Journal of Nepal Geological Society, 2024, Vol. 67, pp. 23–34
https://doi.org/10.3126/jngs.v67i1.74581

Presenting the results of landslide susceptibility mapping using partial least squares regression model: A case study of Khotang District, Koshi Province

Sabin Bhattarai[1], *Suchita Shrestha[2], Rabin Niraula[3] and Kabita Karki[2]

[1] The International Centre for Integrated Mountain Development, Khumaltar, Lalitpur
[2] Department of Mines and Geology, Ministry of Industry, Commerce and Supplies
[3] School of Education, Kathmandu University, Hattiban, Lalitpur, Nepal
**Corresponding author's email: suchitashrestha@gmail.com*

ABSTRACT

Landslides pose a significant threat in mountainous regions globally, causing substantial damage to infrastructure, disrupting livelihoods, and leading to loss of life. This study focuses on the Khotang District, Koshi Province, Nepal, a region highly susceptible to landslides due to its steep terrain, active tectonics and heavy monsoon rainfall. The research aims to assess landslide susceptibility in the district using Partial Least Squares Regression model (PLSR), a robust statistical technique capable of handling complex datasets with correlated variables. The research emphasizes the significant influence of topographic factors on landslide occurrence. Specifically, the Topographic Wetness Index (TWI) and Elevation were found to be the most influential variables, demonstrating the highest importance scores in the PLSR model. The model demonstrated excellent performance in predicting landslide susceptibility, with a balance between fit and generalization. It achieved a testing AUC of 0.740, indicating strong generalization ability and potential for practical applications. The findings of this study indicate the potential use of the PLSR for future landslide susceptibility mapping, owing to its robust predictive power. The study also enhances our understanding of the factors that influence landslide occurrences in the Khotang District. Furthermore, it provides a scientific basis for the implementation of effective mitigation measures to reduce landslide risks.

Keywords: Landslide; Partial Least Squares Regression; Susceptibility; Machine Learning

Received: 1 August 2024 **Accepted:** 26 September 2024

INTRODUCTION

The challenging topography, fragile geology, and complex environment have always impacted livelihoods and development activities in the region. In the flat Terai region, factors are relatively simple to calculate. However, the mountainous areas present costly and complex challenges, and the hilly regions have many poorly understood issues that need to be addressed for development activities to proceed.

The Himalayas are seismically active and very fragile due to their inherently weak geological characteristics. High relief, steep slopes, relatively steep river gradient; fragile, active geology and seismically active zone make the Himalayas highly susceptible to geo-hydrological processes i.e. landslides, erosion, debris torrents, flood, and river channel shifting (Dhital, 2003; Kull and Magilligan, 1994; Shrestha et al., 2017). The mountain environment in Nepal, as in other Mountainous regions of the Hindu-Kush Himalayas is fragile and extremely vulnerable to hazards and disasters whether natural or manmade. Prolonged and high-intensity rains in the monsoon season are the most important factors triggering mass movements, gully erosions and floods (Pradhan and Kim, 2020; Starkel, 1972). Although the main triggering factor of landslides is the monsoonal rainfall associated with extreme weather events, a combination of both natural and anthropogenic factors and processes determines the extent and magnitude of such disasters for any affected areas.

High intensity of soil erosion and high incidence of landslides and frequent floods are the geomorphic processes of environmental and socio-economic concerns of both mountains and the adjacent plains (Eckholm, 1975). Landslide susceptibility mapping is considered the first step in landslide hazard assessment. A variety of methodologies are employed in the evaluation and mapping of landslide susceptibility. These include geomorphological mapping, landslide inventory mapping, statistical approaches, heuristic or knowledge-based methods, physically-based slope stability models, and methods utilizing artificial intelligence, deep learning and classification techniques (Guzzetti et al., 1999; Pradhan et al., 2024; Pradhan and Kim, 2021, 2014; van Westen et al., 2008).

Geomorphological mapping provides a detailed understanding of the terrain and its processes, which is crucial in assessing landslide susceptibility (Guzzetti, 2000). Landslide inventory mapping, on the other hand, records past landslide occurrences and their characteristics, providing valuable data for future susceptibility assessments (Van Den Eeckhaut et al., 2009).

Early attempts at landslide susceptibility mapping relied heavily on qualitative methods based on expert knowledge and field observations. However, the increasing availability of spatial data and advancements in computational capabilities led to the adoption of quantitative methods, particularly statistical approaches. Frequency Ratio (FR) and logistic regression (LR) are among the most widely used statistical models in this domain and are commonly used to identify and quantify the relationships between landslide occurrences and various causative factors (Lee and Talib, 2005). Heuristic or knowledge-based methods rely on the expertise of the analyst to rank and weight different factors based on their perceived

importance in causing landslides (Ayalew and Yamagishi, 2005; Pradhan and Kim, 2016). Rasyid et al. (2016) examined the effectiveness of FR and LR models in generating GIS-based landslide susceptibility maps for Lompobattang Mountain, Indonesia. Their research underscored the influence of data characteristics and model assumptions on the performance of these models, emphasizing the need for careful model selection based on the specific study area. Despite the relative simplicity and interpretability of FR and LR models, their limitations in capturing complex non-linear relationships between landslide conditioning factors and susceptibility have led to the exploration of more sophisticated techniques.

Machine learning algorithms, with their ability to discern intricate patterns from large datasets, have emerged as powerful tools for landslide susceptibility mapping (Pradhan et al., 2024). Chowdhury et al. (2024) demonstrated the application of machine learning algorithms, including LR, Random Forest, and Decision Trees, for generating landslide susceptibility maps in the Chattogram District, Bangladesh. Their study showcased the potential of these algorithms in improving prediction accuracy compared to traditional statistical methods. However, the effectiveness of machine learning models is contingent upon the quality and representativeness of the training data. Gameiro et al. (2022) explored the influence of sampling strategies on the performance of Artificial neural networks for landslide susceptibility mapping. Their research underscored the importance of robust sampling techniques to ensure the reliability and generalizability of the resulting susceptibility maps.

Traditional methods for landslide susceptibility mapping have relied heavily on qualitative approaches, which are based on expert knowledge and field observations. While these methods have provided valuable insights, they are often subjective and limited in their ability to handle the complex interplay of factors contributing to landslides (Sujatha and Sudharsan, 2024) . To minimize those gaps in traditional methods, this study involves a novel Partial Least Squares Regression (PLS). PLS is a robust multivariate regression method that is particularly effective when predictors exhibit collinearity.

STUDY AREA

Khotang District is located in the eastern Nepal which spans 1,591 square kilometers of predominantly hilly terrain. It is situated between 26° 50" N to 27° 28" N and 86° 26" E to 86° 58" E (Fig. 1). The district's elevation ranges from 161 meters to 3,620 meters above sea level (masl). The Sunkoshi and Dudh Koshi Rivers form natural boundaries to the north, west, and south, while a series of hills and smaller waterways delineate its eastern border with Bhojpur District. Forest covers approximately 56% of Khotang, with cultivated land accounting for around 42% of its area.

Khotang District is composed of various sub-watersheds of Koshi Basin. Dudh Koshi River runs from Solukhumbu along the Western Border of Khotang District demarking Khotang from Okhaldhunga towards the west and drains into the Sunkoshi River and runs along the border demarking Udayapur towards the south. Rawakhola sub-watershed, Supsup Khola sub-watershed, Tuwa Khola sub-watershed and Sawa Khola sub-watershed are the major sub-watershed that drain into Sunkoshiriver. Other tributaries of Sunkoshi in the district are Dikhuwa Khola, Tawa Khola, Tap Khola, Buwa Khola, etc. The drainage pattern is dendritic type.

Figure 2 illustrates the number of landslides, the number of deaths, and the estimated financial loss in millions of Nepalese Rupees (NRs) in the Khotang District from 2011 to 2023. The data reveals significant variability over the years. Notably, 2019 stands out with a peak in both landslides and associated financial loss, indicating a severe impact that year. Other years, such as 2011, 2014 and 2017, also show notable occurrences

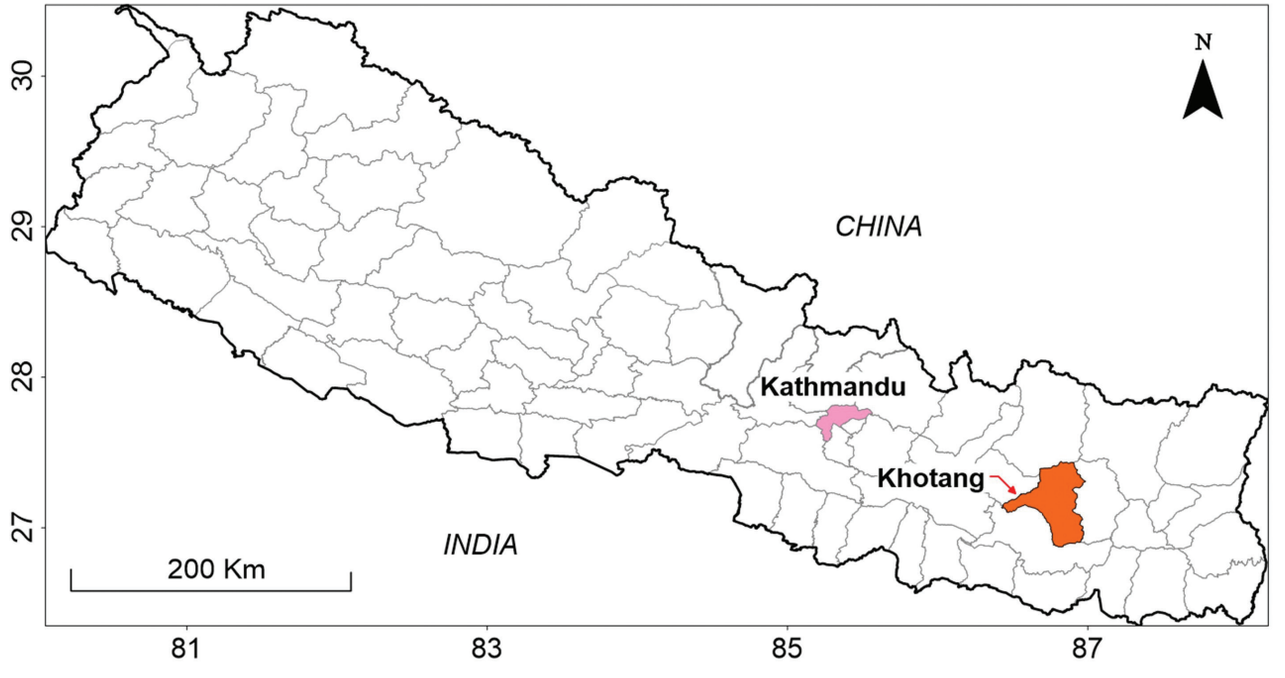

Fig. 1: Location of the study area.

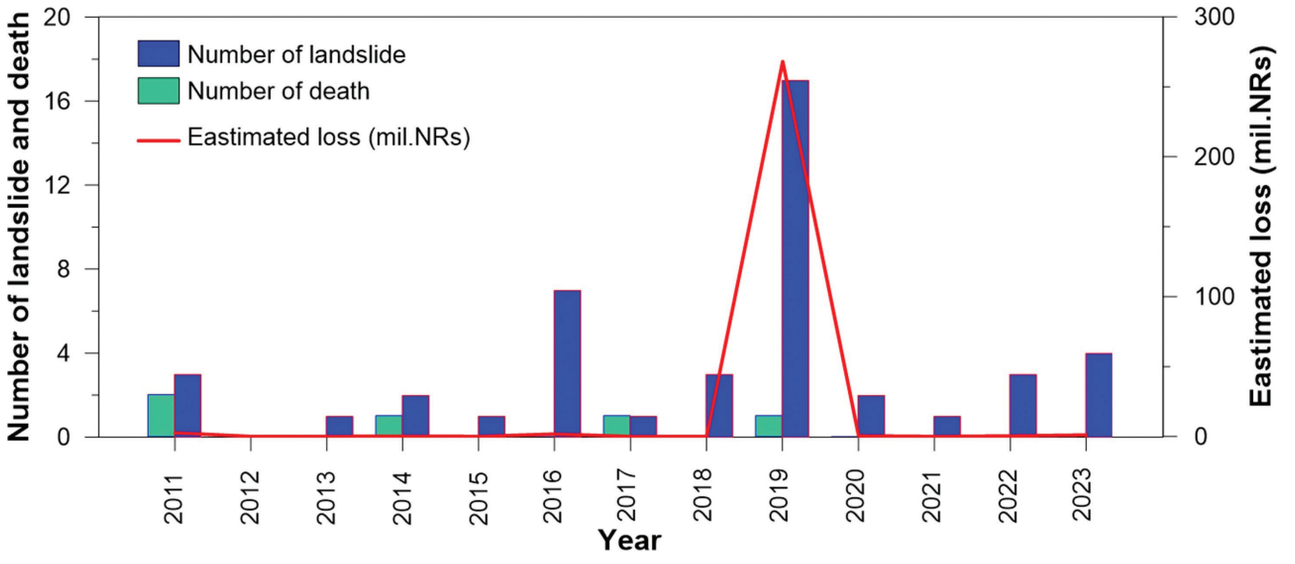

Fig. 2: Number of landslide, death and estimated loss in Khotang District.

of landslides and deaths, though with comparatively lower estimated financial losses. The trend suggests intermittent severe events with varying impacts on human life and economic loss throughout the observed period.

METHODOLOGY AND DATASET

Method

Figure 3 presents a two-phase framework for assessing landslide susceptibility. In Phase-1, Data Collection, various independent variables are gathered, categorized into three groups: Geomorphic (elevation, slope, curvature), Hydrologic (DrainProx, VDCN, TWI, STI, SPI), and Geologic (geology, Faultprox). These variables are used to compile the dependent variable, the landslide inventory. In Phase-2, the Modeling Framework, the collected data is split into training and test sets, with validation ensuring model accuracy. The PLS method is applied to analyze the data, producing coefficients and importance values. These outputs contribute to determining landslide susceptibility.

Dataset

Landslide Inventory

The creation of comprehensive landslide inventories is a critical step in understanding landslide patterns, assessing risks, and guiding mitigation efforts (Guzzetti et al., 2012). Traditionally, these inventories have been developed only through field investigations which are time-consuming, and often limited to accessible areas (Martha and Kerle, 2012). However, the advent of high-resolution satellite imagery has revolutionized this process, providing extensive spatial coverage and enabling the detection of landslides based on changes in land cover, vegetation patterns, and topographic features. Google Earth Pro, with its high-resolution imagery and global coverage, has emerged as a valuable tool for visually identifying and mapping landslides, particularly when leveraging the platform's historical imagery capabilities. The methods for acquiring landslide inventory can be broadly categorized as field surveys and image interpretation techniques. Satellite images, remote sensing and Google Earth™ were used for digitizing 80 landslide polygons in GIS. The digitized landslides were confirmed by field verification in several locations. Figure 4 presents the distribution of landslides in the study area.

Independent variables

This research has been supported by incorporating independent variables derived from terrain analysis into the modeling procedures. All morphometric variables were derived, in the first case study, from a detailed Digital Elevation Model (DEM) produced by the Department of Survey, Government of Nepal (20 × 20 m).

Factors influencing the likelihood and behavior of landslides are categorized as causative variables. These variables encompass a range of features, including topography (elevation, slope, curvature, drainage), geology, geomorphology, and human activities. Essentially, these factors create the preconditions that make landslides possible in a given area. While standardized variables like elevation, slope and drainage are commonly considered, researchers often select causative factors for landslide susceptibility mapping based on subjective assessments and local knowledge. As a result, the selection of landslide causative variables and their classifications is critical in landslide susceptibility modelling research. In this study, a total of 10 variables namely elevation, slope, curvature, DrainProx, VDCN, TWI, STI, SPI, Faultprox and geology were selected based on relevancy and availability. Among 10 variables, geology is categorical so geology was converted into a dummy variable for further analysis.

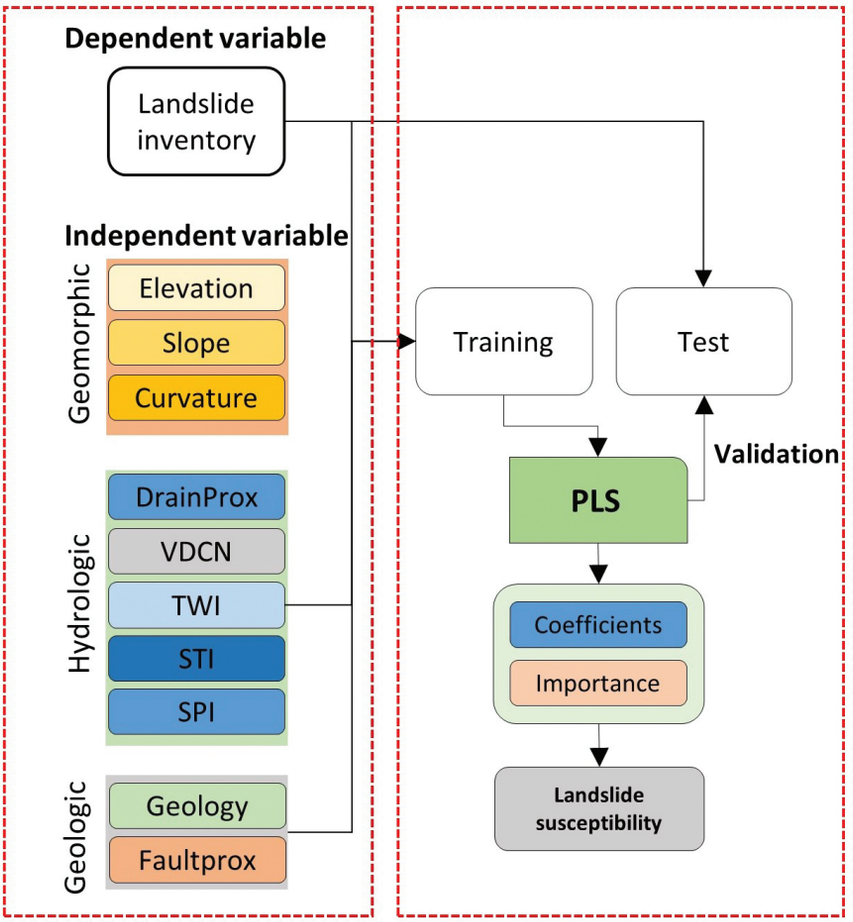

Fig. 3: Architect of the research procedure.

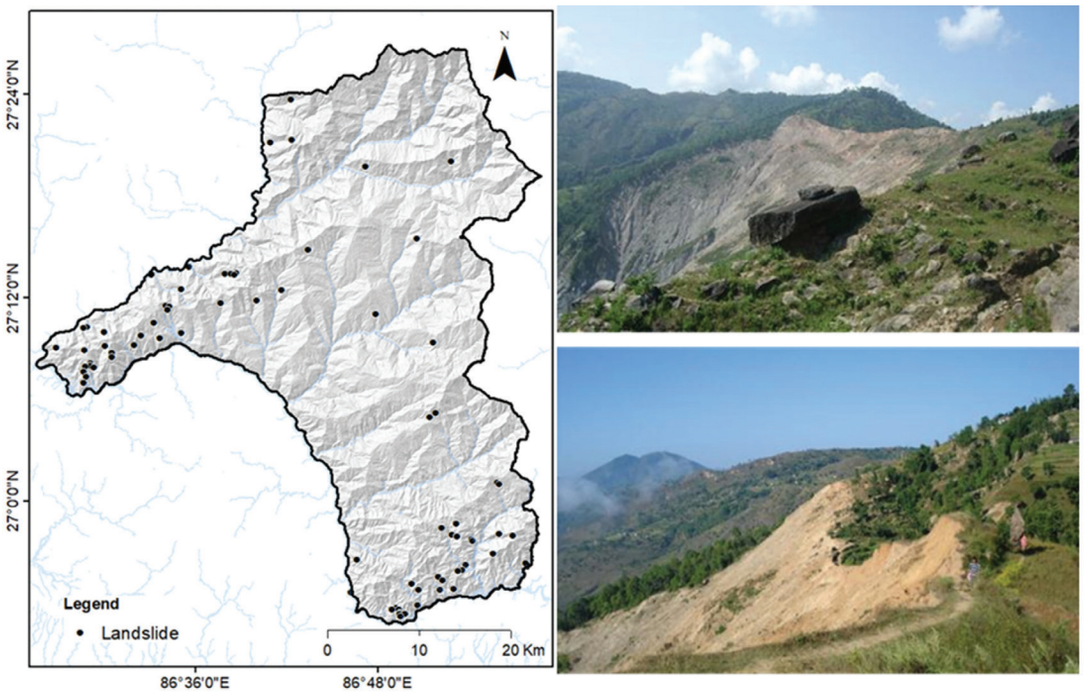

Fig. 4: Landslide inventory map of the study area.

Geomorphic variables

Elevation is a significant factor influencing the occurrence of landslides. The relationship between elevation and landslides is not always straightforward and can vary significantly based on regional factors like climate, geology, and topography. While it's not the sole determinant, it plays a crucial role because higher elevations often receive more rainfall, leading to increased soil saturation. This can weaken soil structure and contribute to landslides. Higher elevations typically have steeper slopes, making them more prone to landslides. The elevation ranges from 161 m above sea level (asl) to 3615 m asl as shown in Fig. 5a.

The slope is a primary factor influencing the occurrence of landslides. The steeper the slope, the greater the potential for gravitational forces to overcome the stability of the materials on it. As the slope angle increases, the shear stress (the force acting parallel to the slope) also increases. This force tends to pull the material downhill. The slope distribution in the study area ranges from 0 to about 69˚ as shown in Fig. 5b.

Curvature and landslides are interconnected in the field of geomorphology, as the curvature of a slope can significantly influence the occurrence and characteristics of landslides. Curvature refers to the degree of bend or the change in slope angle over a specific distance. The distribution of curvature is shown in Fig. 5c.

Hydrologic variables

Drainage proximity is a significant factor influencing landslide susceptibility. The closer an area is to a drainage network (rivers, streams, etc.), the higher the risk of landslides. Proximity to drainage often correlates with higher groundwater levels. This excess water can saturate the soil, reducing its stability and increasing landslide risk. Water flowing through drainage channels can undercut the base of slopes, leading to slope failure. The drainage proximity is presented in Fig. 6a.

Vertical distance to channel network (VDCN) is a crucial geospatial parameter quantifying the elevation difference between a specific point and the nearest river or stream network point. In simpler terms, it measures how high a location is above the closest watercourse. Areas with low VDCN are more susceptible to erosion by the river or stream, which can destabilize slopes. The distribution of VDCN is depicted in Fig. 6b.

The Topographic Wetness Index (TWI) quantifies the potential for water accumulation on a slope, a critical factor in triggering landslides. Higher TWI values indicate areas prone to waterlogging. This excess water can increase soil saturation, reducing its stability and making it more susceptible to landslides. TWI is calculated using the DEM of a terrain (Fig. 6c). It combines information about the slope's steepness and the area contributing to flow at a specific point. The TWI can be calculated using Eq. (1).

$$TWI = \ln\left(\frac{\alpha}{\tan\beta}\right), \tag{1}$$

where is the accumulated catchment area (area contributing flow to the point) and β is the slope angle.

The Sediment Transport Index (STI) is a quantitative measure used to assess the potential of a slope or watershed to transport sediment. It's a valuable tool in hydrology, geomorphology, and environmental science for understanding erosion and sediment yield. It quantifies the potential for sediment movement on a slope. The sediment transport index (STI) depends on the catchment size and slope angle in a nonlinear fashion (Moore and Burch, 1986), as shown in Eq. (2).

$$STI = \left(\frac{A_s}{22.13}\right)^{0.6} \times \left(\frac{\sin\beta}{0.0896}\right)^{1.3}, \tag{2}$$

where A_s is a contributing area and is the slope gradient in radians. The distribution of *STI* is presented in Fig. 6d.

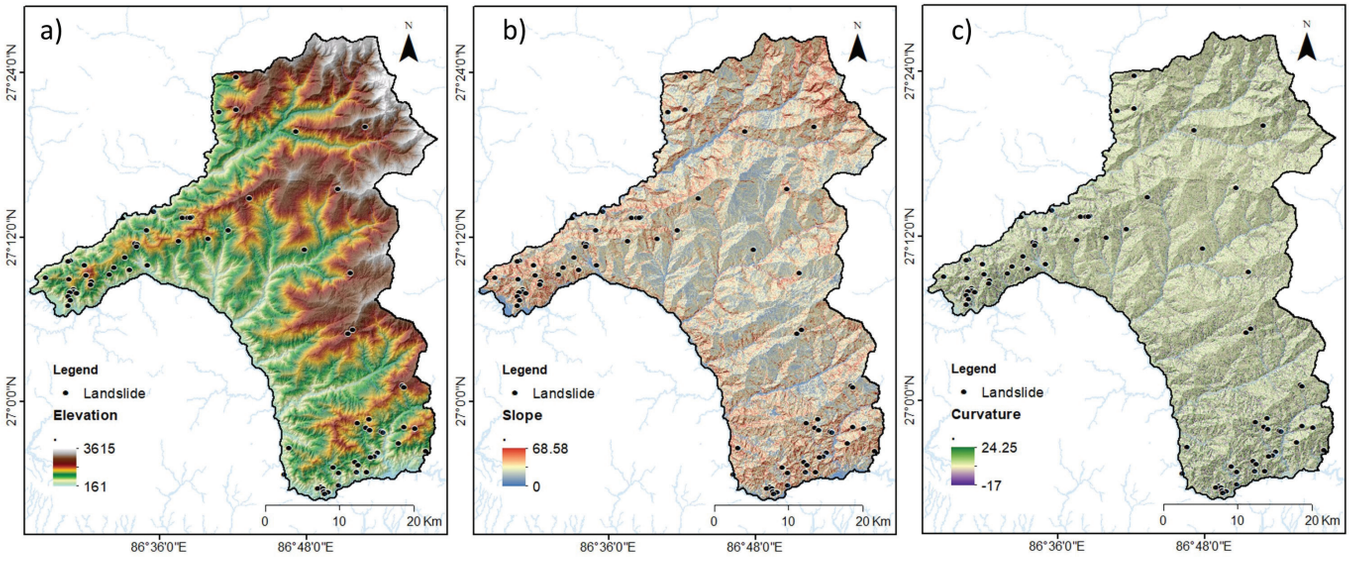

Fig. 5 Geomorphic variables a) elevation, b) slope and c) curvature.

The Stream Power Index (SPI) is a quantitative measure of the erosive power of flowing water. It's a valuable tool in geomorphology and hydrology, helping to understand the potential for erosion and sediment transport within a river system. The SPI can be calculated using the Eq. (3) as given below:

$$SPI_i = \ln\left(DA_i \times \tan\left(G_i\right)\right), \quad (3)$$

where SPI_i is the stream power index at grid cell i, DA_i is the upstream drainage area at grid cell i and G_i is the slope at grid cell i in radians. The spatial distribution of SPI is depicted in Fig. 6e.

Geologic variables

Fault proximity (Fig. 7a) is a critical factor influencing landslide susceptibility, particularly in tectonically active regions. The presence of faults significantly weakens rock masses due to intense fracturing and shearing, reducing their overall strength and stability. This weakening makes slopes more susceptible to weathering and erosion, further exacerbating landslide hazards. Earthquakes, a common occurrence in fault zones, generate strong ground motions that can trigger landslides, especially in areas with steep slopes and weak geological materials. Faults can also influence hydrogeological processes, acting as conduits or barriers to groundwater flow and leading to localized changes in pore water pressure within slopes. Increased pore pressure reduces the effective stress holding soil and rock masses together, further increasing landslide susceptibility. Therefore, accurate landslide susceptibility mapping in tectonically active regions must consider not only the distance to faults but also their activity levels, seismic history, and the surrounding geological context to assess and mitigate landslide hazards effectively.

Lithology, the study of the physical and chemical composition of rocks, plays a pivotal role in landslide susceptibility. The inherent characteristics of different rock types, such as their strength, weathering rates, and permeability, significantly influence slope stability. Variations in mineral composition, grain size, and degree of fracturing within the same rock type can also impact landslide susceptibility.

The study area is geologically dominated by the Tawa Khola Formation (Ta), which is primarily located in the southern part of the area and consists of coarse-grained, dark grey

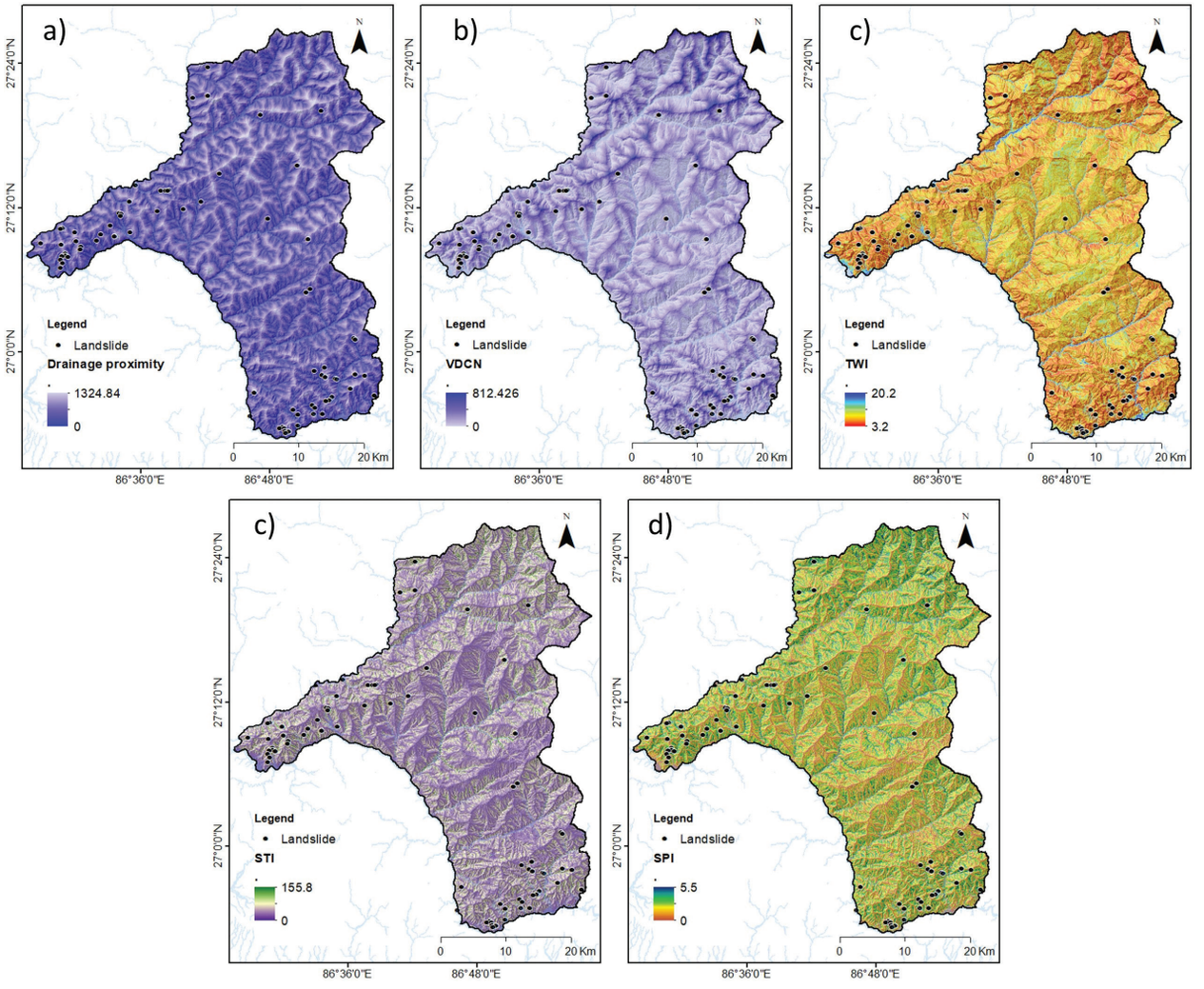

Fig. 6: Hydrologic variables a) drainage proximity, b) VDCN, c) TWI, d) STI and e) STI.

garnetiferous muscovite biotite quartz schists. This formation is followed by the Seti Formation (St), which includes grey to greenish-grey phyllites and gritty quartzites. Next in the sequence is the Shiprin Khola Formation (Sp), characterized by coarse-textured, highly garnetiferous muscovite biotite-quartz schists. Following this is the Sarung Khola Formation (Sk), consisting of fine-textured, dark grey to greenish-white quartz biotite schists. The Ulleri Formation (Ul) comes next, which includes feldspathic schists with augens of feldspar and quartz. The Maksang Formation (Mk), is characterized by grey to grayish-white, fine-grained quartzites. The area also features abundant Granite Intrusions (Gr). After the granite, the sequence continues with the Udaipur Formation (Ud), consisting of grey, grayish-black crystalline limestones. This is followed by the Kushma Formation (Ks), which includes greenish grey, white fine to medium-grained quartzites. Next is the Sangram Formation (Sg), characterized by grey to greenish-grey carbonaceous shales. Lastly, the area includes a small amount of rocks belonging to the Lower Siwalik (Ls), consisting of fine-grained sandstone. The geological map of the study area is presented in Fig. 7b (DMG, 2020).

Partial Least Square Regression Model

Partial Least Squares (PLS) is a method used in machine learning that amalgamates the benefits of principal component analysis, conventional correlation analysis, and linear regression analysis. PLS maps both the predicted and observed variables into a novel space. This is accomplished by identifying pairs of weight vectors that optimize the covariance between the two projections. The PLS regression is an extension of the multiple linear regression model.

This approach connects two data matrices, x and y, using a linear multivariate model. Each parameter in the model is determined by the gradient of a straightforward bivariate regression (least squares) between a column or row of the matrix as the y-variable, and another parameter vector serving as the x-variable (Wold et al., 2001). In its simplest form, a linear model specifies the (linear) relationship between a dependent (response) variable Y, and a set of predictor variables, the *X*'s, so that

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p \quad (4)$$

In this equation, b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables 1 through p) computed from the data.

In PLS modeling, the importance of a predictor for the dependent variables is indicated by the variable importance in the projection (VIP). Factors that have VIP values exceeding 1 are deemed to be the most significant in elucidating the dependent variable. They are viewed as notably influential predictors within the PLS model (Wold, 1995). The VIP and regression coefficients were used to explain the relative influence of each independent variable.

Assessment of the accuracy of the model

Assessment of model landslide susceptibility accuracy involves evaluating its ability to correctly predict landslide occurrence. This is typically done by comparing the model's output to a known dataset of historical landslides. Evaluation metrics were employed to assess the proposed models utilizing contingency matrices, which include True Positives (TP), True Negatives

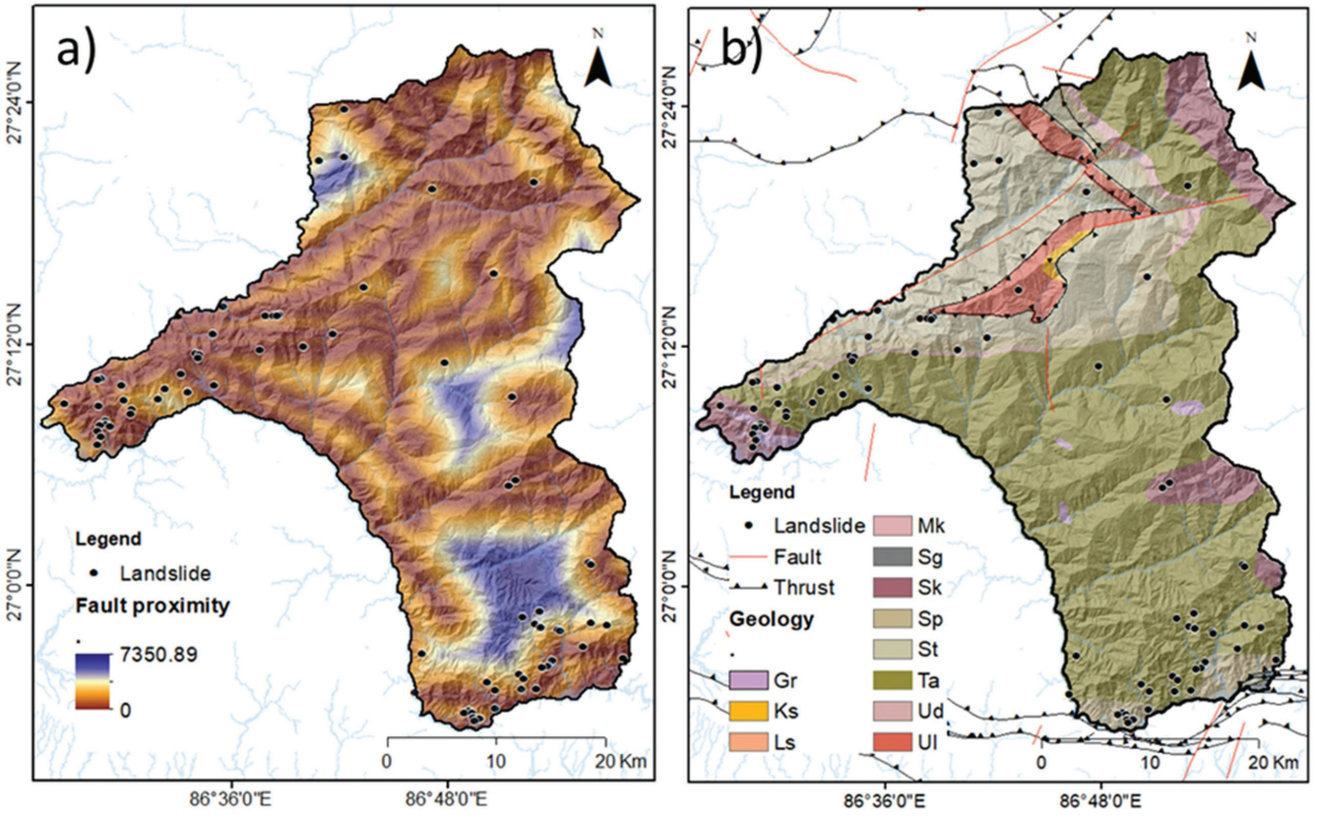

Fig. 7: Geologic variables a) fault proximity and b) geology (DMG 2020).

(TN), False Positives (FP) and False Negatives (FN). The diagnostic ability of the models was depicted using receiver operating characteristic (ROC) curves, which illustrate the true positive rate against the false positive rate. The area under the curve (AUC) represents the prediction values, with larger AUC values indicating better predictions.

This approach involves iteratively splitting a dataset into training and test sets. The training data is utilized to train the model, which is subsequently applied to the test set for evaluation. For this purpose, the entire dataset was divided into training (70%) and test (30%) sets using a random sampling technique. The training set is utilized to instruct a machine learning model to perform a specific task, such as adjusting the system's parameters based on input and output data. The test set is then employed to assess the performance of the machine learning model once it has been trained on the training data.

A ROC curve is a graphical representation of the performance of a binary classification model (Hosmer et al., 2000). In ROC curves, the true positive rate (TPR: sensitivity) is plotted against the false positive rate (FPR: 1-specificity) at different classification thresholds. The TPR is the ratio of true positive predictions to the total number of actual positive cases, and the FPR is the ratio of false positive predictions to the total number of actual negative cases. AUC values are typically measured in the range of 0.5–1. According to Yesilnacar and Topal (2005), there is a relation between prediction accuracy and AUC value that may be classified as follows: 0.5–0.6 (poor), 0.6–0.7 (average), 0.7–0.8 (good), 0.8–0.9 (outstanding) and 0.9–1 (excellent).

RESULTS

TWI has the highest importance score, indicating that it is the most critical factor in the PLS model (Fig. 8). Elevation also shows a high importance score, making it a key predictor. Slope, faultprox, curvature have relatively high importance scores and contribute significantly to the model. Sp, Ta, Gr have moderate importance scores, indicating they are still influential but less critical than the top contributors. UI, Sk, St show a noticeable drop in importance but still play a role in the model. VDCN, Mk, Ud have lower importance scores, suggesting they contribute less to the model's predictive power. Drainprox, Sp and St have the lowest importance scores, indicating minimal contribution to the model. Understanding these variable importance scores can help in focusing on the most impactful factors in future analyses or model improvements.

Each bar has an associated error bar, representing the variability or uncertainty in the importance score. Larger error bars suggest greater uncertainty in the importance estimate. The dashed line across the graph serves as a threshold for significance. Variables with importance scores above this line are considered significant contributors to the model.

As shown in the Fig. 9, the bar chart highlights the relationships between various predictors and the response variable in the PLS model. Positive coefficients suggest variables that increase the response variable when they increase, while negative coefficients indicate the opposite. Variables like Elevation, TWI, and Faultprox show strong positive relationships, whereas Slope and Curvature show strong negative relationships. Variables with coefficients near zero have little to no direct relationship with the response variable, suggesting they are less critical in the model.

The landslide susceptibility map of the study area was prepared using the coefficients of independent variables in Python scripts, which normalize the susceptibility values from 0 to 1. To identify variations in landslide susceptibility across the entire study region, the output values of the model were reclassified into five levels: very low, low, moderate, high and very high, using the natural break classification system (Jenks, 1967) as shown in Fig. 10.

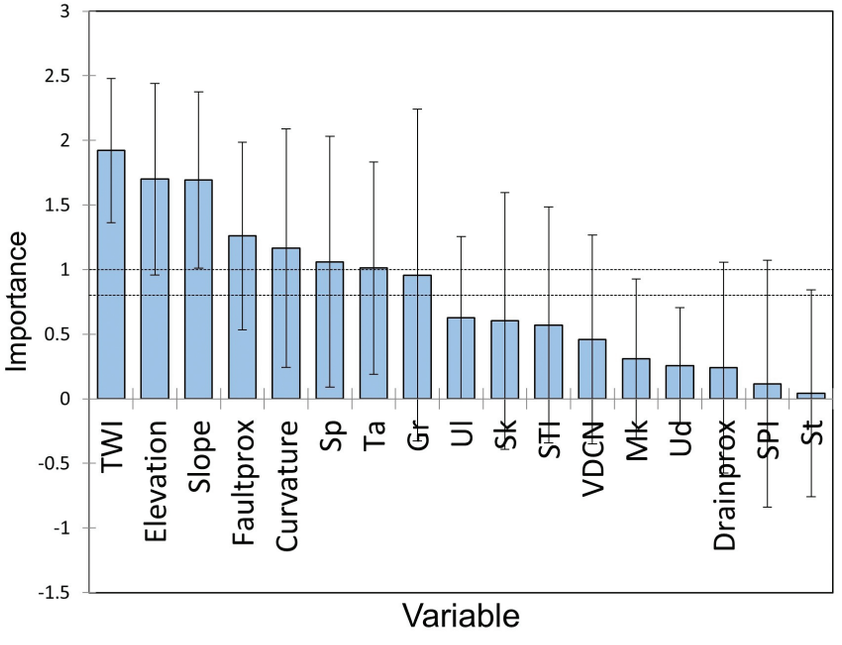

Fig. 8: Variable importance.

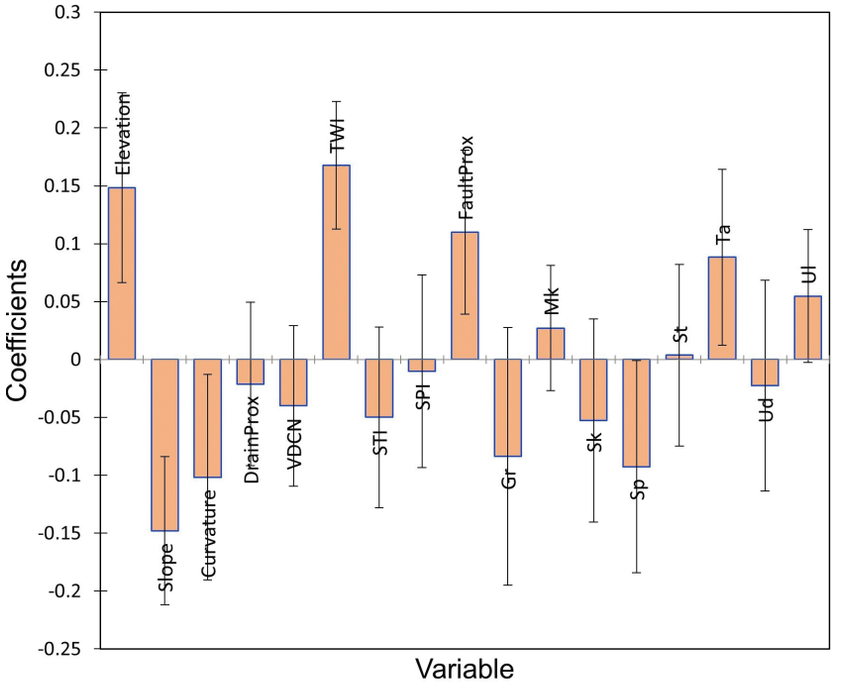

Fig. 9: Bar chart showing the coefficients of each variable.

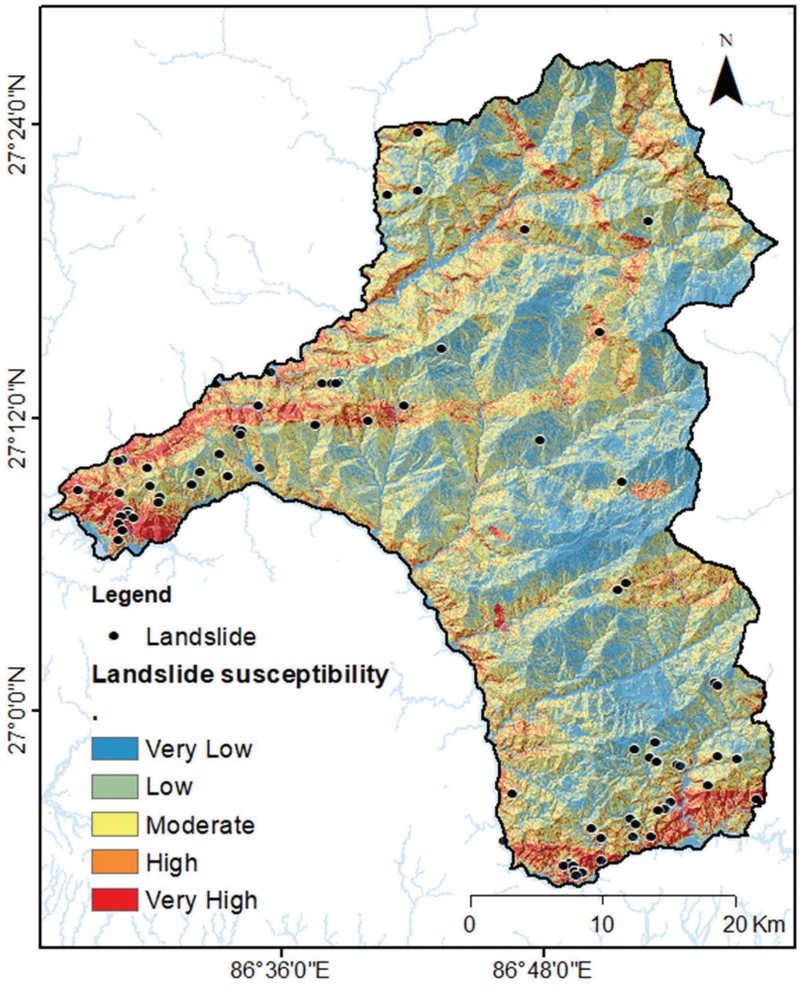

Fig. 10: Landslide susceptibility map of the study area.

Figure 11 depicts the relationship between different susceptibility classes and the corresponding percentages of landslide occurrences. The susceptibility classes are categorized as Very Low, Low, Moderate, High, and Very High. In the Very Low susceptibility class, the area percentage is 21.11%, while the percentage of landslides is significantly lower at 1.266%. The Low susceptibility class has an area percentage of 33.23% and a landslide percentage of 11.39%. For the Moderate susceptibility class, the area percentage is 25.75%, and the landslide percentage is higher at 27.85%. The High susceptibility class shows an area percentage of 15.07% and a landslide percentage of 29.11%. Lastly, the Very High susceptibility class has the lowest area percentage at 4.846%, but the highest percentage of landslides at 30.38%.

The ROC curve has been used to estimate the model's accuracy, which is used as a quantitative measurement. The ROC curves of the model built in this study are shown in Fig. 12. It can be seen that the AUC of the training dataset in the PLS model

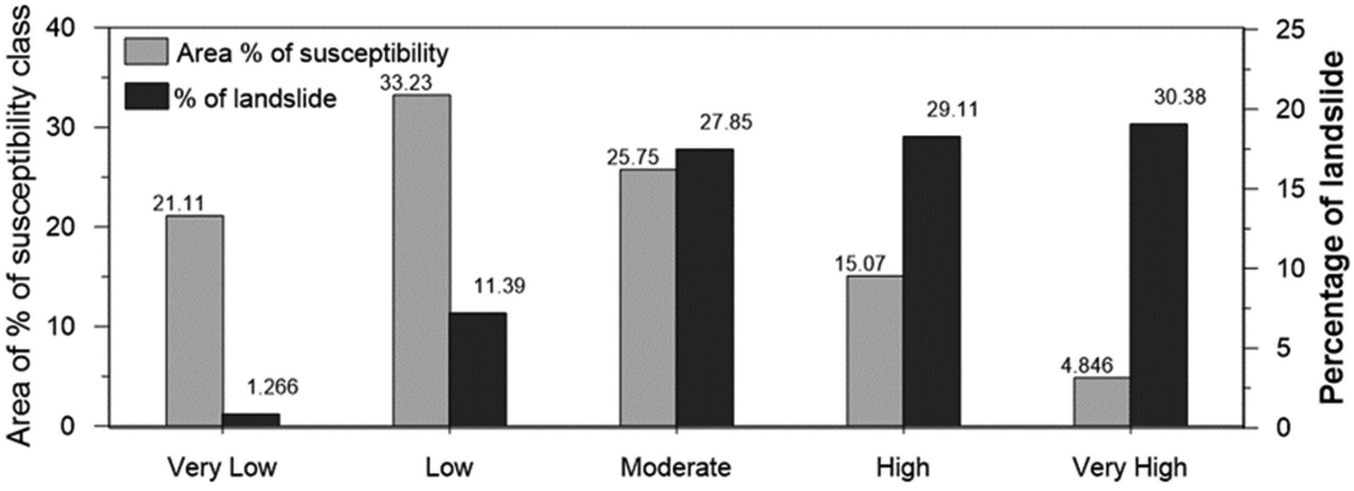

Fig. 11: Terrain percentage and number of landslides.

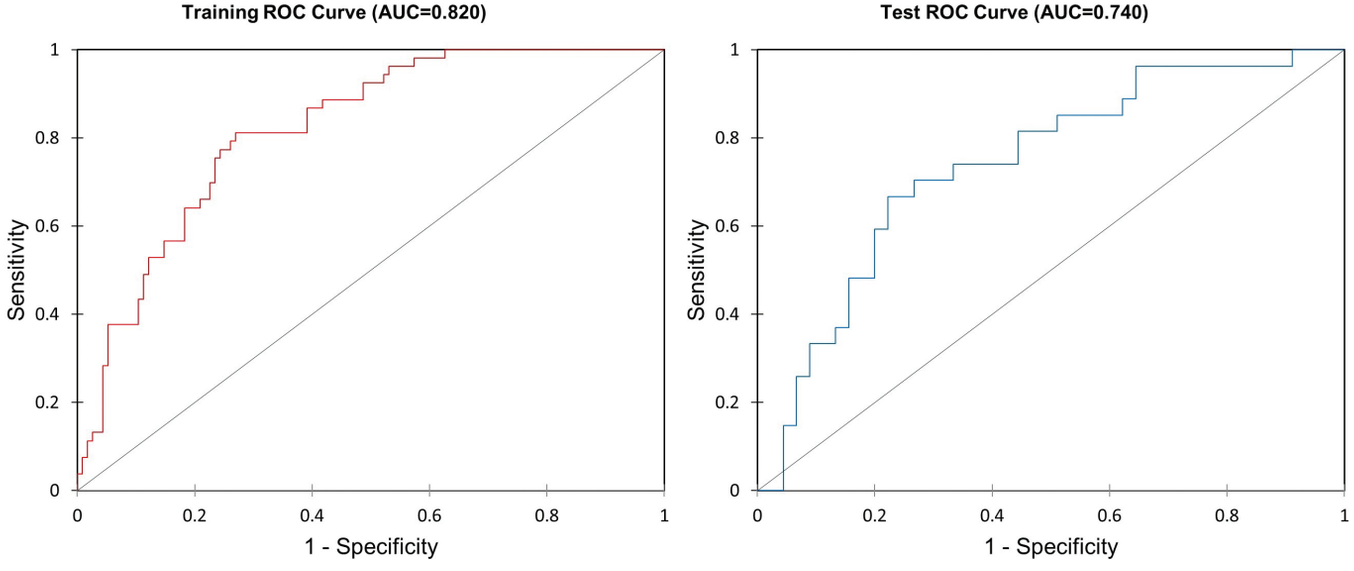

Fig. 12: AUC of training dataset and test dataset.

is 0.82. Thus, the accuracy rate of the PLS model fell within the ‘‘outstanding’’ classification category. The AUC for the test set exhibits 0.74, which fell within the “good” classification category.

DISCUSSION

This study investigated the key factors influencing landslide susceptibility using a PLS model. The findings highlight the critical role of topography, particularly TWI and Elevation, in landslide occurrence within the study area. These variables exhibited the highest importance scores in the PLS model, indicating their strong influence on landslide susceptibility.

The positive coefficients observed for Elevation, TWI, and Faultprox suggest a direct relationship with landslide likelihood. Higher elevations often experience increased precipitation and lower temperatures, potentially leading to increased soil saturation and weathering, thus weakening slope stability. Similarly, higher TWI values indicate areas prone to water accumulation, increasing pore pressure and reducing the shear strength of slopes, making them more susceptible to landslides. The proximity to faults also elevates landslide susceptibility due to the potential for ground shaking during seismic events.

Conversely, Slope and Curvature exhibited a negative relationship with landslide susceptibility. While counterintuitive at first glance, this finding suggests that gentler slopes and more convex terrain in our study area might be associated with more stable geological formations or land cover types, potentially mitigating landslide risk. However, further investigation into the specific geomorphological and geological characteristics of these areas is needed to confirm this hypothesis.

The study findings strongly resonate with existing literature on landslide susceptibility mapping. Chicas et al. (2024) emphasized the importance of slope, elevation and lithology as consistently significant predictors. Notably, they highlighted the consistent ranking of elevation and slope across various studies, aligning with their importance in the model. Nevertheless, among the significant predictors of LSM as mentioned above, they highlighted, that road density, elevation, and slope exhibited the least ranking variability as LSM predictors. Furthermore, Emberson et al. (2022) underscored the predictive power of the average upstream angle and compound topographic index, both closely related to slope steepness and water accumulation potential, reinforcing the significance of TWI in the study area. Migoń and Michniewicz (2016) further validated the utility of TWI in landslide studies, emphasizing its role in identifying preferential drainage pathways within landslide bodies.

The model demonstrated strong performance with a training ROC, indicating the model’s effective learning from the training dataset and its ability to capture the key factors driving landslides in the study area. The model also showed a good generalization ability with a testing AUC of 0.740, suggesting its potential for practical applications in predicting landslide susceptibility in unseen data. These results underscore the model’s strong performance and its ability to handle complex datasets characterized by high dimensionality and multicollinearity, often encountered in landslide susceptibility studies. Unlike traditional statistical methods, such as Logistic Regression, which often struggle with such datasets, PLS effectively manages correlated predictors while simultaneously considering their relationship with the response variable.

CONCLUSIONS

This study has successfully demonstrated the use of the PLS model to assess landslide susceptibility in the Khotang District, Koshi Province, Nepal. The region, characterized by its challenging topography, active geology and heavy monsoon rainfall, frequently experiences landslides. The research focused on understanding the influence of geomorphic, hydrologic and geologic factors on landslide occurrences. The PLSR model demonstrated strong performance with a training AUC of 0.82. This indicates the model’s effective learning from the training dataset and its ability to capture the key factors driving landslides in the study area. The model also showed a good generalization ability with a testing AUC of 0.74, suggesting its potential for practical applications in predicting landslide susceptibility in unseen data. The PLS model shows promising potential for landslide susceptibility mapping, achieving good predictive performance on both training and testing datasets. Hence, the study underscores the robust influence of topographic factors on landslide susceptibility and validates the effectiveness of the PLS method in handling complex datasets characterized by high dimensionality and multicollinearity, often encountered in landslide susceptibility studies.

ACKNOWLEDGMENT

The authors are thankful to anonymous reviewers for their valuable comments that were very useful in bringing the manuscript into its present form.

AUTHOR’S CONTRIBUTIONS

S. Shrestha conceptualized the research. R. Niraula and S. Bhattarai conducted the field study and performed data analysis. The manuscript was drafted by S. Bhattarai and K. Karki under the supervision of S. Shrestha.

REFERENCES

Ayalew, L. and Yamagishi, H., 2005, The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65, pp. 15–31. https://doi.org/10.1016/J.GEOMORPH.2004.06.010

Chicas, S.D., Li, H., Mizoue, N., Ota, T., Du, Y. and Somogyvári, M., 2024, Landslide susceptibility mapping core-base factors and models’ performance variability: a systematic review. Nat. Hazards, pp. 1–21.

Chowdhury, M.S., Rahaman, M.N., Sheikh, M.S., Sayeid, M.A., Mahmud, K.H. and Hafsa, B., 2024, GIS-based landslide susceptibility mapping using logistic regression, random forest and decision and regression tree models in Chattogram District, Bangladesh. Heliyon 10.

Dhital, M.R., 2003, Causes and consequences of the 1993 debris flows and landslides in the Kulekhani watershed, central Nepal, in: Proc. 3rd Intl. Conf. Debris-Flow Hazards Mitigation: Mechanics, Prediction and Assessment, Edited by: Rickenmann, D. and Chen C.-L., Millpress, Rotterdam, Netherlands. pp. 931–942.

DMG, 2020, Department of Mines and Geology, Ministry of Industry, Commerce and Supplies. https://www.dmgnepal.gov.np.

Eckholm, E.P., 1975, The deterioration of mountain environments: ecological stress in the highlands of Asia, Latin America, and Africa takes a mounting social toll. Science (80-), 189, pp.764–770.

Emberson, R., Kirschbaum, D.B., Amatya, P., Tanyas, H. and Marc, O., 2022, Insights from the topographic characteristics of a large global catalog of rainfall-induced landslide event inventories. Nat. Hazards Earth Syst. Sci. 22, pp. 1129–1149.

Gameiro, S., de Oliveira, G.G. and Guasselli, L.A., 2022, The influence of sampling on landslide susceptibility mapping using artificial neural networks. Geocarto Int. pp. 1–23.

Guzzetti, F., 2000, Landslide fatalities and the evaluation of landslide risk in Italy. Eng. Geol. 58, pp. 89–107. https://doi.org/10.1016/S0013-7952(00)00047-8

Guzzetti, F., Carrara, A., Cardinali, M. and Reichenbach, P., 1999, Landslide hazard evaluation: A review of current techniques and their application in a multi-scale study, Central Italy, in: Geomorphology. pp. 181–216. https://doi.org/10.1016/S0169-555X(99)00078-1

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M. and Chang, K.T., 2012, Landslide inventory maps: New tools for an old problem. Earth-Science Rev. https://doi.org/10.1016/j.earscirev.2012.02.001

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2000, Applied logistic regression. Wiley New York.

Jenks, G.F., 1967, The Data Model Concept in Statistical Mapping. Int. Yearb. Cartogr. 7, pp. 186–190.

Kull, C.A. and Magilligan, F.J., 1994, Controls over landslide distribution in the White Mountains, New Hampshire. Phys. Geogr. 15, pp. 325–341.

Lee, S. and Talib, J.A., 2005, Probabilistic landslide susceptibility and factor effect analysis. Environ. Geol. 47, pp. 982–990. https://doi.org/10.1007/s00254-005-1228-z

Martha, T.R. and Kerle, N., 2012, Creation of event-based landslide inventory from panchromatic images by object oriented analysis. Proc. 4th GEOBIA, pp. 7–9.

Migoń, P. and Michniewicz, A., 2016, Topographic Wetness Index and Terrain Ruggedness Index in geomorphic characterisation of landslide terrains, on examples from the Sudetes, SW Poland. https://doi.org/10.1127/zfg_suppl/2016/0328

Moore, I.D. and Burch, G.J., 1986, Physical Basis of the Length-slope Factor in the Universal Soil Loss Equation1. Soil Sci. Soc. Am. J. 50, p. 1294. https://doi.org/10.2136/sssaj1986.03615995005000050042x

Pradhan, A.M.S. and Kim, Y.T., 2021, An artificial intelligence-based approach to predicting seismic hillslope stability under extreme rainfall events in the vicinity of Wolsong nuclear power plant, South Korea. Bull. Eng. Geol. Environ. 80, 3629–3646. https://doi.org/10.1007/s10064-021-02138-0

Pradhan, A.M.S. and Kim, Y.T., 2020, Rainfall-induced shallow landslide susceptibility mapping at two adjacent catchments using advanced machine learning algorithms. ISPRS Int. J. Geo-Information 9, p. 569. https://doi.org/10.3390/ijgi9100569

Pradhan, A.M.S. and Kim, Y.T., 2016, Evaluation of a combined spatial multi-criteria evaluation model and deterministic model for landslide susceptibility mapping. Catena. https://doi.org/10.1016/j.catena.2016.01.022

Pradhan, A.M.S. and Kim, Y.T., 2014, Relative effect method of landslide susceptibility zonation in weathered granite soil: A case study in Deokjeok-ri Creek, South Korea. Nat. Hazards 72, pp. 1189–1217. https://doi.org/10.1007/s11069-014-1065-z

Pradhan, A.M.S., Shrestha, S., Lee, J.H., Hwang, I.T. and Park, H.J., 2024, Utilizing artificial intelligence techniques for soil depth prediction and its influences in landslide hazard modeling. Stoch. Environ. Res. Risk Assess. pp. 1–18.

Rasyid, A.R., Bhandary, N.P. and Yatabe, R., 2016, Performance of frequency ratio and logistic regression model in creating GIS based landslides susceptibility map at Lompobattang Mountain, Indonesia. Geoenvironmental Disasters 3, pp. 1–16.

Shrestha, S., Kang, T.-S. and Suwal, M., 2017, An Ensemble Model for Co-Seismic Landslide Susceptibility Using GIS and Random Forest Method. ISPRS Int. J. Geo-Information 6, p. 365. https://doi.org/10.3390/ijgi6110365

Starkel, L., 1972, The role of catastrophic rainfall in the shaping of the relief of the Lower Himalaya (Darjeeling Hills). Geogr. Pol. 21, pp. 103–147.

Sujatha, E.R. and Sudharsan, J.S., 2024, Landslide Susceptibility Mapping Methods—A Review. Landslide Susceptibility, Risk Assess. Sustain. Appl. Geostatistical Geospatial Model, pp. 87–102.

Van Den Eeckhaut, M., Reichenbach, P., Guzzetti, F., Rossi, M., and Poesen, J., 2009, Combined landslide inventory and susceptibility assessment based on different mapping units: An example from the Flemish Ardennes, Belgium. Nat. Hazards Earth Syst. Sci. 9, pp. 507–521. https://doi.org/10.5194/nhess-9-507-2009

Van Westen, C.J., Castellanos, E. and Kuriakose, S.L., 2008, Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. Eng. Geol. 102, pp. 112–131. https://doi.org/10.1016/J.ENGGEO.2008.03.010

Wold, S., 1995, PLS for multivariate linear modeling. Chemom. methods Mol. Des. pp. 195–218.

Yesilnacar, E. and Topal, T., 2005, Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Eng. Geol. 79, pp. 251–266. https://doi.org/10.1016/j.enggeo.2005.02.002