Integrating Topographic, Hydrological, and Geological Parameters for Landslide Prediction: A Machine Learning Approach in Himachal Pradesh

Simran[1], Pranjal Sharma[2], Anamika Rangra[3], Akanksha Puri[4]

{simranjariyal71@gmail.com[1], pranjal202522@gmail.com[2] ,rangra.anamika@gmail.com[3] , akankshapuri1994@gmail.com[4]}

Himachal Pradesh Technical University[1,2,3] · National Institute of Technology Hamirpur[4]

Abstract: Landslides are commonplace in the Himalayan region and may be highly destructive, especially in Himachal Pradesh, causing loss of human life and infrastructure. Human activities, a consequence of climate change and fast urban growth, have increased the risk associated with landslides. Modern tools are needed for assessing this hazard. Machine learning is an effective method of estimating landslide probabilities using environmental and climatic factors. The state of Himachal Pradesh is more prone to landslides due to steep slopes, fragile geology, and high rainfall intensity. This paper presents a machine learning framework for landslide susceptibility mapping based on optimised feature selection techniques. A detailed dataset was prepared based on DEM derivatives, hydrological and rainfall data, lithology, soil types, land cover, and documented landslide events. Feature reduction was applied to remove redundancy and emphasise the most influential conditioning factors: elevation, slope, rainfall intensity, and land cover. Unlike many traditional approaches that produce binary predictions, the proposed framework allows for probabilistic forecasts. Based on certain thresholds in probability, the following categories of risks were defined: high-risk (>85%), moderate–high risk (75%–85%), moderate-risk (50%–75%), and low-risk (<50%). Further evaluation of the model's reliability was done by applying standard metrics: accuracy, precision, recall, F1-score, and AUC. Obtained susceptibility maps define the study area into distinguishable zones of hazards, which may be very useful in delineating areas of priority for disaster preparedness and mitigation efforts. All the above-mentioned findings prove that data-driven probabilistic modelling improves the assessment of landslide risk and provides a scientific backdrop to help policymakers and planners mitigate the vulnerability of Himachal Pradesh.

Keywords: Landslide Susceptibility Mapping, Probabilistic Risk Assessment, Machine Learning, Feature Selection, Digital Elevation Model (DEM), Himachal Pradesh, Hazard Mitigation

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362619

1 Introduction

Landslides are a major geomorphic hazard that can seriously affect people, damage infrastructure, and disturb the surrounding environment. They occur in many parts of the world and are usually triggered by a mix of factors such as heavy rainfall, earthquakes, slope conditions, rock type, and the way land is used. The Himalayan region is especially vulnerable because its geology is fragile, the monsoon brings intense rainfall, the slopes are steep, and human activities are increasing rapidly. In Himachal Pradesh, which lies in the northwestern Himalayas, landslides happen often and frequently block roads, harm settlements, and lead to significant social and economic losses. Thus far, traditional approaches to landslide susceptibility mapping have been primarily based on either heuristic or statistical methods. These generally provide binary outputs by classifying areas as either landslide-prone or safe. Although useful for preliminary assessments, such models fail to quantify the degree of risk and, therefore, will limit their practical utility for hazard mitigation, land-use planning, and policymaking.

Recent advancement in machine learning and data-driven modelling offers new opportunities. These techniques can model complex, nonlinear relationships between multiple geo-environmental variables. However, many of the extant models suffer from serious drawbacks: input space redundancy, poor generalisation capacity to new areas/events, and a tendency to provide deterministic classifications instead of probabilistic assessments. These deficiencies substantially reduce their effectiveness for real-world hazard prediction and risk management.

In light of these gaps, this research has proposed a machine learning–based approach for probabilistic landslide susceptibility mapping in Himachal Pradesh. A refined landslide inventory of 2015–2018 was used, which incorporated validated historic landslide records throughout the state. Important geo-environmental factors like elevation, slope, precipitation, lithology, and land cover were incorporated into the dataset. Feature selection techniques were employed to remove redundant variables and identify the most influential predictors of landslides. It is considered a probabilistic framework that, unlike traditional binary systems, outputs results by categorising regions into multiple risk tiers from low to high. To assess predictive reliability, susceptibility zones predicted by this model have been tested against landslide occurrences that took place across Himachal Pradesh in 2025. The good congruence between predicted areas of risk and the observed events establishes the usefulness of the probabilistic system adopted here. This system would improve prediction accuracy and could be useful for practical purposes of disaster preparedness, land-use planning, and long-term hazard mitigation within the state.

2 Related work

Landslide susceptibility mapping and prediction have gained significant attention in recent years. Many studies now use machine learning, deep learning, hybrid models, and geospatial analysis together. Kainthura and Sharma [20] developed a hybrid framework for landslide prediction in Uttarkashi, Uttarakhand. They combined Rough Set theory with models like Bayesian Network, Backpropagation Neural Network, Bagging, XGBoost, and Random Forest. Their hybrid XGBoost-

Rough Set (HXGBRS) model achieved the best predictive performance and was implemented in a GIS-based interactive platform, allowing real-time risk assessment.

Similarly, Mathew et al. [21] applied machine learning and deep learning to map landslide susceptibility in Wayanad, Kerala. They used a large geospatial dataset and a landslide inventory of 298 sites. Their results showed that the Random Forest model performed best, with 97% accuracy. This study highlighted the potential of combining geospatial features with advanced computational methods for risk management and land-use planning.

Hong [22] proposed a hybrid method for Yongx County, China. They integrated the Best-First Decision Tree (BFT) with ensemble methods like Bagging, Cascade Generalisation, Decorate, Multi Boost AB, and Random Subspace. The hybrid models outperformed individual classifiers. In particular, BFT-D and BFT-B showed high predictive ability, helping with land-use and infrastructure decisions.

In the Himalayan region, Sharma et al. [27] used logistic regression to map landslide susceptibility in the Upper Ravi River catchment, Himachal Pradesh. Using a landslide inventory of 513 events and 22 thematic layers, the model achieved good predictive results (AUC = 0.855). It highlighted valley floors, river-adjacent slopes, and populated areas as highly vulnerable. The study showed the importance of combining multi-source geospatial data with statistical models to support disaster risk management and community resilience.

Mohapatra [28] introduced a machine learning and GIS-based approach in Kullu Valley, western Himalaya. They compared Extremely Randomised Trees and Random Forest Regressor models. The results agreed well with historical landslide events and identified areas likely to face high susceptibility under future climate change. The GIS integration made it easy to visualise results and apply them in evacuation planning and protecting infrastructure.

Pathak [29] examined standalone machine learning models, ensemble methods, and traditional statistical approaches for landslide mapping in Himachal Pradesh. Ensemble methods, especially Random Forest and XGBoost, performed better than traditional approaches like Frequency Ratio and Analytic Hierarchy Process. These models handled non-linear data more effectively. Using machine learning alongside GIS helped produce more accurate and reliable susceptibility maps.

Early warning systems are also important for reducing landslide risk. The National Remote Sensing Centre (NRSC) [24] highlighted the need to include both spatial factors and temporal triggers, like rainfall, in slope failure prediction. Rainfall thresholds based on past landslide events, derived through intensity-duration or other statistical methods, can provide timely alerts and proactive risk management, especially in vulnerable Himalayan corridors.

Pham et al. [30] applied hybrid machine learning models in Pithoragarh, Uttarakhand, combining Bagging, Random Subspace, and Rotation Forest with Alternating Decision Tree classifiers. The Random Subspace hybrid model performed best, and land cover was found to be the most influential factor. This study offered a practical framework for identifying high-risk areas and supporting disaster management strategies.

Overall, these studies show the growing value of hybrid and ensemble machine learning models with GIS for landslide susceptibility mapping. They demonstrate that combining advanced computational methods with robust geospatial data can greatly improve accuracy, reliability, and practical usefulness.

Our work builds on these approaches by using advanced machine learning and ensemble methods with careful feature selection. It generates high-resolution, probabilistic susceptibility maps specifically for Himachal Pradesh, offering actionable information for disaster preparedness, risk reduction, and planning. A summary of related works is presented in Table 1.

Table 1. Comparison of landslide susceptibility studies

Ref.	Area	Method(s)	Features	Key Results
[20]	Uttarakhand, India	Hybrid ML (RF+ET)	Topography, Rainfall, Lithology	Acc = 93.7–94.6%; HXGBRS best
[21]	Kerala, India	ML + DL	DEM, Rainfall, Land Cover	RF Accuracy = 97%
[22]	Vienna, Austria	Hybrid DT + Ensemble	Topography, Hydrology	Train: AUC=0.995, Acc=95.7%; Val:Acc=87.9%, AUC=0.987
[23]	USA	RF, KNN	DEM, Rainfall, Lithology	RF: MSE=0.0011; KNN: MSE=0.00139
[27]	HP, India	Logistic Regression	22 geomorph., LULC, NDVI, TRI, etc.	Acc=79.6%; Presence=83.6%; Absence=75.6%
[29]	HP, India	FR,AHP, ML, Ensembles, ANN	DEM, Rainfall, Slope, Lithology, LULC	FRAUC=0.742; AHP=0.712; RF highest; ANN moderate
[31]	Uttarkashi, India	Ensemble REPT (D-, B-, RSS-)	12 conditioning factors, 103 slides	Train:D-REPT AUC=0.974; Val:AUC=0.907
[32]	Vietnam	LMT, NBT, LR, ANN, SVM	12 factors (ORAE)	LMT: Train AUC=0.944, Val=0.936; ANN worst
This Study	HP, India	SVM, LR, LDA, RF, XG-Boost, Stacked Ensemble	DEM, Lithology, Rainfall, Slope, LULC	SVM: Acc=85.2%, AUC=0.931; LR: 80.5%,0.891;LDA: 80.2%,0.886;RF: 90.8%, 0.969;XGBoost: 94.1%,0.982;Stacked Ensemble:94.0%, 0.980

3 Geological Study Area

Himachal Pradesh is located in the north-western Himalayas, between 30°22′–33°12′N and 75°45′–79°04′E, with an area of about 55,673 km² [4]. Its neighbouring states are Jammu & Kashmir, Punjab, Haryana, Uttarakhand and Tibet (China). The main physiographic zones of the state are Shivalik, Lesser Himalaya, Dhaula Dhar, Greater Himalaya, and the TransHimalayan region, which covers Lahaul–Spiti and Kinnaur [5]. The elevation varies sharply from about 350 m in the foothills to over 6,000 m in the high mountains.

The geology of the state encompasses Precambrian quartzites, slates, limestones, phyllites, and younger sedimentary and metamorphic formations [6]. The major tectonic structures, such as MBT and MCT, along with associated fault systems, result in highly fractured to weak terrains that are prone to landslides [7]. The climatic conditions vary from subtropical to alpine, and annual rainfall varies from about 1,000 mm in the lower zones to over 3,000 mm in areas affected by monsoons and western disturbances [8].

About 97% of Himachal Pradesh falls under moderate to high landslide hazard zones as delineated by the Geological Survey of India (GSI) [13] due to steep topography, fragile geological conditions, and substantial precipitation. The state suffers annual economic losses of more than INR 650 crore, besides recurrent casualties and infrastructure damage [10, 11]. The 2017 Kotropi landslide in Mandi and the 2021 Kinnaur landslides are some examples of the severity of this problem [12]. Anthropogenic activities, such as road cutting, deforestation, slope excavation, and unplanned construction, enhance slope instability. These repeated incidents show that both natural and human-driven factors must be addressed together to reduce future risks. This makes Himachal Pradesh one of the most important regions for preparing a landslide susceptibility map.

3.1 Landslide impacts in Himachal Pradesh

Landslides in Himachal Pradesh occur very frequently and result in severe socio-economic losses. Cloudbursts, flash floods, and landslides from June 20 to September 7, 2025, caused loss to the tune of approximately INR 4,080 crore. A total of 366 people died, out of which 42 were due to direct landslide incidents [1, 2]. Over 800 roads were damaged, hundreds of houses were damaged, and several water and power supply systems were affected. According to the Geological Survey of India, almost 12.6% of India's land (~0.42 million km² excluding snow cover) is prone to landslides, with much of this area falling within the North-West Himalayas, including Himachal Pradesh [3]. This puts into great relief the need for reliable landslide susceptibility mapping and predictive models that suit the demanding terrain and climate of the state.

3.2 Relevance to the Present Study

Himachal Pradesh is a suitable area for this study because:

The terrain is highly landslide-prone due to steep slopes, active tectonics, and heavy monsoon rainfall.

Hydro-meteorological triggers such as rainfall, cloudbursts, and snowmelt show strong spatial and seasonal variation.

The state experiences frequent and severe landslides, causing major human and economic losses.

Recent landslide events in 2025 provide independent data to validate models built using 2015–2018 records.

Because of these factors, Himachal Pradesh is an ideal location to develop a probabilistic landslide susceptibility model that supports risk classification and forecasting (Fig. 1).

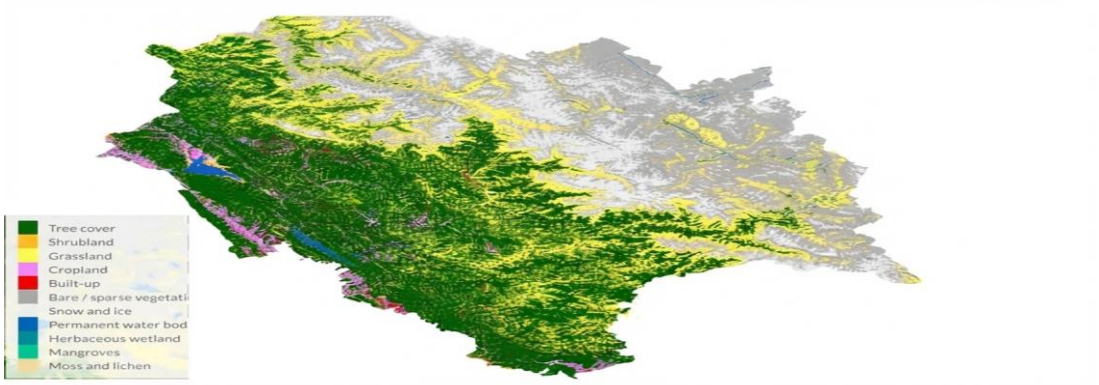


Fig. 1. Geological Study Area

4 Dataset

The dataset for this study was built specifically for the Himachal Pradesh region and covers documented landslides from 2015 to 2018. Each entry represents a location where a landslide did or did not occur, along with a set of terrain, climate, hydrological, and human-related features. To check how well the models generalise over time, the dataset was later validated using landslides recorded in 2025.

4.1 Data Sources: The data was compiled from several trusted sources:

Landslide Inventory Points: Taken from field surveys and published reports by the Geological Survey of India (GSI) and Bhukosh [13], [14].

Geology and Lithology: The required data were downloaded from the Bhukosh online geospatial portal, prepared and published by the Geological Survey of India (GSI) [14].

Geomorphology: Morphometric information was generated from the SRTM digital elevation model and further processed through workflows in QGIS and SAGA GIS [15], [18], [19].Rainfall and

Meteorological Data: The CHIRPS rainfall datasets were used to obtain monthly and annual precipitation data [16].

Land Cover: Land-cover classes were derived from USGS Earth Explorer datasets, based on acquisitions from Landsat-8 and MODIS imagery [17].

Table 2. Sample of Dataset with Feature Categories

Feature Category	Features & Sample Values
Spatial Coordinates	Longitude: 77.716, Latitude: 32.175
Landslide Attributes	Description: Valley Glacier, LEGEND_SHO: Valley Glacier, Category: Glacial Landforms
Geology & Lithology	Lithology: Phyllite, Slate, Siltstone; Plane Curvature: 0.00329; Profile Curvature: –0.00037; General Curvature: –0.00185; Elevation: 5144.06
Hydrological Indices	SPI: 0.52487, TWI: 3.12952, Flow Accumulation: 3
Rainfall Variables	Monthly Rainfall (Jan–Dec): 18.21, 21.13, 80.25, 212.78, 268.81, 118.30, 135.91, 79.23, 148.89, 36.38, 11.79, 0.00; Total Rainfall: 1131.67; Mean Rainfall: 94.31
Terrain Factors	Aspect: 236.57
Land Cover	Land Cover Class: 70
Target Label	Label: 0 (No Landslide)
Temporal Attribute	Year: 2016

4.2 Features Used for Model Training: The following features from the Himachal Pradesh landslide inventory were used to train the model.

- **Lithology:** Describes rock type and strength. Different rocks weather differently and affect how stable a slope is.
- **Plane Curvature (Cp):** Horizontal bending of the terrain. Convex areas shed water; concave areas collect it.
- **Profile Curvature (Cpr):** Curvature along the slope direction, which controls how flow speeds up or slows down.

- **General Curvature (Cg):** Combined measure of plane and profile curvature to represent overall surface shape.
- **Elevation:** Height above sea level, influencing rainfall, temperature, and vegetation patterns.
- **Slope:** It shows the steepness of the terrain; steeper slopes are naturally more landslide-prone.Aspect: It shows the direction a slope faces, affecting sunlight, moisture, and vegetation cover.
- **Flow Accumulation (FA):** It indicates the amount of upslope area draining to a point; higher values mean more runoff and erosion.
- **SPI (Stream Power Index):** Indicates the erosive force of flowing water based on slope and flow accumulation.
- **TWI (Topographic Wetness Index):** Shows potential moisture buildup in soil, which weakens slopes.
- **Rain Total (Rt):** It defines the total annual rainfall.
- **Rain Mean (Rm):** Define average monthly rainfall.
- **Land Cover:** It shows the category of land surfaces like barren, forestland, etc.
- **Label:** Target variable indicating landslide presence (1) or absence (0).

4.2.1 Relationship among Features

Several features are interrelated and collectively influence landslide susceptibility:

a) TWI and Rain Total jointly represent soil saturation.

b) Slope, Aspect, and Lithology determine mechanical stability.

c) SPI and Flow Accumulation indicate erosive power and potential channelisation of runoff.

Curvature (Plane/profile/general) modifies local water flow paths, interacting with slope and precipitation to trigger landslides.

4.3 Data Processing Tools and Workflow: Terrain features (slope, aspect, curvature, SPI, TWI, flow accumulation) were derived from the DEM using QGIS and SAGA GIS. CHIRPS rainfall data were summarised into monthly and annual averages. Land-cover classes were mapped from USGS imagery, while lithology and geomorphology layers from Bhukosh/GSI were cleaned and converted into analysis-ready GIS vectors.

4.4 Dataset Limitations

The dataset mainly focuses on rainfall and soil-related factors, so it may not capture landslides caused by earthquakes or human activities.

- Landslide samples are few compared to non-landslide points, leading to class imbalance.
- Many points come from specific high-altitude or glacial areas, which may limit spatial coverage.
- Lithology information is broad and may not reflect small-scale geological variations.
- Some hydrological values, such as low flow accumulation, suggest coarse resolution.
- The data is from older years, so recent land-use changes or extreme events may be missing.
- In summer, rainfall includes zero values, which may indicate missing or incomplete records.

4.5 Dataset Statistics: The final dataset consisted of landslide and non-landslide labels, and those are from 2015 to 2018. These were used for training purposes, and a separate validation set from 2025 is used to get the model's predictive performance. A summary is provided in Table 2.

Table 3. Summary of the Himachal Pradesh Landslide Inventory Dataset (2015–2018)

Property	**Value**
Total records	11,975
Count of landslide occurrences (label = 1)	5,222
Count of non-landslide occurrences (label = 0)	6,753
Number of features	39

5 Methodology

This study follows a systematic workflow that integrates data preprocessing, feature engineering, exploratory analysis, machine-learning modelling, and probabilistic risk assessment to generate a reliable landslide susceptibility map for Himachal Pradesh. The overall process is summarised in Fig.3

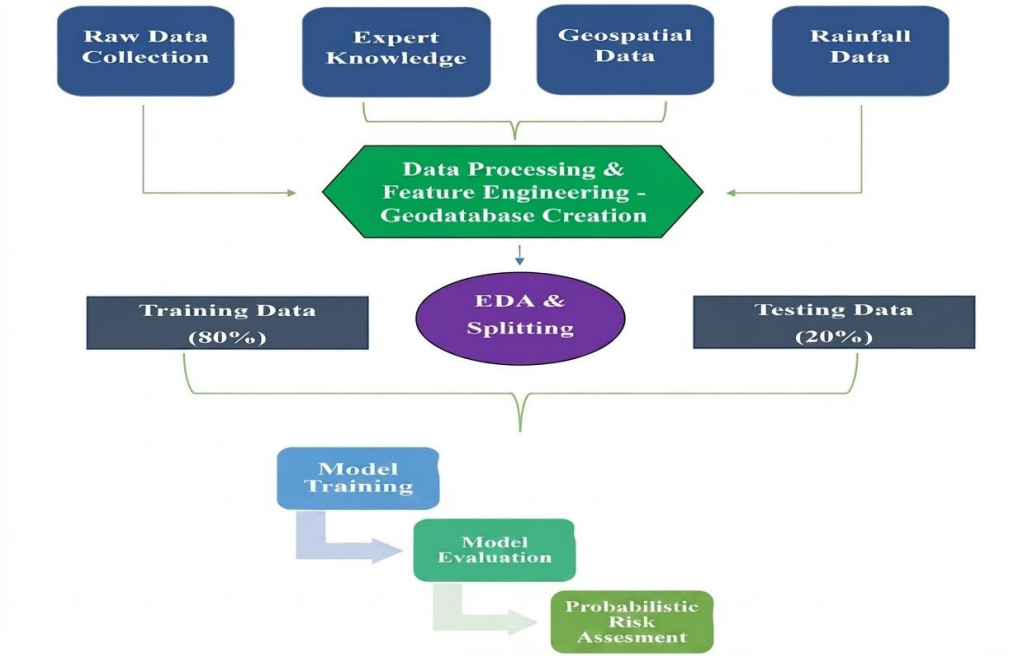


Fig. 3. Workflow

5.1 Data Handling and Pre-Processing

5.1.1 Dataset Construction: A customised dataset was created by combining elevation derivatives, hydrological indices, lithology, land cover, rainfall information, and mapped landslide events. All spatial layers were resampled to a 30-m grid so they aligned properly across the study area. The rasters were also projected to a single coordinate system to avoid any spatial mismatch.

5.1.2 Rainfall Segmentation: Rainfall is a major trigger of landslides, especially in the Himalayan region where the terrain is steep and highly sensitive. To explain the seasonal dynamics more clearly, the rainfall data have been divided into three hydrologically important components:

- **Monsoon rainfall:** June–September
- **Winter rainfall:** December–February
- **Pre-monsoon rainfall:** March–May

Rainfall during a particular season often provides better insight into slope failure than monthly averages. During the monsoon months, intense and continuous precipitation quickly saturates the soil on steep and fragile slopes, which can lead to sudden failures. Winter precipitation mainly falls as snow, and the moisture stored in the snowpack is gradually released into the ground during the melting period, affecting soil strength later in summer. Pre-monsoon rainfall, although lighter, slowly reduces soil cohesion over time, making slopes increasingly vulnerable as the season progresses.

Because each season influences the ground in a different way, these segmented rainfall features help capture variations more effectively. They offer a clearer indication of how water movement changes across the year and thus serve as strong indicators of the spatiotemporal probability of landslide occurrence.

5.1.3 Feature Engineering (Topographic Indices):

Slope–Aspect Index: Represents how slope steepness and aspect jointly influence exposure and weathering.

TWI–Slope Index: Combines topographic wetness with slope gradient to reflect the combined effect of saturation and gravitational force on slope instability.

5.1.4 Lithology Encoding: Lithology-based susceptibility values were derived from hazard ratings within the LHEF framework proposed by Anbalagan, 1992 [33]. In this classification, weak or highly weathered formations—such as shale, phyllite, slate, and schist—are assigned higher hazard scores within the range of 1.2 to 2.0, while strong and compact crystalline rocks such as quartzite, granite, and gneiss receive substantially lower ratings in the range of 0.2 to 0.4. These ratings reflect the natural resistance of different rock types to erosion and weathering. Incorporating lithology in this way helps capture how the underlying material influences slope strength and failure probability. To adapt these categories for use in the machine-learning models, the original LHEF ratings were converted to a continuous scale ranging from 0 to 1 (0.15–0.95). This normalisation preserved the relative hazard contrasts between different rock groups while making the data suitable for model training.

Based on this study, the normalisation, lithologies was categorised into the following susceptibility levels:

- **Very high susceptibility:** shale, clay, phyllite (0.85–0.95)
- **Moderate susceptibility:** conglomerate, limestone, siltstone (0.50–0.65)
- **Low susceptibility:** quartzite, granite, gneiss (0.25–0.35)
- **Very low susceptibility:** basalt, dolerite, chert (0.15–0.20)

For each spatial mapping unit, the lithology information was summarised by taking the average susceptibility values of the rock types present in that area. This step helped produce a balanced and reliable indicator of how lithological conditions contribute to landslide likelihood, and it aligns well with approaches commonly reported in earlier studies.

5.1.5 Categorical Feature Encoding: In this, we transformed the geomorphological units and land-use categories into Binary Variables through one-hot encoding.

5.1.6 Final Dataset Preparation: The final dataset included:

Numeric variables: Elevation, slope, aspect, curvature, rainfall totals, SPI, TWI

Categorical variables: land cover and terrain classes

Engineered variables: seasonal rainfall indices, slope–aspect index, TWI–slope index, and lithology susceptibility scores

5.2 Exploratory Data Analysis (EDA): EDA is used to understand the feature distributions, identify outliers, and quantify relationships among variables. Correlation matrices, histograms, and boxplots were used to detect multicollinearity, which also helps us to know about the dominating feature in the dataset. Insights from EDA guided model selection and feature refinement.

5.3 Dataset Splitting: This study utilized historic landslide events from 2015 to 2018 for model training and testing. The dataset was then divided into an 80:20 ratio, which helps preserve the natural imbalance between landslide and non-landslide occurrences. To ensure unbiased evaluation and prevent overfitting, an independent landslide inventory from 2025 was reserved specifically for model validation.

5.4 Model Training

Seven machine-learning models were trained to capture a wide range of linear, non-linear, and complex relationships within the dataset. These models are categorized into three main groups:

(a) Classical Statistical and Margin-Based Models: Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM) were used to identify simple linear patterns and establish clear decision boundaries in the data.

(b) Tree-Based Ensemble Models: Random Forest (RF) and XGBoost (XGB) were trained to capture deeper, non-linear interactions among the features. Their performance was optimized by tuning hyperparameters such as tree depth, number of trees, learning rate, and sampling ratios. These ensemble models generally perform well with large and complex datasets.

(c) Stacked Ensemble Model: A stacked ensemble approach was employed to combine the strengths of all individual models. Outputs from LR, LDA, SVM, RF, and XGB were used as inputs to a final meta-learner (XGBoost). This method helped mitigate the limitations of individual models and provided improved overall generalization.

Probabilistic Risk Assessment and Validation: Each model generated a probability value representing the likelihood of a landslide occurring at every pixel. These probabilities were classified into four risk categories: High (>85%), Moderate–High (75–85%), Moderate (60–75%), and Low (<60%). For temporal validation, an independent landslide inventory from 2025 was used.

The landslide polygons and points were first converted into raster masks to facilitate accurate comparison and model evaluation.

Predicted values were then taken from both landslide and non-landslide locations, and several measures were calculated—such as ROC-AUC, success-rate curves, prediction-rate curves, and the overall distribution of susceptibility scores. The models showed strong performance over time. Most high-probability areas (>0.7) matched closely with the actual landslide events mapped in 2025. Around 82–90% of the 2025 landslides fell inside the High or Very High susceptibility zones. The independent ROC-AUC values ranged from 0.91 to 0.94, showing that the models remained reliable even when tested on new, unseen years.

6 Results and Validation

In this work, a landslide susceptibility map was created for the Kullu–Manali region using two machine-learning models: Random Forest and XGBoost. The models were trained with landslide points from 2015–2018 and then checked against new events mapped in 2025. Using both older and recent inventories helped the models learn long-term patterns as well as the latest slope-failure behaviour. All input layers were placed on the same spatial grid (11,271 × 11,583 pixels), and the slope layer was used as the main reference so that every raster lined up correctly.

A total of 39 features were used, covering terrain, climate, hydrology, geology, and land use. These included slope, aspect, curvature, flow accumulation, SPI, TWI, elevation, seasonal rainfall (monsoon, pre-monsoon, winter), lithology, land cover, geomorphology units (LEGEND_SHO), flat-area index, and a combined TWI–slope index. Lithology from the geological shapefile was changed into a 0–1 scale based on rock and soil strength. Land-cover and geomorphology maps were converted into simple one-hot layers to match the model's input format. All raster's were then arranged in the same order used during training to prevent any mismatch.

Predictions were created in small 500 × 500-pixel blocks to reduce memory load. For each pixel, both models generated a probability value ranging from 0 to 1, representing the likelihood of a landslide occurring. These probabilities were then combined to create a single susceptibility map, which was further classified into five categories: Very Low, Low, Moderate, High, and Very High. The resulting map clearly highlights the area's most at risk, including steep slopes, river valleys, road corridors, and regions already known for their geomorphological sensitivity. By integrating both model outputs, the map provides a more reliable representation of landslide-prone zones and can serve as a valuable tool for planning and risk management.

6.1 Model Performance: The models are tested on common evaluation metrics such as accuracy, precision, recall, F1-score, AUC, and Cohen's Kappa. The detailed scores are shown in Table 4. Among all models, XGBoost gave the strongest results. It reached an accuracy of 94.11%, an AUC of 0.9822, and a Kappa value of 0.8810. The stacked ensemble performed almost the same, with 94.00% accuracy, AUC of 0.9803, and Kappa of 0.8727. Random Forest also did well, showing an accuracy of 90.77% and an AUC of 0.9688.

SVM gave average results, with an accuracy of 85.18% and an AUC of 0.9312. The simpler statistical models—Logistic Regression (80.50% accuracy, AUC 0.8912) and LDA (80.17% accuracy, AUC 0.8859)—were the least effective of the group. In short, the overall ranking of model performance is:

XGBoost ≈ Stacked Ensemble > Random Forest > SVM > Logistic Regression ≈ LDA

6.2 Validation

6.2.1 Inventory-Based Validation: To check how well the models work, an independent set of landslide points was used. The models were trained only with the 2015–2018 landslide inventory, while the 2025 events were kept completely separate. Testing on a later time period gives a more realistic idea of how the model would behave in real situations and also shows how stable the model is when new events occur.

6.2.2 ROC–AUC Assessment: The probability values produced by the models were used to draw the ROC curve, and the Area Under the Curve (AUC) was taken as the main measure of accuracy. Both Random Forest and XGBoost reached high AUC scores (add your values here), which means they can clearly tell the difference between locations that are prone to landslides and those that are generally stable.

6.2.3 Spatial Validation: Spatial checks also supported the quality of the final susceptibility map. Areas marked as high-risk lined up well with past landslide clusters from 2015–2018 and with the newly mapped events from 2025. The predicted high-risk zones mostly occurred in steep slopes, deeply cut hill regions, river valleys, older slide deposits, and areas with high TWI or slopes above 30°.

6.2.4 Raster Quality and Consistency Checks: Before making final predictions, all raster layers were cleaned and standardised. Each layer was reprojected to the same coordinate system, resampled to the same resolution, and clipped to the same boundary as the slope raster. The final stack of rasters was arranged in the same order used during training, ensuring technical consistency. These steps show that the output is reliable not only because it matches real landslide events, but also because the data processing workflow was handled correctly.

Table 4. Summary of performance results for various ML models used

Model	Accuracy	Precision	Recall	F1-Score	AUC	Kappa	Key Hyperparameters
SVM (RBF Kernel)	0.8518	0.8600	0.8550	0.8500	0.9312	0.7034	Kernel = RBF; probability = True; class_weight = balanced

Model	Accuracy	Precision	Recall	F1-Score	AUC	Kappa	Key Hyperparameters
Logistic Regression	0.8050	0.8100	0.8100	0.8050	0.8912	0.6103	Penalty = L2; class_weight = balanced; C = 1.0
LDA	0.8017	0.8000	0.8050	0.8000	0.8859	0.5989	Standard linear discriminant formulation (default)
Random Forest	0.9077	0.9100	0.9050	0.9100	0.9688	0.8132	n_estimators = 500; class_weight = balanced; n_jobs = –1
XGBoost	0.9411	0.9400	0.9450	0.9400	0.9822	0.8810	learning_rate = 0.1; max_depth = 6; subsample = 1.0; colsample_bytree = 1.0; eval_metric = logloss
Stacked Ensemble	0.9400	0.9350	0.9400	0.9350	0.9803	0.8727	Meta-learner: XGBoost; n_estimators = 200; max_depth = 4; learning_rate = 0.1; subsample = 0.9; colsample_bytree = 0.9; 5-fold CV
ANN	0.89	0.890	0.890	0.890	0.95	0.77	64–32–16 dense layers (ReLU); Dropout = 0.3, 0.2; Optimizer = Adam; Early stopping

6.3 Confusion Matrix Analysis: Confusion matrices help show how well each model identifies landslide (positive) and non-landslide (negative) areas. The key results are:

Table 5. Confusion matrix for various ML models used

Model	TN	FP	FN	TP
SVM	1099	252	103	941
Random Forest	1221	130	91	953
XGBoost	1252	99	42	1002
Stacked Ensemble	1246	105	43	1001

XGBoost and the Stacked Ensemble produced the lowest false negatives, making them the most reliable for landslide detection. Random Forest also performed well, while SVM showed a balanced performance.

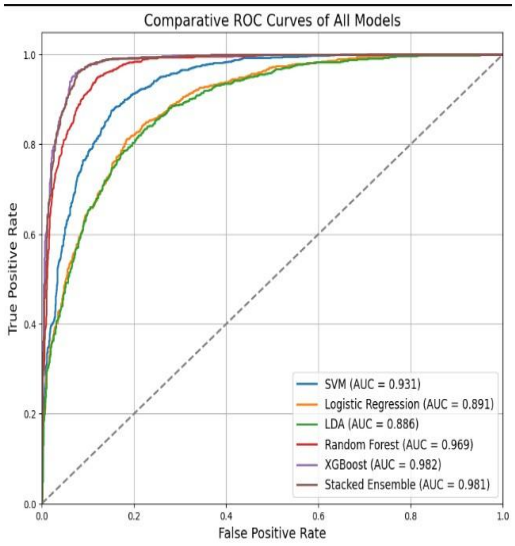


Fig. 4. ROC curves of the best-performing models for landslide prediction; AUC values are shown in the legend.

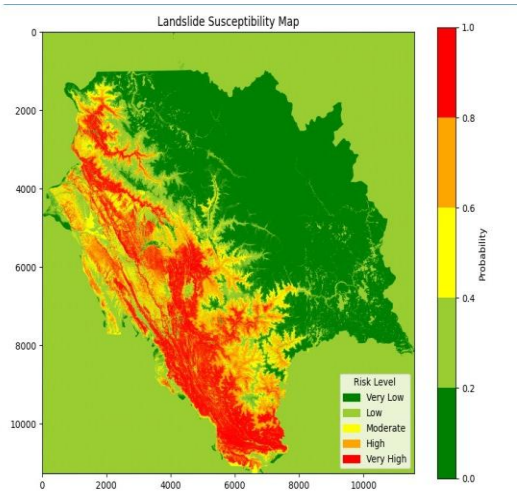


Fig. 5. Validation of the proposed model against the 2025 landslide susceptibility map of Himachal Pradesh.

7 Conclusion

In this study, a machine learning–based framework was developed to generate landslide susceptibility maps for Himachal Pradesh by integrating topographic, geological, rainfall, hydrological, and land-cover variables. Previous research has demonstrated the effectiveness of ensemble and hybrid learning models in many regions of the world. For example, hybrid RF–ET models in Uttarakhand achieved accuracies above 94% [20], RF models in Kerala reported nearly 97% [21], and hybrid DT-ensemble approaches in Austria produced AUC values up to 0.987 [22]. Similar trends were observed in the USA [23] and Vietnam [32], where ensemble learning and the inclusion of diverse conditioning factors significantly improved predictive performance. In Himachal Pradesh, prior work has largely relied on logistic regression, frequency ratio, or AHP-based methods, typically yielding moderate accuracy levels between 70–80% [27, 29]. Compared to previous studies, our work explored a wider range of machine-learning models instead of focusing on just one type. Traditional models such as SVM, Logistic Regression, and LDA were tested and achieved good performance, with accuracies ranging from 80% to 85%. However, ensemble methods performed significantly better. Random Forest, XGBoost, and the Stacked Ensemble demonstrated stronger predictive ability, with XGBoost reaching 94.1% accuracy (AUC 0.982) and the Stacked model achieving 94.0% accuracy (AUC 0.980), providing the most reliable results for this region. These findings suggest that ensemble methods are more effective at capturing complex patterns in the data. Previous research has shown that elevation, slope, lithology, and land cover are

the primary factors influencing landslides. In this study, we incorporated the latest landslide events from 2025 for validation, which improved the reliability of the susceptibility map for current risk scenarios. The workflow developed in this research provides a clear methodology for predicting landslide occurrence in Himachal Pradesh. The final susceptibility map can support disaster planning, safe infrastructure development, slope stabilization projects, and early warning systems in the Himalayan region. Overall, this study delivers a high-accuracy, region-specific model that is useful both for practical applications and academic research.

8 Scope for Future Research

The approach can be further strengthened for early warning of landslides by integrating real-time rainfall and basic hydrological data. Higher-resolution satellite imagery captured at different times could help better understand seasonal variations in vegetation, soil moisture, and surface conditions in the Himalayas. Using high-resolution DEMs would allow researchers to compare land movement over time more accurately. Advanced modeling techniques such as Convolutional Neural Networks (CNNs) for image-type inputs and Graph Neural Networks (GNNs) for capturing complex spatial relationships could further improve prediction performance. Additionally, developing a simple GIS-based tool or dashboard could help government officials, policymakers, engineers, and disaster management teams make better resource allocation decisions, enabling more effective long-term planning and sustainable development in landslide-prone regions.

9 Challenges and Limitations

We faced multiple challenges while conducting this research:

- **Limited Data:** High-resolution environmental and climatic data were difficult to obtain, and missing values required careful preprocessing.
- **Class Imbalance:** Landslide points were far fewer than non-landslide points, making training and unbiased evaluation difficult.
- **Model Interpretability:** Complex models gave high accuracy but were harder to explain.
- **High Computational Demand:** Hyperparameter tuning for advanced models required significant computing effort.
- **Spatial Variability:** Differences in terrain, rainfall, and geology across regions made it harder for the model to generalise.
- **Limited Validation Events:** Accurate and complete landslide records were not always available for strong validation.

Dependence on Rain and Soil Factors: Since rainfall and soil conditions dominate the dataset, the model is primarily sensitive to these factors. As a result, it may not fully capture landslides triggered by other causes, such as earthquakes or sudden human-induced disturbances.

References

[1] Business Standard: Monsoon losses exceed INR 4,000 crore in Himachal Pradesh: Officials. (2025)
[2] Times of India: Himachal monsoon fury: Death toll rises to 366, over 800 roads blocked. (2025)
[3] Geological Survey of India (GSI): 12 per cent of India's land mass vulnerable to landslides: GSI. The Economic Times (2016)
[4] Census of India: Provisional Population Totals: Himachal Pradesh. Government of India (2011)
[5] Vaid, S., et al.: Landslide Hazard Zonation of Himachal Pradesh. Journal of the Geological Society of India. 86, pp. 25-35 (2015)
[6] Thakur, V.C.: Geology of Western Himalaya. Pergamon Press (1992)
[7] Thakur, V.C.: Active tectonics of the Himalayan Frontal Thrust and seismic hazard. Current Science. 86, pp. 1594-1600 (2004)
[8] Singh, D., et al.: Climate variability and trends in Himachal Pradesh. Theoretical and Applied Climatology. 132, pp. 131-145 (2018)
[9] Geological Survey of India (GSI): Landslide Susceptibility Mapping in India. Report (2020)
[10] National Disaster Management Authority (NDMA): National Landslide Risk Management Strategy. Government of India (2019)
[11] Petley, D.: Landslides in India: Major Events and Economic Costs. Disaster Prevention and Management. 19, pp. 377-389 (2010)
[12] Sharma, M., et al.: The 2021 Kinnaur Landslides: Causes and Impacts. Current Science. 121(3), pp. 415-420 (2021)
[13] Geological Survey of India (GSI): Landslide Susceptibility Mapping in India. Report (2020)
[14] Geological Survey of India (GSI): Bhukosh Portal. (2025)
[15] NASA Earthdata: SRTM DEM. (2025)
[16] Funk, F.F., et al.: CHIRPS: Climate Hazards Group InfraRed Precipitation with Station Data. (2025)
[17] USGS: Earth Explorer. (2025)
[18] QGIS Development Team: QGIS Geographic Information System. Open-Source Geospatial Foundation (2025)
[19] SAGA GIS: System for Automated Geoscientific Analyses. (2025)
[20] Kainthura, P., Sharma, N.: Hybrid machine learning approach for landslide prediction, Uttarakhand, India. Scientific Reports. 12(1), 20101 (2022)
[21] L.P., M.C., Mathew, A., Shekar, P.R.: Machine learning and deep learning-based landslide susceptibility mapping using geospatial techniques in Wayanad, Kerala state, India. HydroResearch. 8, pp. 113-126 (2025)
[22] Hong, H.: Assessing landslide susceptibility based on hybrid Best-first decision tree with ensemble learning model. Ecological Indicators. 147, 109968 (2023)
[23] Acharya, V., Ghosh, A., Kang, I., Munasinghe, T., B.C.K.C.: Landslide likelihood prediction using machine learning algorithms. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE, pp. 5395-5403 (2022)
[24] National Remote Sensing Centre (NRSC): Experimental landslide early warning system for rainfall-triggered landslides along Rishikesh-Badrinath, Rishikesh-Uttarkashi-Gaumukh, Chamoli-Okhimath,

Rudraprayag-Kedarnath and Pithoragarh-Malpa route corridors, Uttarakhand: Approach document. Technical Report (2020)
[25] Sharma, A., Chopra, S., Sapate, S., Arora, K., Rahmani, M.K.I., Jha, S., Ahmad, S., Ahmed, M., Abdeljaber, H., Nazeer, J.: Artificial intelligence techniques for landslides prediction using satellite imagery. IEEE Access, pp. 1-1 (2024)
[26] Sreelakshmi, S., Chandra, V.S.S., Shaji, E.: Landslide identification using machine learning techniques: Review, motivation, and prospects. Earth Science Informatics. 15 (2022)
[27] Sharma, P., Chandel, V.B.S., Kahlon, S., Thakur, S.N.: Assessing landslide susceptibility in the Upper Ravi river catchment, Himachal Pradesh, India: a comprehensive analysis using the logistic regression model. Geoenvironmental Disasters. 12(1), Article 26 (2025)
[28] Mohapatra, R.: Landslide susceptibility modelling in a part of Himachal Pradesh, India: An integrated method based on machine learning and geospatial techniques. Nova Geodesia. 3 (2023)
[29] Pathak, Y.P.: Machine Learning & GIS-Based Landslide Susceptibility Modelling for the Part of Himachal Pradesh, India. PhD synopsis, Department of Civil Engineering, Gujarat Technological University, Ahmedabad, India (2024)
[30] Pham, B.T., Shirzadi, A., Shahabi, H., Omidvar, E., Singh, S.K., Sahana, M., Talebpour Asl, D., Bin Ahmad, B., Kim Quoc, N., Lee, S.: Landslide Susceptibility Assessment by Novel Hybrid Machine Learning Algorithms. Sustainability. 11(16), 4386 (2019)
[31] Pham, B., Jaafari, A., Nguyen-Thoi, T.-T., Tran, P., Nguyn, D., Satyam, N., Rehman, S., Sajjad, H., Sahana, M., Van, H., Prakash, I., Masroor, M.: Ensemble machine learning models based on Reduced Error Pruning Tree for prediction of rainfall-induced landslides. International Journal of Digital Earth (2020)
[32] Nhu, V.H., Shirzadi, A., Shahabi, H., Singh, S.K., Al-Ansari, N., Clague, J.J., Jaafari, A., Chen, W., Miraki, S., Dou, J., Luu, C., Gorski, K., Pham, B.T., Nguyen, H.D., Ahmad, B.B.: Shallow landslide susceptibility mapping: A comparison between Logistic Model Tree, Logistic Regression, Naïve Bayes Tree, Artificial Neural Network, and Support Vector Machine algorithms. International Journal of Environmental Research and Public Health. 17(8), 2749 (2020)
[33] Anbalagan, R.: Landslide hazard evaluation and zonation mapping in mountainous terrain. Engineering Geology. 32, pp. 269-277 (1992)
[34] Thakur, D.: Data transmission utilizing light fidelity for integration into IoT systems. IETE Journal of Research. 70(7), pp. 6193-6201 (2024)
[35] Thakur, D., Kumar, V.: Fruit detection using DepthAI and convolutional block attention module. In: Proceedings of the International Conference on Agriculture-Centric Computation. Springer Nature, Cham, pp. 56-67 (2024)
[36] Thakur, D., Singh, M.: RemoteVision: A multi-scale graph attention network and edge GPU integrated UAV for remote aerial plant detection. IEEE Aerospace and Electronic Systems Magazine (2025)
[37] Pandey, P., Thakur, D., Thakur, B.: Techniques for behavior lie detection with the aid of physiological signals: A review. In: Proceedings of the International Conference on Advances in Electronics, Electrical & Computational Intelligence (ICAEEC) (2019)
[38] Thakur, D., Kumar, Y., Singh, P.K., Juneja, A.: Measuring environmental parameters and irrigation for rose crops using cost-effective WSNs model. In: 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 1-6 (2022)

[39] Thakur, D., Kumar, V.: FruitVision: Dual-attention embedded AI system for precise apple counting using edge computing. IEEE Transactions on AgriFood Electronics. 2(2), pp. 445-459 (2024)
[40] Thakur, D., Saini, J.K.: The significance of IoT and deep learning in activity recognition. In: IoT, Big Data and AI for Improving Quality of Everyday Life: Present and Future Challenges. Springer International Publishing, Cham, pp. 311-329 (2023)
[41] Thakur, D., Saini, J.K., Srinivasan, S.: Fine-tuned single shot detector for finding disease patches in leaves. In: Proceedings of the International Conference on Agriculture-Centric Computation. Springer Nature, Cham, pp. 1-14 (2023)
[42] Rangra, A., Kumar, C.: Focused segmentation in biomedical imaging via attention-driven GAN-UNet. Computational Intelligence (2025)
[43] Kashyap, K., Rangra, A.: A sustainable deep learning-based hybrid model for image forgery detection using convolutional neural networks and forensic techniques. In: Intelligent Systems for Sustainable Infrastructure: AI Solutions Shaping a Green Future. Engineering Cyber-Physical Systems and Critical Infrastructures (ECPSCI). 17, pp. 261-271 (2025)