliability of trained models are crucial for susceptibility mapping. In this study, the model's performance is evaluated using various metrics, including the Area Under the Receiver Operating Characteristic Curve (AU-ROC), Sensitivity, Specificity, Accuracy, and F1 score. A higher AU-ROC value indicates a better-performing model that can effectively differentiate positive (subsidence) and negative (non-subsidence) cases. The formulae for the metrics are presented in Equations 9 to 10:

$$Sensitivity = \frac{TP}{TP + FN}$$

Equation 8

$$Specificity = \frac{TN}{TN + FP}$$

Equation 9

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 10

Where *TP*, *TN*, *FP*, and *FN* are true positives, true negatives, false positives, and false negatives, respectively. The values calculated for the above metrics for all five trained models are shown in Table 3. The evaluation of the five models reveals significant differences in their ability to predict land subsidence susceptibility in Kakinada. As a benchmark linear model, the LR algorithm achieved an AUC of 0.76 and 69.5% overall Accuracy. It was outperformed by the more complex, non-linear models across all metrics. It also had the highest number of False Negatives (*FN* = 26), indicating a lower ability to identify actual subsidence cases when compared to other models.

Table 3: Evaluation metrics of models on the test dataset

Model	Accuracy	Precision	Sensitivity	F1-score	Specificity	AUC
Logistic Regression	0.695	0.650	0.714	0.680	0.678	0.76
MLP	0.715	0.642	0.846	0.729	0.605	0.79
Random Forest	0.815	0.776	0.835	0.804	0.798	0.88
XG Boost	0.790	0.747	0.813	0.779	0.770	0.88
GBM	0.820	0.784	0.835	0.809	0.807	0.88

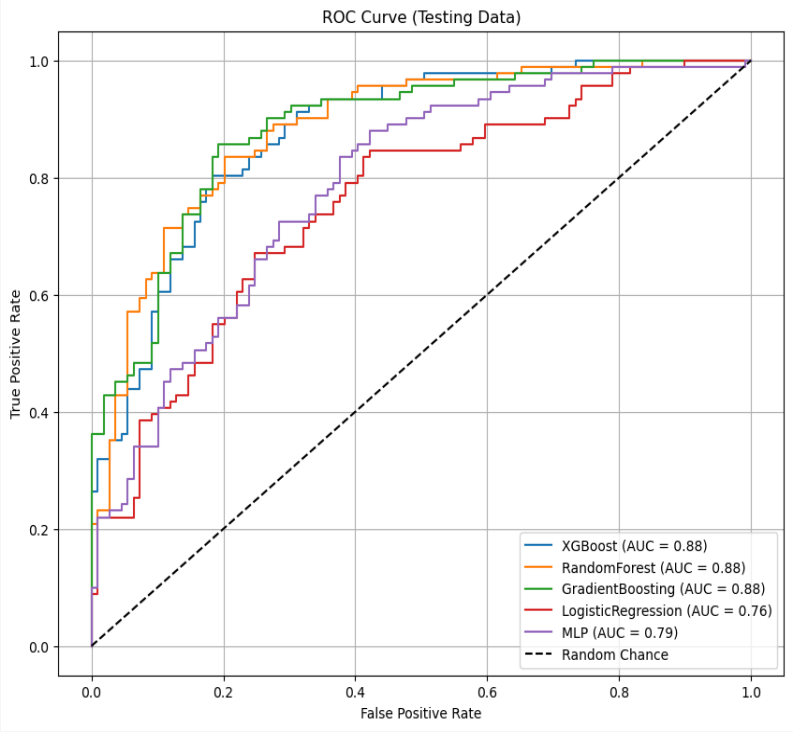

Figure 5: ROC curves and AUC metrics for test dataset

The MLP, a type of artificial neural network, showed improved performance over Logistic Regression with an AUC of 0.79. It achieved the highest Sensitivity (0.846) of all models, meaning it was very effective at identifying actual subsidence pixels (low FN=14). However, it also had the lowest Specificity (0.605) and the highest number of False Positives (FP=43), suggesting that the model was over-predicting subsidence risk in stable areas. Random Forest achieved an accuracy of 81.5%. With 76 true positives and 87 true negatives, this model showed a good balance between sensitivity (0.835) and specificity (0.798). The strong performance of Random Forest aligns with previous studies that have found this algorithm effective for land subsidence mapping due to its capacity to identify non-linear relationships between multiple variables [18]. XG Boost showed an accuracy of 79.0% and an AUC of 0.88. While slightly lower in accuracy than Random Forest, it maintained a good balance between sensitivity (0.813) and specificity (0.770). This performance is consistent with findings from other studies that highlight XGBoost's effectiveness in predicting complex geospatial phenomena[35]. The GBM model emerged as the top-performing algorithm in this study. It achieved the highest Accuracy (82.0%), the highest F1-score (0.809), and the highest Specificity (0.807). It scored the highest AUC value of 0.88 with RF and XGBoost. The combination of high sensitivity and the highest specificity indicates that the GBM was not only excellent at identifying true subsidence zones but was also the best at correctly classifying stable zones, thus minimising false alarms. The AUROC scores for all five models are presented in the Figure 5.

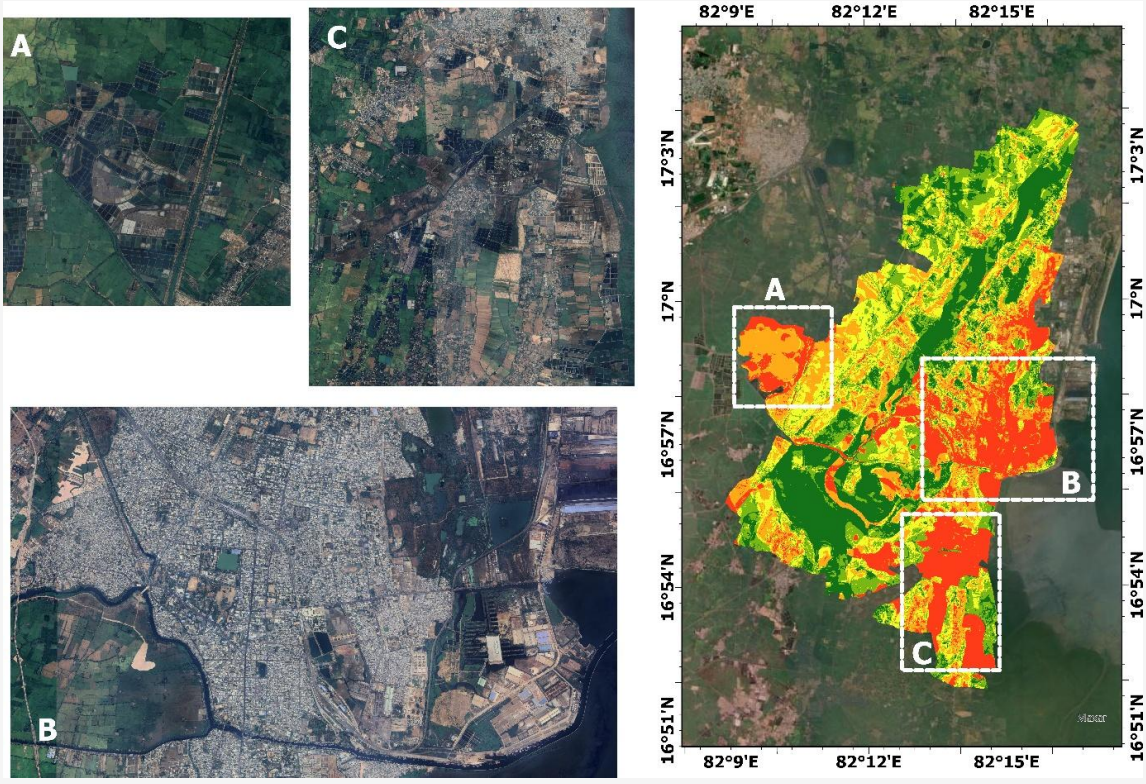

Figure 6: High susceptibility zones predicted by the GBM model

6. Discussion

The susceptibility maps revealed distinct spatial patterns of subsidence risk across Kakinada. Three areas emerged as particularly vulnerable: the densely populated urban core of Kakinada city, and two peripheral regions characterized by intensive aquaculture practices and salt pan operations were classified as high-risk zones. These findings aligned with a study by [19] where high susceptibility was concentrated in urban and aquaculture areas in a similar deltaic setting, suggesting that intensive loads and groundwater extraction activities increase vulnerability to subsidence. Figure 6 displays this susceptibility map, overlaid on Google Earth imagery, highlighting regions where the model predicts a higher risk. The city centre of Kakinada emerged as a primary high-risk zone in all models, particularly in the GBM and Random Forest predictions. This area is characterized by dense urban development, significant groundwater extraction, and coastal plain geomorphology with unconsolidated sediments. The concentration of buildings and infrastructure creates loading pressure on these underlying sediments, potentially accelerating compaction processes. This finding aligns with the PS-InSAR data, which showed significant displacement in this region. The port area, which falls within the high-risk zone, presents a particularly interesting case. Many structures in this area were constructed on the land reclaimed from mangrove ecosystems. The combination of unconsolidated reclaimed soil, removal of natural buffers (mangroves), and heavy infrastructure makes this area vulnerable.

The other two high-risk regions, Achuthapuram to the west and Gurajanapalli to the south, are dominated by aquaculture and salt production activities. The continuous extraction of groundwater to maintain brackish water conditions in aquaculture ponds appears to be a significant driver of subsidence in these areas. This pattern was consistently identified across all tree-based models, indicating a robust relationship between aquaculture practices and subsidence susceptibility. Around 45 sq. km. was classified as having low susceptibility by both GBM and Random Forest models. These regions are primarily covered by agriculture, natural vegetation, and low-density built-up areas north of the city. The low susceptibility classification of these areas aligns with their stable signature in the PS-InSAR data. The better performance of ensemble models in the present study also aligned with [17], where AdaBoost (an ensemble method) outperformed Logistic Regression and MLP models. By using a displacement threshold of 10 mm/year to identify subsidence points, a

quantitative and consistent definition of subsidence was established over the study area. This classification is valuable in a region like Kakinada, where systematic field-based subsidence monitoring has been limited. Past studies have confirmed that the region has subsided at a pace of approximately 10 mm/yr, validating our threshold selection.

The results demonstrate that machine-learning susceptibility mapping can reliably predict vulnerable zones in coastal delta cities, such as Kakinada. To our knowledge, this is the first study to apply machine-learning susceptibility modelling for land subsidence in India. Many delta cities and coastal areas in India are at risk due to rapid urbanisation, excessive groundwater use, and climate change. Tools like this can help planners and decision-makers take action to protect these vulnerable areas.

7. Conclusion

This study successfully developed a land subsidence susceptibility map for Kakinada by integrating PS-InSAR data with multiple conditioning factors (NDVI, geomorphology, lithology, LULC, distance to roads, and drainage density) through machine learning models. The analysis revealed varying degrees of subsidence susceptibility across the study area, with notable patterns correlating to specific geological and anthropogenic features. Tree-based ensemble models, particularly GBM and Random Forest, performed well over other algorithms, achieving high accuracy and AUC scores. GBM emerged as the best model with of 82% accuracy and an AUC score of 0.88, closely followed by Random Forest. The study identified approximately 50 sq.km. area as high risk, primarily in the urban core and two peripheral regions characterised by intensive aquaculture and salt pans. These findings can be integrated into comprehensive land use planning and environmental impact assessments for sustainable urban development and hazard mitigation in the Kakinada region. While acknowledging its valuable contributions, the study also indicates that future research could be improved by incorporating additional conditioning factors, such as groundwater levels and soil properties.

References

[1] Pedretti, L., Giarola, A., Korff, M., Lambert, J. and Meisina, C., (2024). Comprehensive Database of Land Subsidence in 143 Major Coastal Cities Around the World: Overview of Issues, Causes, and Future Challenges. *Frontiers in Earth Science*, Vol. 12. 1–28. https://doi.org/10.3389/feart.2024.1351581.

[2] Davydzenka, T., Tahmasebi, P. and Shokri, N., (2024). Unveiling the Global Extent of Land Subsidence: The Sinking Crisis. *Geophysical Research Letters*, Vol. 51(4), 1–11. https://doi.org/10.1029/2023GL104497.

[3] Agarwal, V., Kumar, A., Gee, D., Grebby, S., Gomes, R. L. and Marsh, S., (2021). Comparative Study of Groundwater-Induced Subsidence for London and Delhi Using PSInSAR. *Remote Sensing*, Vol. 13(23). https://doi.org/10.3390/rs13234741.

[4] Hu, B., Wang, H. S., Sun, Y. L., Hou, J. G. and Liang, J., (2014). Long-Term Land Subsidence Monitoring of Beijing (China) Using the Small Baseline Subset (SBAS) Technique. *Remote Sensing*, Vol. 6(5), 3648–3661. https://doi.org/10.3390/rs6053648.

[5] Ng, A. H. M., Ge, L., Li, X., Abidin, H. Z., Andreas, H. and Zhang, K., (2012). Mapping Land Subsidence in Jakarta, Indonesia Using Persistent Scatterer Interferometry (PSI) Technique with ALOS PALSAR. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 18(1), 232–242. https://doi.org/10.1016/j.jag.2012.01.018.

[6] Gambolati, G., Teatini, P. and Ferronato, M., (2005). Anthropogenic Land Subsidence. *Encyclopedia of Hydrological Sciences*. https://doi.org/10.1002/0470848944.hsa164b.

[7] Baldi, P., Casula, G., Cenni, N., Loddo, F. and Pesci, A., (2009). GPS-Based Monitoring of Land Subsidence in the Po Plain (Northern Italy). *Earth and Planetary Science Letters*, Vol. 288(1–2), 204–212. https://doi.org/10.1016/j.epsl.2009.09.023.

[8] Berardino, P., Fornaro, G., Lanari, R. and Sansosti, E., (2002). A New Algorithm for Surface Deformation Monitoring Based on Small Baseline Differential SAR Interferograms. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40(11), 2375–2383. https://doi.org/10.1109/TGRS.2002.803792.

[9] Ferretti, A., Prati, C. and Rocca, F., (2000). Nonlinear Subsidence Rate Estimation Using Permanent Scatterers in Differential SAR Interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 38(5 I), 2202–2212. https://doi.org/10.1109/36.868878.

[10] Thammaboribal, P., Triapthti, N., and Lipiloet, S. (2025). Using of Analytical Hierarchy Process (AHP) in Disaster Management: A Review of Flooding and Landslide Susceptibility Mapping. *International Journal of Geoinformatics*, Vol. 21(4), 177–196. https://doi.org/10.52939/ijg.v21i4.4091..

[11] Akay, H., (2021). Flood Hazards Susceptibility Mapping Using Statistical, Fuzzy Logic, and MCDM Methods. *Soft Computing*, Vol. 25(14), 9325–9346. https://doi.org/10.1007/s00500-021-05903-1.

[12] Phetprayoon, T. (2025). Flood Susceptibility Mapping using GIS and Evidential Belief Function Model in Lam Chiang Krai Watershed, Nakhon Ratchasima Province, Thailand. *International Journal of Geoinformatics*, Vol. 21(6), 29–46. https://doi.org/10.52939/ijg.v21i6.4231.

[13] Ali, I., Khatibi, B., and Karimzadeh, S. (2024). A Comparative Study of Groundwater Recharge Mapping Using Analytical Hierarchy Process, Fuzzy-Analytical Hierarchy Process, and Frequency Ratio Models: A Case Study from Quetta Region, Pakistan. International Journal of Geoinformatics, Vol. 20(7), 111–133. https://doi.org/10.52939/ijg.v20i7.3411.

[14] Hasan, M. F., Smith, R., Vajedian, S., Pommerenke, R. and Majumdar, S., (2023). Global Land Subsidence Mapping Reveals Widespread Loss of Aquifer Storage Capacity. *Nature Communications,* Vol. 14(1). https://doi.org/10.1038/s41467-023-41933-z.

[15] Bui, D. T., Shahabi, H., Shirzadi, A., Chapi, K., Pradhan, B., Chen, W., Khosravi, K., Panahi, M., Ahmad, B. B., and Saro, L., (2018). Land Subsidence Susceptibility Mapping in South Korea Using Machine Learning Algorithms. *Sensors (Switzerland)*, Vol. 18(8), 2464. https://doi.org/10.3390/s18082464.

[16] Mohammady, M., Pourghasemi, H. R. and Amiri, M., (2019). Assessment of Land Subsidence Susceptibility in Semnan Plain (Iran): A Comparison of Support Vector Machine and Weights of Evidence Data Mining Algorithms. *Natural Hazards*, Vol. 99(2), 951–971. https://doi.org/10.1007/s11069-019-03785-z.

[17] Hakim, W. L., Achmad, A. R. and 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*, Vol. 12(21), 1–26. https://doi.org/10.3390/rs12213627.

[18] Sekkeravani, M. A., Bazrafshan, O., Pourghasemi, H. R. and Holisaz, A., (2022). Spatial Modeling of Land Subsidence Using Machine Learning Models and Statistical Methods. *Environmental Science and Pollution Research International*, Vol. 29(19), 28866–28883. https://doi.org/10.1007/s11356-021-18037-6.

[19] Tran, A. V., Brovelli, M. A., Ha, K. T., Khuc, D. T., Tran, D. N., Tran, H. H. and Le, N. T., (2024). Land Subsidence Susceptibility Mapping in Ca Mau Province, Vietnam, Using Boosting Models. *ISPRS International Journal of Geo-Information*, Vol. 13(5). https://doi.org/10.3390/ijgi13050161.

[20] Yu, H., Arabameri, A., Costache, R., Crăciun, A. and Arora, A., (2022). Land Subsidence Susceptibility Assessment Using Advanced Artificial Intelligence Models. *Geocarto International*, Vol. 37(27), 18067–18093. https://doi.org/10.1080/10106049.2022.2136265.

[21] Zhao, R., Arabameri, A. and Santosh, M., (2024). Land Subsidence Susceptibility Mapping: A New Approach to Improve Decision Stump Classification (DSC) Performance and Combine it with Four Machine Learning Algorithms. *Environmental Science and Pollution Research International*, Vol. 31(10), 15443–15466. https://doi.org/10.1007/s11356-024-32075-w.

[22] Kumar, L. R., Vaka, D. S. and Rao, Y. S., (2017). Mapping Land Subsidence of Krishna - Godavari Basin Using Persistent Scatterer Interferometry Technique. *38th Asian Conference on Remote Sensing - Space Applications: Touching Human Lives, ACRS 2017*, *2017-Octob*.

[23] Reshma, K. N., Agrawal, R., Ramakrishnan, R., Sreejith, K. M. and Rajawat, A. S., (2023). Land Subsidence Studies in the Godavari Delta Regions of the East Coast of India Using ALOS and Sentinel 1 Data. *Ecological Informatics,* Vol. 78. https://doi.org/10.1016/j.ecoinf.2023.102373.

[24] Li, H., Zhu, L., Guo, G., Zhang, Y., Dai, Z., Li, X., Chang, L. and Teatini, P., (2021). Land Subsidence Due to Groundwater Pumping: Hazard Probability Assessment Through the Combination of Bayesian Model and Fuzzy Set Theory. *Natural Hazards and Earth System Sciences*, Vol. 21(2), 823–835. https://doi.org/10.5194/nhess-21-823-2021.

[25] Rezaei, M., Yazdani Noori, Z. and Dashti Barmaki, M., (2022). Land Subsidence Susceptibility Mapping Using Analytical Hierarchy Process (AHP) and Certain Factor (CF) Models at Neyshabur Plain, Iran. *Geocarto International*, Vol. 37(5), 1465–1481. https://doi.org/10.1080/10106049.2020.1768596.

[26] Eghrari, Z., Delavar, M. R., Zare, M., Beitollahi, A. and Nazari, B., (2023). Land Subsidence Susceptibility Mapping Using Machine Learning Algorithms. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 10(4/W1-2022), 129–136. https://doi.org/10.5194/isprs-annals-X-4-W1-2022-129-2023.

[27] Chai, L., Wei, L., Cai, P., Liu, J., Kang, J. and Zhang, Z., (2024). Risk Assessment of Land Subsidence Based on GIS in the Yongqiao Area, Suzhou City, China. *Scientific Reports*, Vol. 14(1), 1–13. https://doi.org/10.1038/s41598-024-62108-w.

[28] Osama, N., Shao, Z. and Freeshah, M., (2023). The FABDEM Outperforms the Global DEMs in Representing Bare Terrain Heights. *Photogrammetric Engineering and Remote Sensing*, Vol. 89(10), 613–624. https://doi.org/10.14358/PERS.23-00026R2.

[29] DeMaris, A., (2012). Logistic Regression: Basic Foundations and New Directions. *Handbook of Psychology, Second Edition*. https://doi.org/10.1002/9781118133880.hop202019.

[30] da Silva, I. N., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L. H. B. and dos Reis Alves, S. F., (2017). Multilayer Perceptron Networks. In *Artificial Neural Networks: A Practical Course*. Springer International Publishing. (55–115). https://doi.org/10.1007/978-3-319-43162-8_5.

[31] Breiman, L., (2001). Random Forests. *Machine Learning*, Vol. 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.

[32] Scornet, E., Biau, G., and Vert, J. P. (2015). Consistency of Random Forests. *Annals of Statistics*, Vol. 43(4), 1716–1741. https://doi.org/10.1214/15-AOS1321

[33] Natekin, A. and Knoll, A., (2013) Gradient Boosting Machines, a Tutorial. *Front. Neurorobot*, Vol. 7. https://doi.org/10.3389/fnbot.2013.00021.

[34] Chen, T. Q. and Guestrin, C., (2016) Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785.

[35] Liu, J., Liu, W., Allechy, F. B., Zheng, Z., Liu, R. and Kouadio, K. L., (2024). Machine Learning-Based Techniques for Land Subsidence Simulation in an Urban Area. *Journal of Environmental Management*, Vol. 352. https://doi.org/10.1016/j.jenvman.2024.120078.