Landslide Prediction Based on Soil Characteristics Using Data Mining Techniques: Comparing Topsoil vs Subsoil in Risk Zones

Muhammad Raihan*, An Naura Erwana Dwi Putri*, Resky Auliyah Kartini Askin*,
Amalia Diah Ramadani*, Michael Gabriel Bida*
*Hasanuddin University, Makassar, Indonesia
raihanm23h@student.unhas.ac.id, putrianed23h@student.unhas.ac.id, askinrak23h@student.unhas.ac.id,
ramadaniad23h@student.unhas.ac.id, bidamg21h@student.unhas.ac.id

***Abstract*—Landslides are one of the most frequent natural disasters, and often pose serious threats and significant losses. While most studies have focused on external triggers like rainfall and slope, this research investigated the influence of internal factors namely, soil characteristics on landslide risk. This study investigated the role of properties from both topsoil and subsoil layers on landslide risk and developed a predictive model using data mining techniques. Three main datasets were integrated for the analysis: the Harmonized World Soil Database (HWSD), the Global Landslide Catalog (GLC), and administrative boundaries from Natural Earth. First, comparative statistical analyses were conducted to confirm significant differences in soil properties between landslide and non-landslide areas. Subsequently, an XGBoost algorithm was applied to model the landslide risk. The results showed that models using topsoil, subsoil, and a combination of both layers yielded consistent AUC values of approximately 0.73. However, analysis of F1 scores revealed that subsoil characteristics, especially chemical composition and texture exhibited slightly greater predictive power than topsoil. The combined model, integrating both layers, achieved an F1 score of 0.71, an increase of approximately 3%. These findings indicate that although subsoil characteristics are slightly more dominant, topsoil information provides an essential complementary contribution to achieve optimal model performance.**

***Index Terms*—Landslide, Soil Characteristics, Data Mining, XGBoost, Machine Learning, Subsoil, Topsoil**

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. According to data from the Center for Research on the Epidemiology of Disasters (CRED), landslides ranked as the fourth most frequent natural disaster in 2024 [1]. One of the most recent landslides disasters occurred in Papua New Guinea, where a major landslide in Enga Province resulted in one of the country's most severe disasters in recent memory. with United Nations agencies estimating that there were around 670 fatalities. This devastating event underscores the critical need to understand the underlying causes of such disasters.

Landslides are hazardous phenomena that occur when slopes become unstable. This instability is widely attributed to the synergistic effect of an area's inherent physical conditions and the influence of external events. According to regional disaster system theory, the occurrence of a landslide is the result of this interplay between spatial static susceptibility and temporal dynamic inducibility [2].

These causal factors are broadly classified into two categories. The first are conditioning factors (or preparatory/non-variable factors), which represent the inherent spatial susceptibility of an area, such as its geology, soil type, topography, and lithology. The second group are the triggering factors, which are the dynamic, temporal events directly responsible for initiating the slope failure. These triggers, which include high-intensity rainfall, earthquakes, volcanic activity, and disruptive human interventions (like deforestation or improper construction), are often considered the most immediate cause of a landslide.

Although external factors such as heavy rain, earthquakes, and deforestation are often considered the dominant trigger, many prediction models have been developed relying exclusively on these external agents. For example, a systematic review by [3] analyzed numerous articles on rainfall thresholds, finding that most models depend solely on meteorological data. Similarly, [4] developed empirical thresholds using only external factors, and even advanced machine learning approaches by [5] still focus exclusively on external rainfall metrics. This reliance on external triggers highlights a critical gap: while the role of triggers is well-documented, the independent predictive power of internal soil properties (such as texture, permeability, and density) is less understood. Most systems ignore these subsurface dynamics, leading to potential inaccuracies. Therefore, to better isolate and quantify the contribution of these overlooked factors, this study addresses this gap by developing a predictive model that focuses exclusively on internal soil characteristics, specifically analyzing the unique contributions of surface soil and subsoil properties to landslide risk.

Our research focuses on two main soil layers, topsoil and subsoil, which possess distinct characteristics and hydrological functions critical to slope failure mechanisms. The topsoil, typically rich in organic matter and more porous, facilitates initial rainwater infiltration. In contrast, the subsoil is generally denser, with higher clay content and lower permeability, which

impedes drainage. This structural difference is a key factor in rainfall-induced landslides. As reviewed by [6], water seeping through the topsoil accumulates at the interface with the less permeable subsoil during intense rainfall. This buildup elevates pore water pressure, which in turn diminishes effective stress and reduces the soil's internal shear strength, acting as a primary trigger for slope failure. The significance of this mechanism is highlighted by studies referenced in their review, which show that soil saturation from infiltration can reduce the shear modulus by up to 50%, directly linking these layered properties to landslide initiation.

This study aims to develop an accurate landslide risk prediction model while elucidating the specific contribution of pedological features.

II. Methodology

A. Study Area

The data used in this study was acquired 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 (GLC). Fig [1] illustrates the distribution of landslide events and soil diversity (represented by the count of unique Soil Mapping Units) for the selected countries.

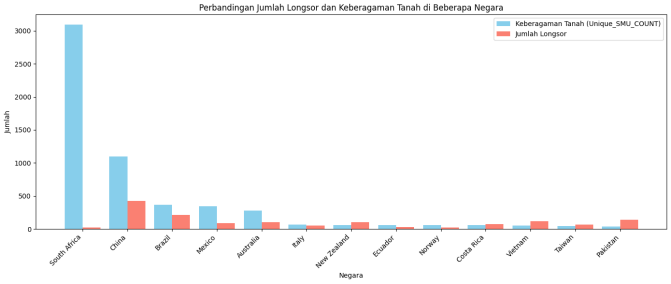

Fig. 1. Landslide events vs. unique soil characteristics.

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. The broad geographical scope was intentionally chosen to validate the model beyond a single, localized area. This approach ensures the inclusion of significant variations in key landslide-triggering factors such as diverse climate patterns, topography, and land cover which is essential for developing a more robust and globally generalizable model.

B. Datasets

This study integrates data from three major data sources. First, soil characteristics were derived from the Harmonized World Soil Database (HWSD) v1.2 [7]. This 30 arc-second raster database harmonizes soil information from various regional and national sources, such as the European Soil Database (ESDB), the 1:1,000,000 scale Soil Map of China, and legacy Food and Agriculture Organization (FAO) soil maps. Structurally, the HWSD consists of a raster image linked to an attribute database via a unique soil mapping unit code. This attribute table provides information on soil composition (dominant and accompanying soils) for each mapping unit. Crucially for this study, the database provides quantitative data on physical and chemical properties for two depth layers: 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 [I] describes the classification of data sources and their data types, while Table [II] describes general information on the soil mapping unit composition.

Second, historical landslide event data were sourced from the GLC [8]. The GLC, developed and maintained by the NASA Goddard Space Flight Center since 2007, is a comprehensive global repository specifically designed to identify rainfall-triggered landslide events, regardless of size, impact, or location. It compiles reports from diverse sources, including media, disaster databases, and scientific reports. For this study, we extracted all documented rainfall-triggered landslide events from the GLC database corresponding to the period of 2006-2017. Finally, all data were geographically contextualized using the Natural Earth 'Admin 0 – Countries' dataset (v5.1.1, 1:10m scale). This dataset provides de facto national administrative boundaries, which reflect actual territorial control, to define the study area [9].

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).

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	√

The process of integrating the three data sets was guided by two main references: the first one is HWSD Official Technical Report and Instructions, this report was essential as it provides

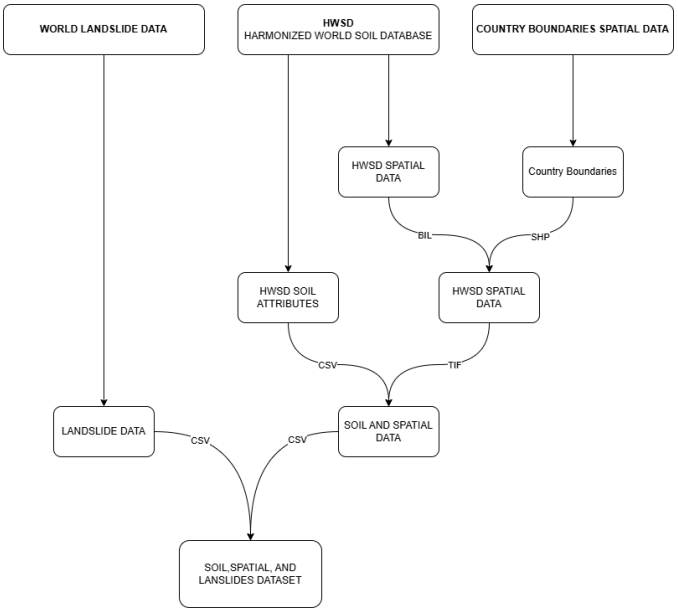

Fig. 2. Data Integration Workflow

the technical methodology for utilizing the HWSD data. The HWSD itself is a comprehensive global dataset developed by the FAO and IIASA precisely to address the long-standing lack of reliable, harmonized soil information, which has historically hampered environmental studies and predictive modeling. And the second one is a technical note by Professor D.G. Rossiter from Cornell University, Processing the HWSD Version 1.2 in R, This note explains how to access and query the HWSD data. This allows integration of the HWSD with any other geographic coverage, as well as statistical summaries. A detailed flowchart of the data integration procedure is presented in Fig[2], where the final result of this integration process is a dataset containing detailed soil characteristics of two layers of topsoil and subsoil in various countries, as well as landslide occurrence data with accurate geographic mapping.

C. Method

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 [10]. 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 method. A machine learning-based approach using the XGBoost algorithm was employed to assess the importance of each remaining feature [11]. The results of this analysis, visualized in Fig[3], identified two features `AWC_CLASS` and `S_CASO4` with an importance score of zero. Consequently, these non-contributory features 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 [12]. This study employs two primary analytical approaches:

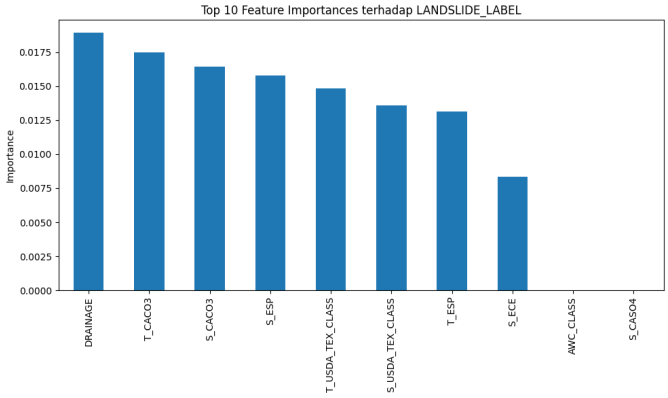

Fig. 3. Feature Importance Analysis

statistical analysis and machine learning modeling.

1) statistical analysis: Statistical anala comparative method was used to identify significant differences in soil characteristics between two soil groups: locations where landslides occurred (labeled as 1) and locations where landslides did not occur (labeled as 0). To achieve this, both parametric and non-parametric hypothesis tests were applied. If the data was normally distributed, the independent samples t-test was used, and if it was not normally distributed, the Mann-Whitney U test was used [13].

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 (1), while the formula for the Mann-Whitney U test is detailed in (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 XGBoost model [14]. 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 (CART) to create a single, robust predictive model. The core of the algorithm aims to minimize the regularized objective function, as formulated in (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 (7), penalizes the complexity of the model to avoid overfitting. The overall algorithmic process is described by (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 (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 [15]. 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 issue of significant class imbalance, the SMOTE method was applied to the training data to perform oversampling on the minority class [16] . In addition to oversampling, a weight adjustment was applied to further balance the influence of the minority class during model training . Furthermore, hyperparameter optimization was performed using Bayesian Optimization, implemented through BayesSearchCV [17]. This method facilitates an effective search for the optimal parameter combination within a specified search space. The Confusion Matrix served as the foundation for performance analysis [18], categorizing predictions into four distinct outcomes: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), as illustrated in 4.

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 employed to visualize the model's discriminative ability across various classification thresholds. This curve was generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). The Area Under the Curve (AUC) was subsequently calculated from the ROC curve, providing a single aggregate metric to quantify the model's performance across all possible thresholds.

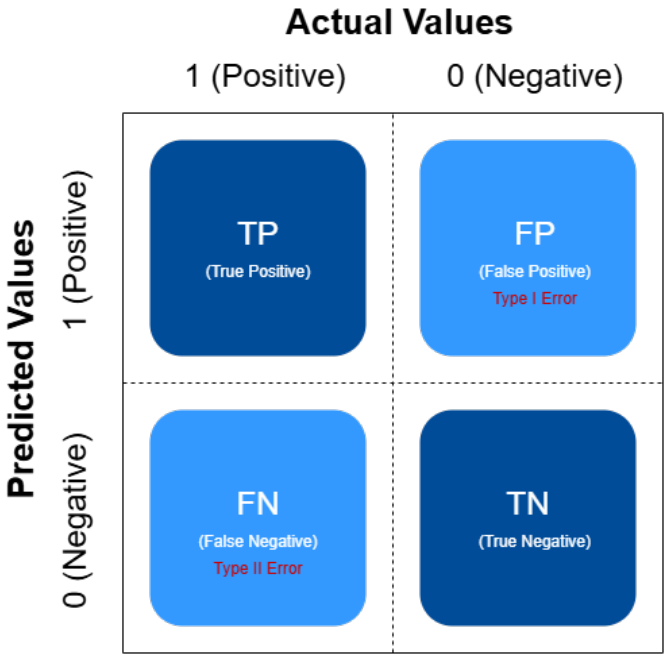

Fig. 4. Confusion matrix for evaluating classification performance.

The sensitivity (True positive or Recall) tells the proportion of positive class (landslides locations) that are correctly classified as landslides (12). In contrast, the specificity (True Negative Rate) tells the proportion of negative class (non-landslides

locations) that are correctly classified as non-landslides (13). Between sensitivity and specificity lies False Negative Rate (FNR), which signifies the proportion of landslide points wrongly classified as landslides (14). The False Positive Rate (FPR) tells the proportion of non-landslides incorrectly classified as non-landslides (15).

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

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

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

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

III. Results

By comparing the data distribution of soil characteristics in both the landslide and no landslide groups, as illustrated in Fig[5], 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 indicates that landslide-prone areas tend to have higher clay content. This finding is consistent with the research by [19], which found that "high thickness of soil and high content of clay play an important role in determining landslide sensitivity".

To support this observation, a statistical analysis was conducted to calculate the *p*-value using (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.

After analyzing the differences in soil characteristics between landslide and non-landslide areas, and proving that the two classes do indeed have different characteristics, the next step is to evaluate the predictive ability and contribution of soil features at two different depths, namely topsoil and subsoil.

To this end, separate classification analyses were conducted using topsoil and subsoil features. The discriminative performance of these two feature sets was then evaluated using (ROC) curves and (AUC) values. The results, as shown in Fig[6], 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.

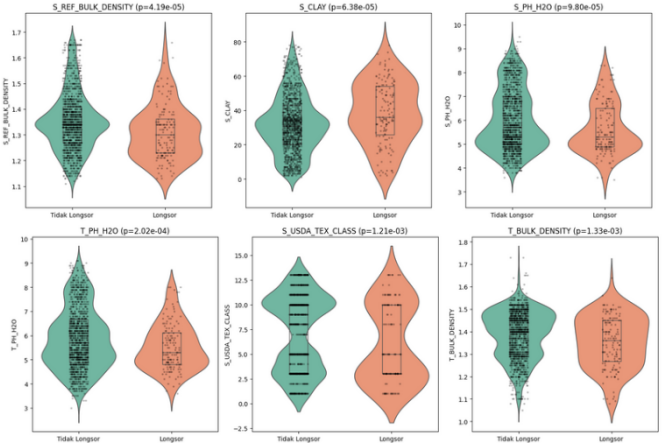

Fig. 5. Violin Plot of Topsoil vs Subsoil.

To further investigate the performance difference between the two soil layers, a feature importance analysis was conducted, as illustrated in Fig 7. These findings underscore the critical role of subsoil characteristics, particularly its chemical properties and texture, in predicting landslide susceptibility. This is consistent with a study by [20], which found that high infiltration water ECe (around 1600–5100 μS/cm) triggers ion exchange (related to TEB) in smectite minerals. This process results in material swelling and a decrease in shear strength in the subsoil layer (depth 4–6 m), thereby increasing susceptibility to creeping landslides.

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.

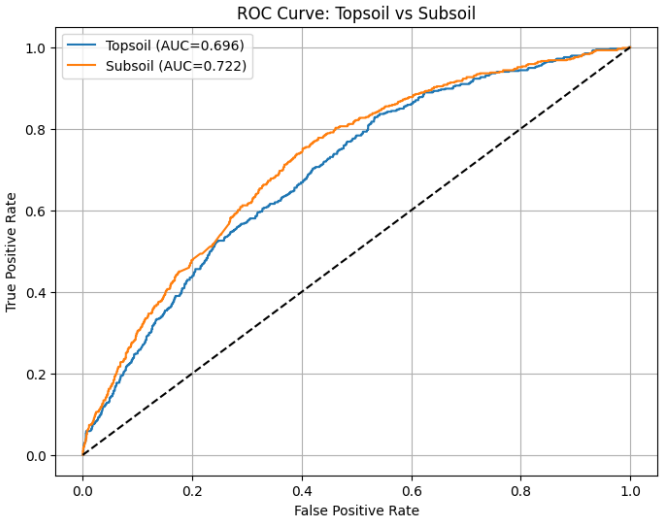

Fig. 6. ROC-AUC of Topsoil and Subsoil Layers.

As a final step, a comprehensive model was trained by combining features from the topsoil and subsoil layers to

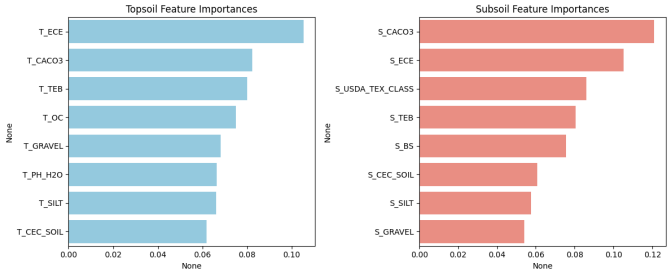

Fig. 7. Feature importance of Topsoil and Subsoil Layers.

evaluate their synergistic effects. The evaluation results of this combined model, visualized in the ROC curve in Fig[8], 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 to distinguish between the two classes.

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.

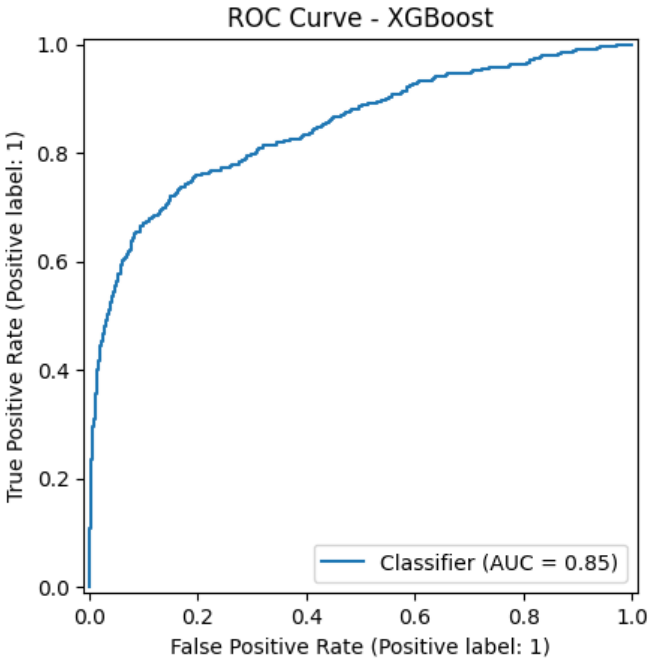

Fig. 8. ROC-AUC curve for all layers.

Fig 9 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).

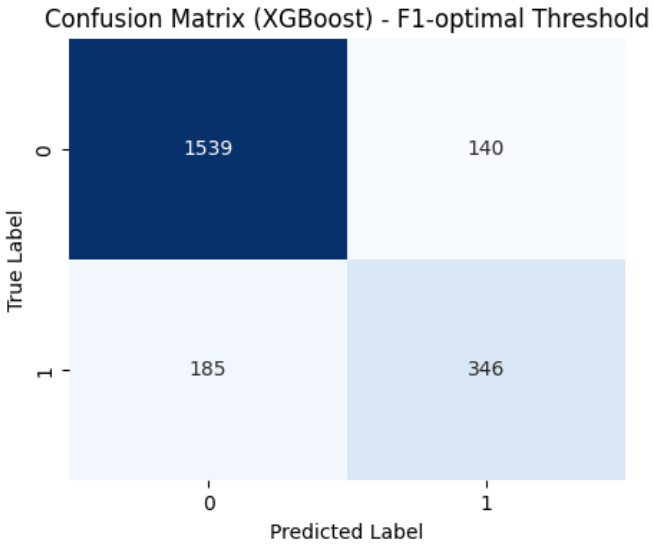

Fig. 9. Confusion matrix.

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. 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, [21] 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 [22] 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 Fig 5 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 [23], 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 ROC, AUC, and evaluation matrices were used to validate the predictive capability of the XGBoost algorithm. These indicators follow evaluation standards similar to those employed by [24], 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 [25], who investigated how different soil layers influence mass movement events.

As shown in Fig6, 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 [26], 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 Fig8, 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 [27], 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

The results of this study demonstrate that stratifying soil characteristics by depth topsoil and subsoil offers a more nuanced understanding of their respective contributions to landslide susceptibility. The XGBoost model trained on subsoil features outperformed the topsoil-based model (AUC = 0.7205 vs. 0.6984), while the combined model yielded the highest performance (AUC = 0.85; F1-score = 0.68). These findings confirm that incorporating a holistic soil profile enhances predictive accuracy.

The application of SMOTE to address class imbalance, coupled with Bayesian Optimization for hyperparameter tuning, proved effective in increasing the model's sensitivity to landslide events. Nonetheless, limitations persist in the generalizability of the optimized parameters, as the best-performing configuration in one cross-validation fold may not consistently perform well across others.

Feature importance analysis indicated that subsoil chemical properties such as calcium carbonate content and electrical conductivity play a pivotal role in landslide prediction. Additionally, soil textural attributes and cation exchange capacity further contributed to improved model performance, underscoring the complex interplay between physical and chemical soil factors in landslide initiation.

This study advances the development of soil-based early warning systems and provides valuable insights into key indicators of landslide risk. Future research should consider incorporating additional environmental variables, such as rainfall intensity, slope gradient, and land cover, to construct more comprehensive and spatially robust predictive models.

References

[1] Centre for Research on the Epidemiology of Disasters (CRED), *2024 Disasters in Numbers*. Brussels, Belgium: CRED, 2025. [Online]. Available: https://files.emdat.be/reports/2024_EMDAT_report.pdf

[2] F. C. G. Gonzalez, M. do C. R. Cavacanti, W. N. Ribeiro, M. B. de Mendonça, and A. N. Haddad, "A systematic review on rainfall thresholds for landslides occurrence," *Heliyon*, vol. 10, no. 1, p. e23247, 2024, doi: 10.1016/j.heliyon.2023.e23247.

[3] M. Ehsan, M. T. Anees, A. F. B. A. Bakar, and M. et al., "A review of geological and triggering factors influencing landslide susceptibility: artificial intelligence-based trends in mapping and prediction," *International Journal of Environmental Science and Technology*, vol. 22, pp. 17347–17382, Dec. 2025, doi: 10.1007/s13762-025-06741-6.

[4] G. Jordanova, S. L. Gariano, M. Melillo, S. Peruccacci, M. T. Brunetti, and M. J. Auflič, "Determination of empirical rainfall thresholds for shallow landslides in Slovenia using an automatic tool," *Water*, vol. 12, no. 5, p. 1449, May 2020, doi: 10.3390/w12051449.

[5] V. Menon and S. Kolathayar, "Empirical and machine learning-based approaches to identify rainfall thresholds for landslide prediction: a case study of Kerala, India," *Discover Applied Sciences*, vol. 7, p. 203, Mar. 2025, doi: 10.1007/s42452-025-06636-8.

[6] K. M. P. Ebrahim, S. M. M. H. Gomaa, T. Zayed, and G. Alfalah, "Recent phenomenal and investigational subsurface landslide monitoring techniques: a mixed review," *Remote Sensing*, vol. 16, no. 2, p. 385, Jan. 2024, doi: 10.3390/rs16020385.

[7] Food and Agriculture Organization of the United Nations (FAO), "Harmonized World Soil Database v1.2," *FAO Soils Portal*, 2008. [Online]. Available: https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/. Accessed: Nov. 8, 2025.

[8] NASA Scientific Visualization Studio, "Global Landslide Catalog (update 2019)," *NASA Goddard Space Flight Center*, Mar. 13, 2019. [Online]. Available: https://svs.gsfc.nasa.gov/4710. Accessed: Nov. 8, 2025.

[9] Natural Earth, "Admin 0 – Countries (version 5.1.1)," *Natural Earth Data*, Sep. 25, 2009. [Online]. Available: https://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-countries/. Accessed: Nov. 8, 2025.

[10] S. B. Bhoite, "Innovative Inter Quartile Range-based Outlier Detection and Removal Technique for Teaching Staff Performance Feedback Analysis," *Journal of Engineering Education Transformations*, vol. 37, no. 3, Mar. 2024, doi: 10.16920/jeet/2024/v37i3/24013.

[11] H. Wang, Q. Liang, J. T. Hancock, et al., "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *Journal of Big Data*, vol. 11, no. 44, Mar. 2024, doi: 10.1186/s40537-024-00905-w.

[12] F. Pargent, F. Pfisterer, J. Thomas, *et al*., "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Computational Statistics*, vol. 37, pp. 2671–2692, Nov. 2022, doi: 10.1007/s00180-022-01207-6.

[13] R. Wall Emerson, "Mann-Whitney U test and t-test," *Journal of Visual Impairment & Blindness*, vol. 117, no. 1, pp. 99–100, 2023, doi: 10.1177/0145482X221150592.

[14] S. Badola, V. N. Mishra, and S. Parkash, "Landslide susceptibility mapping using XGBoost machine learning method," presented at the Conference Paper, Jan. 2023, doi: 10.1109/MIGARST353.2023.1006486.

[15] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross validation optimation on machine learning for prediction," Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 7, no. 4, pp. 11792, Oct. 2022, doi: 10.33395/sinkron.v7i4.11792.

[16] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning*, vol. 113, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.

[17] K. Yang, L. Liu, and Y. Wen, "The impact of Bayesian optimization on feature selection," *Scientific Reports*, vol. 14, no. 3948, Feb. 2024, doi: 10.1038/s41598-024-54515-w.

[18] S. Swaminathan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, Nov. 2024, doi: 10.53555/AJBR.V27I45.4345.

[19] M. J. Froude and D. N. Petley, "Global fatal landslide occurrence from 2004 to 2016," *Natural Hazards and Earth System Sciences*, vol. 18, pp. 2161–2181, 2018, doi: 10.5194/nhess-18-2161-2018.

[20] A. Baldermann, M. Dietzel, and V. Reinprecht, "Chemical weathering and progressing alteration as possible controlling factors for creeping landslides," *Science of The Total Environment*, vol. 778, p. 146300, 2021, doi: 10.1016/j.scitotenv.2021.146300.

[21] 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.

[22] 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

[23] 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

[24] 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.

[25] 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.

[26] 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.

[27] 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.