Depressive Disorder Classification Using Machine Learning Techniques

Muhammad Raihan*, An Naura Erwana Dwi Putri*, Amalia Diah Ramadani*,
Resky Auliyah Kartini Askin*, Isra Nirwana Nur N. Kalau*, Michael Gabriel Bida*
*Hasanuddin University, Makassar, Indonesia
reskyauliyah@unhas.ac.id, annaura@unhas.ac.id, amalia.ramadani@unhas.ac.id,
raihan.m@unhas.ac.id, isra.nur@unhas.ac.id, michael.bida@unhas.ac.id

***Abstract*—Landslides are common natural disasters in hilly regions, particularly in Indonesia, often causing significant economic and social damage. While rainfall is widely recognized as a key trigger, soil physical properties such as texture, bulk density, pH, and organic matter content also play a critical role in landslide susceptibility. This study investigates the influence of soil characteristics on landslide risk and develops a predictive model using data mining, enhanced by an interactive dashboard visualization. Two primary datasets, HWSD and WLD, were integrated for analysis. Statistical tests (*t*-test and Mann–Whitney *U*) and the XGBoost algorithm were applied. Results show that subsoil properties contribute more to landslide risk than topsoil (AUC 0.7205 vs. 0.6984), with the combined model yielding the best performance (AUC 0.85). These findings highlight the importance of subsoil properties in slope stability and landslide prediction, offering valuable insights for improving early warning and mitigation systems.**

I. Introduction

Landslides are one of the deadliest natural disasters in the world, especially in tropical and mountainous regions such as South Asia and Southeast Asia. Based on data from the Center for Research on the Epidemiology of Disasters (CRED), landslides have caused more than 55,000 deaths globally between 1998 and 2017, and more than 75The Southeast Asian region, including Indonesia, the Philippines and Myanmar, is particularly vulnerable to landslides due to the influence of the monsoon climate as well as anthropogenic activities such as deforestation and development on steep slopes. According to a UNDRR report, about 60While rainfall is often considered the main trigger, recent research has shown that the physical characteristics of the soil-such as texture, permeability, moisture content and density-contribute significantly to slope stability and landslide potential [3]. However, most early warning systems are still dominated by meteorological and topographic-based approaches, while pedological information is often neglected [4]. Along with the development of data-driven approaches, machine learning offers great potential in improving the accuracy of landslide risk prediction. Algorithms such as Random Forest, XGBoost, and LightGBM can utilize hundreds of soil and environmental features in robust classification models. However, challenges such as class imbalance often arise, as the amount of landslide data (label 1) is generally much less than non-landslide data (label 0). To overcome this, the SMOTE (Synthetic Minority Over-sampling Technique) [5] is used, which is effective in balancing the data distribution before model training. This study adopts a pipeline-based approach with SMOTE integration, model training with Random Forest, XGBoost, and LightGBM, and hyperparameter tuning using Bayesian Optimization [6]. Model evaluation is conducted through Stratified K-Fold Cross Validation with key metrics such as F1-score and ROC-AUC, and exploration of optimal thresholds to improve recall as part of disaster risk mitigation strategies. Through this approach, the research aims to build a soil property-based landslide risk prediction model that is not only accurate, but also spatially informative and practical to be implemented in early warning systems in disaster-prone areas.

II. Methodlogy

A. Study Area

This study is based on data from 13 countries across five continents: Australia, Brazil, China, Costa Rica, Ecuador, Italy, Mexico, New Zealand, Norway, Pakistan, South Africa, Taiwan, and Vietnam. The selection of these nations was guided by two primary criteria: a significant diversity of soil types , as indexed by the Harmonized World Soil Database (HWSD) and a high frequency of recorded landslides between 2006 and 2017, as documented in NASA's Global Landslide Catalog. Figure 1 illustrates the distribution of landslide events

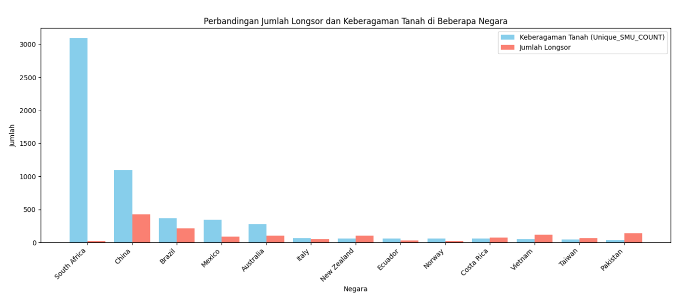

Fig. 1. Landslide events vs. unique soil characteristics.

and soil diversity (represented by the count of unique Soil Mapping Units) for the selected countries. For most nations, a general correspondence is observed between the two metrics. South Africa, however, presents a notable exception, exhibiting exceptionally high soil diversity relative to its number of recorded landslides. This broad geographical scope inherently introduces significant variations in other key landslide-triggering factors, such as climate patterns (ranging from tropical to temperate), topography (from steep mountainous terrain to gentle hills), and land cover, which is essential for developing a globally generalizable model.

B. Datasets

This study integrates data from three primary sources. Soil characteristics were derived from the Harmonized World Soil Database (HWSD), a 30 arc-second raster database providing presents soil information collected from various regional and national sources, including the European Digital Map (ESDE), the 1:1,000,000 scale Soil Map of China, and various legacy soil maps from FAO. Data from these different sources were harmonized using standard procedures. The database consists of a raster image that is linked to the attribute database through a unique code of the soil mapping unit. For each mapping unit, the database provides information on the composition of the soil types present in it (dominant soils and accompanying soils). It also contains quantitative data on soil physical and chemical properties for two depth layers, namely topsoil (0-30 cm) and subsoil (30-100 cm). These attributes include organic carbon content, pH, water holding capacity, soil depth, cation exchange capacity, clay fraction, salinity, and soil texture. Table 1 describes the classification of data sources and their data types and table 2 describes general information on the soil mapping unit composition.

TABLE I
DATA SOURCES AND OUTPUTS IN THE HARMONIZED WORLD SOIL DATABASE

No.	Data Source	Format	Output
	Database Name	***Data Type***	***Resolution/Quantity***
1	ESDB	Geo. DB	Raster ∼1 km
2	Soil Map China	Digital Map	Raster ∼1 km
3	SOTER (SOTWIS)	Soil	Raster ∼1 km
4	Soil Map World	Digital Map	Raster ∼1 km
5	Soil Profile DB	Profile Data	9607 profiles

[a]ESDB: European Soil Database. Geo. DB: Geographic Database.
[b]Soil Map of China; SOTER: World Soils and Terrain Database.
[c]Soil Profile DB: Soil Profile Database.
[d]Input Scale/Resolution for raster data (1:1M or 1:10M).
[e]Input Scale/Resolution for SOTER databases (1:2.5M – 5M).

Historical landslide event data were obtained from the NASA Global Landslide Catalog (GLC), from which we extracted all documented rainfall-triggered landslides occurring between 2006 and 2017. Finally, all data were geographically contextualized using the Natural Earth “Admin 0 – Countries” vector dataset (1:10m scale) to define the national administrative boundaries for the study area.

C. Method

The methodology adopted in this study is systematically illustrated in Figure 2. The research commenced with an extensive literature review on soil characteristics and landslide occurrences to gather relevant data. Subsequently, three distinct datasets were compiled and integrated.

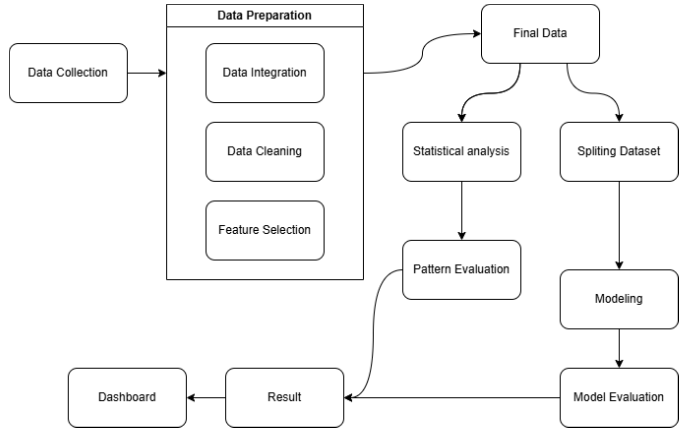

Fig. 2. research workflow

This integration process was guided by two primary references: the HWSD Technical Report and Instructions and a technical note by D.G. Rossiter, Processing the Harmonized World Soil Database (Version 1.2) in R. A detailed flowchart of the data integration procedure is presented in Figure 3. From the study area, a total of 2,210 landslide and non-landslide data, each corresponding to different soil characteristics, were identified.

TABLE II
FIELD AVAILABILITY IN SOIL DATABASES (SIMPLIFIED)

Field	Description	DSMW
General		
ID	Database ID	√
MU_GLOBAL	Global Unit ID	√
COVERAGE	Coverage	√
ISSOIL	Soil indicator	√
SHARE	Share in Unit	√
SU_SYMBOL	Symbol	√
Phases and Additional		
PHASE1	Phase 1	√
ROOTS	Root obstacles	
AWC_CLASS	AWC Class	√

Following data integration, a comprehensive data preprocessing phase was conducted. This began with data cleaning, in which missing values were imputed using the median of their respective columns. Subsequently, outlier handling was performed using the Interquartile Range (IQR) method. This analysis revealed numerous outliers across most feature columns; these were also imputed using the column-specific median in order to maintain data integrity. The next stage focused on feature selection. Features were removed based on three criteria: identifier columns (as listed in Table II), columns containing only a single unique value, and those with low feature importance scores. A machine learning-based approach using the Random Forest algorithm was employed to assess the importance of each remaining feature. The results of this analysis, visualized in Figure 4, identified two features—`AWC_CLASS` and `S_CASO4`—with an importance score of zero. Consequently, these non-contributory features

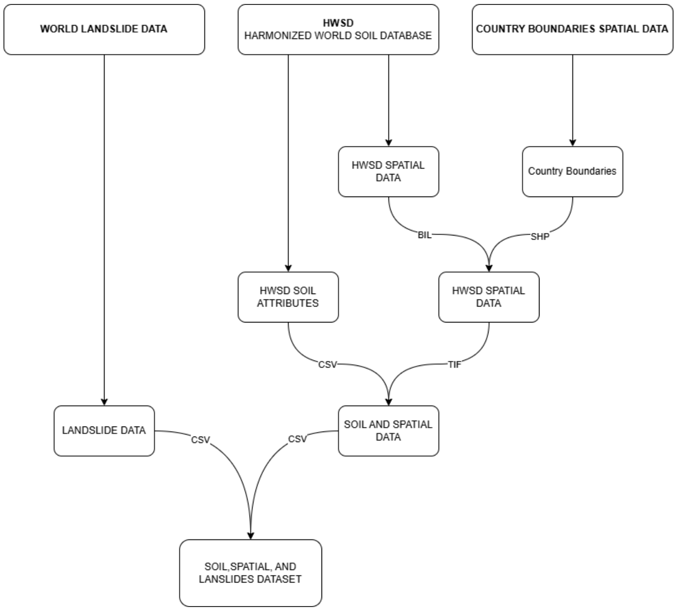

Fig. 3. Data Integration Workflow

were removed from the dataset.

The final preprocessing step was data transformation. Label encoding was applied to the categorical feature 'COUNTRY'. This method was selected for its suitability with tree-based models, which are invariant to the ordinal relationships that might be artificially introduced by other encoding techniques. This study employs two primary analytical approaches: statistical analysis and machine learning modeling.

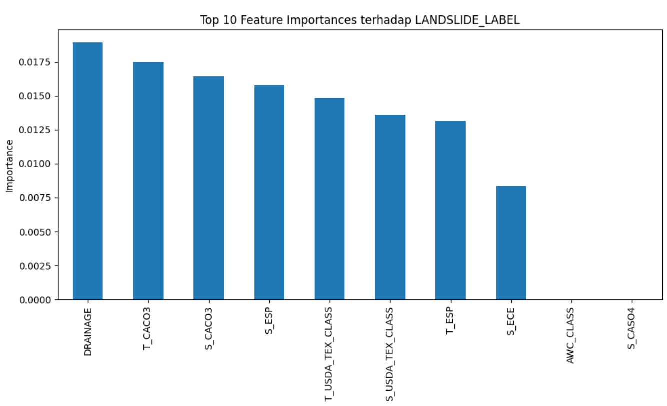

Fig. 4. Landslide events vs. unique soil characteristics.

1) statistical analysis: Statistical analysis was conducted to identify significant differences in soil characteristics between two predefined groups: locations where landslides occurred (labeled as 1) and locations where they did not (labeled as 0). To achieve this, both parametric and non-parametric hypothesis tests were applied, specifically the independent samples t-test and the Mann-Whitney U test.

The main objective of these tests was to calculate the p-value for each soil feature. The p-value quantifies the probability that an observed difference between the groups is merely due to random chance. A lower p-value indicates a more statistically significant difference, suggesting that the corresponding soil feature is a meaningful differentiator between landslide and non-landslide conditions.

The mathematical formulation for the independent samples t-test is presented in Equation 1, while the formula for the Mann-Whitney U test is detailed in Equation 3.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1}$$

where s_p is the pooled standard deviation, calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2}$$

For the Mann-Whitney U test, the U statistic is given by:

$$U = \min(U_1, U_2) \tag{3}$$

where:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{4}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \tag{5}$$

2) machine learning modeling: For the primary classification task, this study utilized the Extreme Gradient Boosting (XGBoost) model. XGBoost is a highly efficient and scalable implementation of the gradient tree boosting algorithm, widely regarded as a state-of-the-art machine learning method. It employs a regularized boosting technique, which effectively mitigates overfitting and thereby enhances model accuracy and generalization performance. The selection of XGBoost was motivated by its numerous advantages, including its scalability across diverse scenarios, inherent capability to handle sparse data, low computational resource requirements, high-performance speed, and ease of implementation. The fundamental principle of the boosting algorithm is to sequentially combine the outputs of multiple weak learners—in this case, Classification and Regression Trees (CARTs)—to create a single, robust predictive model. The core of the algorithm aims to minimize the regularized objective function, as formulated in Equation 6. This function is composed of two main parts: a loss function and a regularization term. The loss function measures the discrepancy between the actual target (y_i) and the prediction ($\hat{y}_i$). The second component, the regularization term detailed in Equation 7, penalizes the complexity of the model to avoid overfitting. The overall algorithmic process is described by Equations 6 through 9.

$$L(\Phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{6}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{7}$$

where:

- T: the number of leaves in the tree;

- w: the score of each leaf;
- γ, λ: the regularization degrees.

$$L^{(t)}(\Phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{8}$$

In order to speed up the optimization process, second order Taylor expansion is applied to the objective. After removing the constant terms, a simplified objective function at step t is given in Equation 9.

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{9}$$

where:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{10}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{11}$$

The preprocessed dataset was partitioned for training and validation using a 5-fold StratifiedKFold cross-validation scheme. This approach effectively splits the data into an 80% training set and a 20% testing set during each fold, while critically preserving the original class distribution (landslide vs. non-landslide) across all folds. This stratification is essential for ensuring reliable model evaluation on an imbalanced dataset. To address the class imbalance issue, the SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training data to synthetically oversample the minority (landslide) class.

Three classification algorithms were evaluated in this study: Random Forest, XGBoost, and LightGBM. To enhance the models' sensitivity to the positive class (landslide), a class weight adjustment mechanism was integrated into the training process. Hyperparameter optimization was conducted efficiently using Bayesian Optimization implemented via BayesSearchCV, facilitating an effective search for the optimal parameter combination within a predefined search space.

Model performance was assessed using a suite of standard classification metrics. The Confusion Matrix served as the foundation for performance analysis, categorizing predictions into four distinct outcomes: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), as illustrated in Figure 5.

The primary evaluation metric for this study was the F1-Score, which is the harmonic mean of precision and recall, providing a balanced measure of performance on imbalanced data. Furthermore, the classification threshold was tuned to achieve a desired level of recall. The Receiver Operating Characteristic (ROC) curve was also utilized to visualize the discriminative ability of the models across various thresholds. This curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR), where the area under the curve (AUC-ROC) provides an aggregate measure of performance across all classification thresholds.

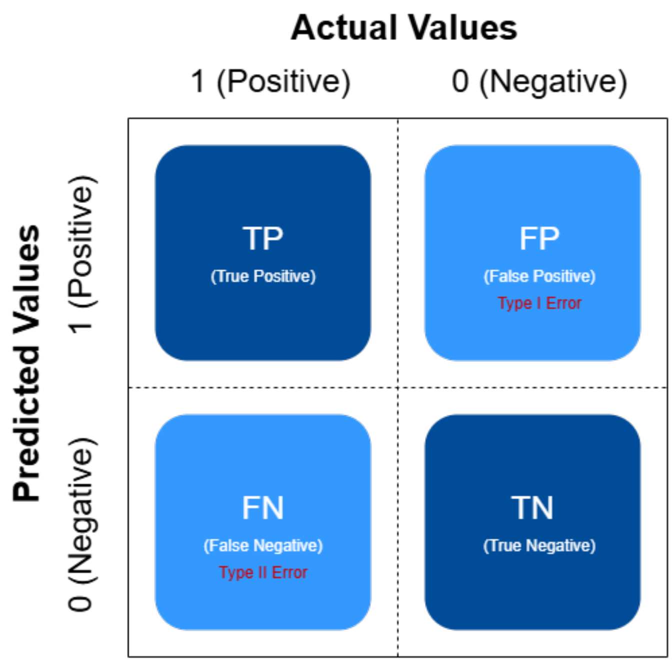

Fig. 5. Landslide events vs. unique soil characteristics.

The sensitivity (True positive or Recall) tells the proportion of positive class (landslides locations) that are correctly classified as landslides (Equation (8)). In contrast, the specificity (True Negative Rate) tells the proportion of negative class (non-landslides locations) that are correctly classified as non-landslides (Equation (9)). Between sensitivity and specificity lies False Negative Rate (FNR), which signifies the proportion of landslide points wrongly classified as landslides (Equation (10)). The False Positive Rate (FPR) tells the proportion of non-landslides incorrectly classified as non-landslides (Equation (11)).

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{13}$$

$$\text{FNR} = \frac{FN}{TP + FN} \tag{14}$$

$$\text{FPR} = \frac{FP}{TN + FP} = 1 - \text{specificity} \tag{15}$$

III. Result

By comparing the data distribution of soil characteristics in both the landslide and no landslide groups, as illustrated in Figure 6, it can be observed that the two groups exhibit differences in distribution, although not always significant. Taking the *S_CLAY* feature as an example—which represents the clay content in the subgrade—it is evident that the no landslide group shows a data concentration between 20 and 40, with a median value tending to lie below 40. In contrast, the landslide group exhibits a concentration between 30 and

60, with a higher median than the no landslide group. This suggests that landslide-prone areas tend to have higher clay content compared to non-landslide areas.

To support this observation, a statistical analysis was conducted to calculate the *p*-value using Equations 1 or 3 (the formula used depends on the data distribution, if the data distribution is normal then use t-test and u whitney-u otherwise). The `S_CLAY` feature yielded a *p*-value of 6.38×10^{-5} (or 0.0000638), which is much smaller than the significance threshold of 0.05. This indicates a statistically significant difference between the two groups.

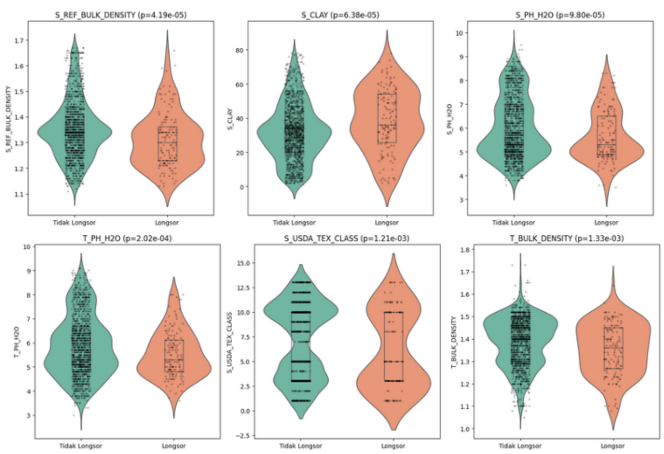

Fig. 6. Violin Plot of Topsoil vs Subsoil.

To evaluate the relative predictive contribution of different soil layers, separate classification analyses were conducted for topsoil and subsoil features. The discriminative performance of these two feature sets was then evaluated using ROC (Receiver Operating Characteristic) curves and AUC (Area Under the Curve) values. The results, as shown in Figure 7, indicate that the model trained using the subsoil features achieved slightly superior predictive performance with an AUC value of 0.7205, compared to the model using the topsoil features with an AUC of 0.6984.

To further investigate the performance difference between the two soil layers, a feature importance analysis was conducted, as illustrated in Figure 8. In the topsoil layer, the most influential features were primarily related to soil chemical properties, including Electrical Conductivity (T_ECE), Calcium Carbonate (T_CACO3), Total Exchangeable Bases (T_TEB), and Organic Carbon (T_OC). Meanwhile, in the subsoil layer, Calcium Carbonate (S_CACO3) and Electrical Conductivity (S_ECE) were the top contributors, followed by texture classification (S_USDA_TEX_CLASS) and Total Exchangeable Bases (S_TEB). These findings highlight that subsoil characteristics—particularly chemical and textural properties—play a more critical role in predicting landslide susceptibility.

Collectively, these findings suggest that while both soil layers contribute valuable information, subsoil characteristics—particularly those related to chemical composition and texture classification—exhibit greater predictive power in landslide risk modeling compared to topsoil features.

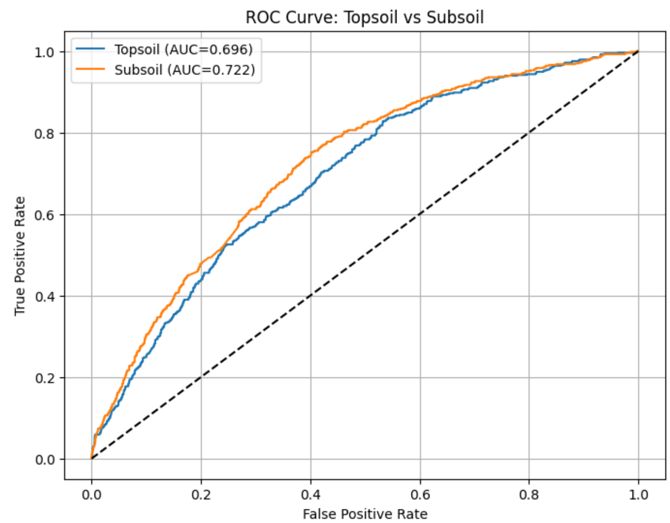

Fig. 7. ROC-AUC of Topsoil and Subsoil Layers.

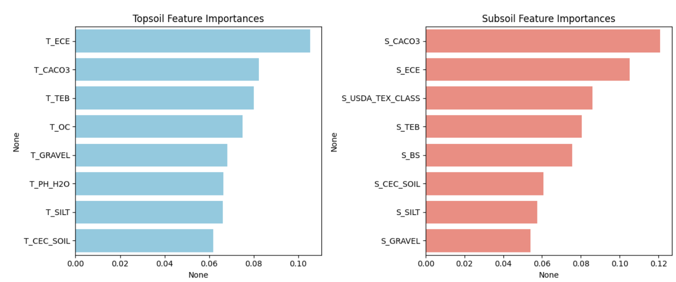

Fig. 8. Feature importance of Topsoil and Subsoil Layers.

As a final step, a comprehensive model was trained by combining features from the topsoil and subsoil layers to evaluate their synergistic effects. The evaluation results of this combined model, visualized in the ROC curve in Figure 9, show a significant improvement in performance, achieving an AUC value of 0.85. This value far surpasses the performance of the model using only one of the layers separately.

This substantial improvement confirms that although the subsoil characteristics have a slightly more dominant predictive power, information from the topsoil also makes an essential complementary contribution. Thus, it can be concluded that holistic soil profile analysis—considering the interaction between both layers—is the most effective approach for landslide risk modeling with the highest accuracy.

In Figure 10 the confusion matrix shows that out of a total of 531 actual landslide events (class 1), the model correctly identified 346 of them (True Positive), while 185 events were missed (False Negative). For the non-landslide class (class 0), the model demonstrated strong performance with 1539 correct predictions (True Negative) and only 140 incorrect predictions (False Positive).

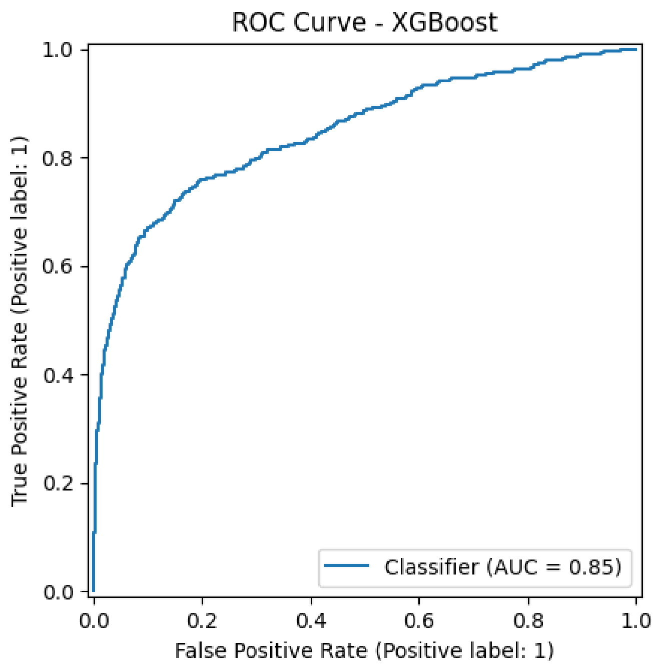

Fig. 9. ROC-AUC curve for all layers.

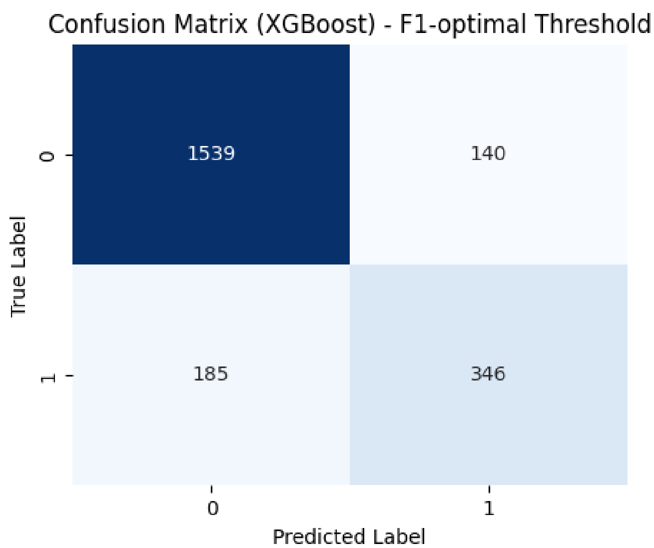

Fig. 10. Confussion matrix.

This is reflected in the classification report, where the model achieved a recall of 0.65 and a precision of 0.71 for the landslide class, resulting in an F1-score of 0.68. The overall accuracy of the model reached 85.29%, indicating generally reliable predictive capability. With a weighted average F1-score of 0.85, the model demonstrates an effective balance between precision and recall across the dataset, validating the success of the optimization strategy on imbalanced data.

IV. Discussion

A statistical significance test was conducted to examine the differences in soil characteristics between landslide and non-landslide groups, as illustrated in the violin plots (Fig. *ref*). The aim of this analysis was to verify whether there were significant distinctions in soil attributes between the two groups. For this purpose, both the *t*-test and Mann–Whitney U test were employed. The use of such statistical tests follows similar methodologies adopted in prior studies. For instance, [1] applied the Mann–Whitney U test to compare soil properties such as clay content, bulk density, and pH between landslide-affected and unaffected areas in the Three Gorges Reservoir, while [2] employed the Mann–Whitney U test for non-normally distributed data and the *t*-test for normally distributed data to evaluate differences in soil porosity, organic content, and texture in landslide-prone areas of the Himalayas.

The results depicted in the violin plots reveal several soil characteristics exhibiting significant differences, as indicated by p-values less than 0.05. Notably, variables such as `s_clay` and `s_ref_bulk density` demonstrate statistically significant differences between the landslide and non-landslide groups. These findings are consistent with the observations reported by [3], who stated that thick clay layers with low bulk density, due to their loose structure, render the soil more susceptible to landslides.

Performance indicators such as the Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), and evaluation matrices (Fig. *ref*) were used to validate the predictive capability of the XGBoost algorithm. These indicators follow evaluation standards similar to those employed by [4], who used ROC-AUC to assess landslide susceptibility mapping using XGBoost. Furthermore, the model separates the analysis between topsoil and subsoil layers, following the approach of [5], who investigated how different soil layers influence mass movement events.

As shown in Fig. *ref*, the comparison of XGBoost model performance between the two soil layers yielded an AUC of 0.6984 for topsoil and 0.7205 for subsoil. Although the difference is not statistically significant, the subsoil layer exhibited slightly superior predictive performance. This aligns with findings reported by [6], who noted that subsoil layers exhibit more stable properties—such as organic content and texture—over time, particularly after landslide events, thereby making them more suitable for long-term landslide analysis.

An additional model evaluation was conducted by integrating both topsoil and subsoil data, as shown in Fig. *ref*, resulting in a significant increase in AUC to 0.85. This suggests that combining both soil layers enhances model performance, with subsoil data contributing more strongly to the prediction. Landslide analysis using data mining approaches has proven to be a valuable tool for spatial planning and land management. However, it remains a challenge to achieve high model prediction performance using soil properties alone. As demonstrated by [7], incorporating other environmental factors such as rainfall and slope gradients can significantly improve model performance, achieving an AUC of 0.89, even though the geographical context differs.

V. Conclusion

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] Z. Guo, L. Chen, K. Yin, D. P. Shrestha, and L. Zhang, "Quantitative risk assessment of slow-moving landslides from the viewpoint of decision-making: A case study of the Three Gorges Reservoir in China," Engineering Geology, vol. 273, Art. no. 105667, Aug. 2020.

[2] Z. Habib, A. Kumar, R. A. Mir, I. M. Bhat, W. Qader, and R. K. Mallik, "Geotechnical analysis and landslide susceptibility of overburden slope material in the Jammu and Kashmir, Western Himalaya," Geosystems and Geoenvironment, in press, Art. no. 100413, May 2025. DOI: 10.1016/j.geogeo.2025.100413

[3] J. Sartohadi, N. A. H. J. Pulungan, M. Nurudin, and W. Wahyudi, "The ecological perspective of landslides at soils with high clay content in the Middle Bogowonto Watershed, Central Java, Indonesia," Applied and Environmental Soil Science, vol. 2018, Art. ID 2648185, May 2018. DOI: 10.1155/2018/2648185

[4] S. Badola, V. N. Mishra, and S. Parkash, "Landslide susceptibility mapping using XGBoost machine learning method," in Proc. 2023 Int. Conf. Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS), Hyderabad, India, 2023, pp. 1–4, doi: 10.1109/MIGARS57353.2023.10064496.

[5] A. Parasyris, L. Stankovic, and V. Stankovic, "A Machine Learning-Driven Approach to Uncover the Influencing Factors Resulting in Soil Mass Displacement," Geosciences, vol. 14, no. 8, p. 220, Aug. 2024. doi: 10.3390/geosciences14080220.

[6] M. S. Kim, Y. Onda, J. K. Kim, and S. W. Kim, "Effect of topography and soil parameterisation representing soil thicknesses on shallow landslide modelling," Quaternary International, vol. 384, pp. 91–106, Oct. 2015. doi: 10.1016/j.quaint.2015.03.057.

[7] Y. Han and S. J. Semnani, "Important considerations in machine learning–based landslide susceptibility assessment under future climate conditions," [Journal Name Unavailable], vol. 20, pp. 475–500, 2025. Published: Aug. 3, 2024.