scientific reports

OPEN

Hybrid machine learning approach for landslide prediction, Uttarakhand, India

Poonam Kainthura[1,2]✉ & Neelam Sharma[2]

Natural disasters always have a damaging effect on our way of life. Landslides cause serious damage to both human and natural resources around the world. In this paper, the prediction accuracy of five hybrid models for landslide occurrence in the Uttarkashi, Uttarakhand (India) was evaluated and compared. In this approach, the Rough Set theory coupled with five different models namely Bayesian Network (HBNRS), Backpropagation Neural Network (HBPNNRS), Bagging (HBRS), XGBoost (HXGBRS), and Random Forest (HRFRS) were taken into account. The database for the models development was prepared using fifteen conditioning factors that had 373 landslide and 181 non-landslide locations that were then randomly divided into training and testing locations with a ratio of 75%:25%. The appropriateness and predictability of these conditioning factors were assessed using the multi-collinearity test and the least absolute shrinkage and selection operator approach. The accuracy, sensitivity, specificity, precision, and F-Measures, and the area under the curve (AUC)-receiver operating characteristics curve, were used to evaluate and compare the performance of the individual and hybrid created models. The findings indicate that the constructed hybrid model HXGBRS (AUC = 0.937, Precision = 0.946, F1-score = 0.926 and Accuracy = 89.92%) is the most accurate model for predicting landslides when compared to other models (HBPNNRS, HBNRS, HBRS, and HRFRS). Importantly, when the fusion is performed with the rough set method, the prediction capability of each model is improved. Simultaneously, the HXGBRS model proposed shows superior stability and can effectively avoid overfitting. After the core modules were developed, the user-friendly platform was designed as an integrated GIS environment using dynamic maps for effective landslide prediction in large prone areas. Users can predict the probability of landslide occurrence for selected region by changing the values of a conditioning factors. The created approach could be beneficial for predicting the impact of landslides on slopes and tracking landslides along national routes.

Natural disasters are important problems all around the world, and many governments spend a large portion of their annual budget trying to regulate and prevent them[1]. Landslides are widespread in hilly and mountainous places, and they result in significant losses of life and property, having a devastating impact on the socioeconomic situation of a region[2]. Landslides occur when shear pressure in the inclination exceeds shear quality[3]. It has a significant impact on slope modification, particularly in terms of height, steepness, and slope shape[4]. Landslides can be found on all continents and play a significant role in landscape change. They also pose a severe threat to the inhabitants in many locations[5]. Landslides are frequently influenced by both natural and manmade factors[6]. The escalation of human engineering operations in recent years has not only aided in the development of society and the economy but has also contributed to climate instability and the potential increase in the socio-economic value of landslides[7,8]. Therefore, early spatial prediction of landslides is desirable to avoid landslides and, as a result, damage[9]. There were 4862 landslides reported worldwide between 2004 and 2016, resulting in 55,997 deaths[10]. In this aspect, rigorous planning is required to reduce landslide hazard risk, loss, and slope instability mitigation[11]. Developing a landslide mapping model is a fundamental step toward creating catastrophe assessment and mitigation measures in high-risk locations[12–15]. The study of landslide prediction modeling has become a major scientific topic around the world due to its enormous importance[16–18].

Researchers have devised many ways for identifying landslide-prone areas as well as solutions for reducing the negative effects of landslides[19–26]. Landslide susceptibility mapping (LSM) is one of the most effective approaches for predicting landslide-prone zones in certain areas[27]. In general, future landslides in a given location are

[1]School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India. [2]Department of Computer Science, Banasthali Vidyapith, Jaipur, Rajasthan, India. ✉email: poonamkainthura@gmail.com

expected to occur under comparable conditions as in the past[28]. As a result, a spatial relationship between the elements that influence the occurrence of landslides is necessary to identify and predict future landslide locations[29].

The terminology and methods used in the field of landslide prediction modeling have evolved over the years, and they now encompass both qualitative (inventory-based and knowledge-driven)[15,28] and quantitative (data-driven and machine learning) [30–32] techniques. However, the best strategy for determining landslide occurrence is still up for debate[33]. Experts utilize qualitative methods to assess landslide vulnerability zones based on their knowledge[34]. Expert knowledge is derived from a combination of field studies and theoretical understanding of physical processes[34]. Knowledge-based models[35,36], such as the weighted linear combination, rank the relevance of landslide conditioning elements based on expert opinions and experience[37]. Additionally, the qualitative procedures such as the Analytical hierarchy process (AHP)[38] are based on the opinions of one or more experts[39–41]. The complex non-linear link between landslide and landslide producing causes is recovered in a knowledge-driven model by giving weight to elements based on expert knowledge[42]. Expert knowledge is crucial to the success of the knowledge-driven paradigm[15,43]. These methods, on the other hand, are costly and require a high level of geology and geomorphology knowledge[44]. However, it is difficult to objectively assess or evaluate the quality of the outcomes using these methods[28].

On the other hand, quantitative approaches[45] generate numerical estimates, i.e., probabilities of landslide occurrence in any susceptibility zone[46]. Methods such as frequency ratio (FR)[47], logistic regression (LR)[48], statistical index (SI)[49], the weight of evidence (WoE)[50], evidential belief function (EBF)[51], information value (IV)[52], certainty factors (CF)[53], multivariate regression (MR)[54], are based on strong mathematical rules, regardless of individual decision[55]. The strategies listed above have been utilized by researchers all across the world to predict landslides[37,41,56–58]. These data-driven or statistical models are built on simple principles that can outperform simple univariate and multivariate linear tasks[10]. Quantitative techniques for landslide prediction modeling have risen in prominence over the last two decades[15]. Even though various methods for assessing landslide dangers have been developed, no method has been acknowledged as the standard technique for analysis and prediction[59]. Conventional statistical approaches do not perform well for complex and high-dimensional nonlinear issues[60]. As a result, the data nature is likely to change in many circumstances, lowering the model accuracy[61].

Machine Learning approaches are the most effective at handling complicated and nonlinear high-dimensional[62] data sets in landslide research. Tree inductions[63], probabilistic approaches[21,22], Artificial Neural Network (ANN)[23], and Support Vector Machine (SVM)[19] are among the methods used. Machine learning together with artificial intelligence[64,65] algorithms has shown to be an effective and promising tool in many geotechnical applications[66]. In landslide forecasting, machine learning methods are being utilized to improve model accuracy and, in particular, the flexibility of such models to handle a wide range of conditioning factors[7,67]. Landslide mapping has vastly improved in the recent decade, largely in major part to machine learning techniques[68,69]. These models, however, still include some shortcomings, and these shortcomings are impacting the prediction performance of single models[70]. When training data is limited, for example, there is a risk of underfitting, which can lead to erroneous model development when employed alone[71]. Ensemble approaches like Bagging[72], Boosting (AdaBoost[73], Gradient Boosting Machine (GBM)[24], Extreme Gradient Boosting (XGBoost)[74]), and Random Forest[75] were developed in the field of machine learning to help solve this challenge. Ensemble techniques are a type of machine learning methodology that integrates numerous base models to create a single best-fit predictive model. By gaining a better knowledge of the data and rules from multiple models, an ensemble can reduce variance and bias[70]. Ensemble methods aid in the reduction of over-fitting issues in models[76]. It also works well with data with a range of dimensions. Furthermore, the computation performance is unaffected by missing values in the dataset[77]. Additionally, it can also deal with problems in unbalanced data and error reduction[70]. Compared to other ensemble models, the XGBoost technique provides many advantages[78]. Outliers have a negligible effect. There is no need to scale or normalize the data, and it can even manage missing values. The training time is greatly reduced by parallelizing the entire boosting procedure[79].

Model ensembles, on the contrary hand, are not always preferable[80]. Bagging has one limitation: instead of reporting specific values for the classification or regression model, it focuses its final prediction on the mean predictions from the subgroup trees[81]. Overfitting is possible in boosting if parameters are not tuned properly[71]. Overall, it is observed that the capability of the individual model has some limitations[82].

Therefore, hybrid machine learning algorithms have recently generated promising results in landslide prediction modeling[23,83,84]. The base model prediction performance is improved by using a hybrid technique[10]. Moreover, the hybrid models higher performance in predicting landslide revealed that the landslide modeling may be improved by factor optimization[85]. Many studies have demonstrated the value of a hybrid strategy in landslide situations all over the world[10,82,83,86]. Due to regional geological and geomorphological causes, slope unsteadiness, including landslides, is a common problem in Uttarkashi, Uttarakhand[87–89]. According to NASA statistics, 958 landslides killed 6779 persons in India between 2008 and 2015, with Uttarakhand leading the way with 5,226 deaths[90]. Various scholars discovered the reason for landslides in the Uttarkashi region after extensive research[68,91–93]. In the research area, the progression of road construction and changes in land-use patterns have been recognized as the major causes of landslides[94]. Several well-known and well-tested machine learning algorithms, such as SVM, Naïve Bayes, Logistic Regression, Bayesian Network, BPNN, and Random forest were used in this study because they produce great results in the situation of landslide assessment[68,93]. The ability to forecast using various combinations of machine learning technologies is always improving[83]. To select the optimal model, we investigated five hybrid models namely XGBoost-Rough Set (XGBRS), Backpropagation Neural Network-Rough Set (BPNNRS), Bayesian Network-Rough Set (BNRS), Bagging-Rough Set (BRS), and Random Forest-Rough Set (RFRS) in the landslide prediction modeling of the Uttarkashi district, Uttarakhand, India. Additionally, as a result of geological and geomorphological changes induced by climate change, landslide risk may appear in non-landslide areas of study area[95]. Therefore, utilizing GIS platforms such as Openstreet maps[96], the hybrid model with optimal prediction capabilities is used to predict landslide event based on changing

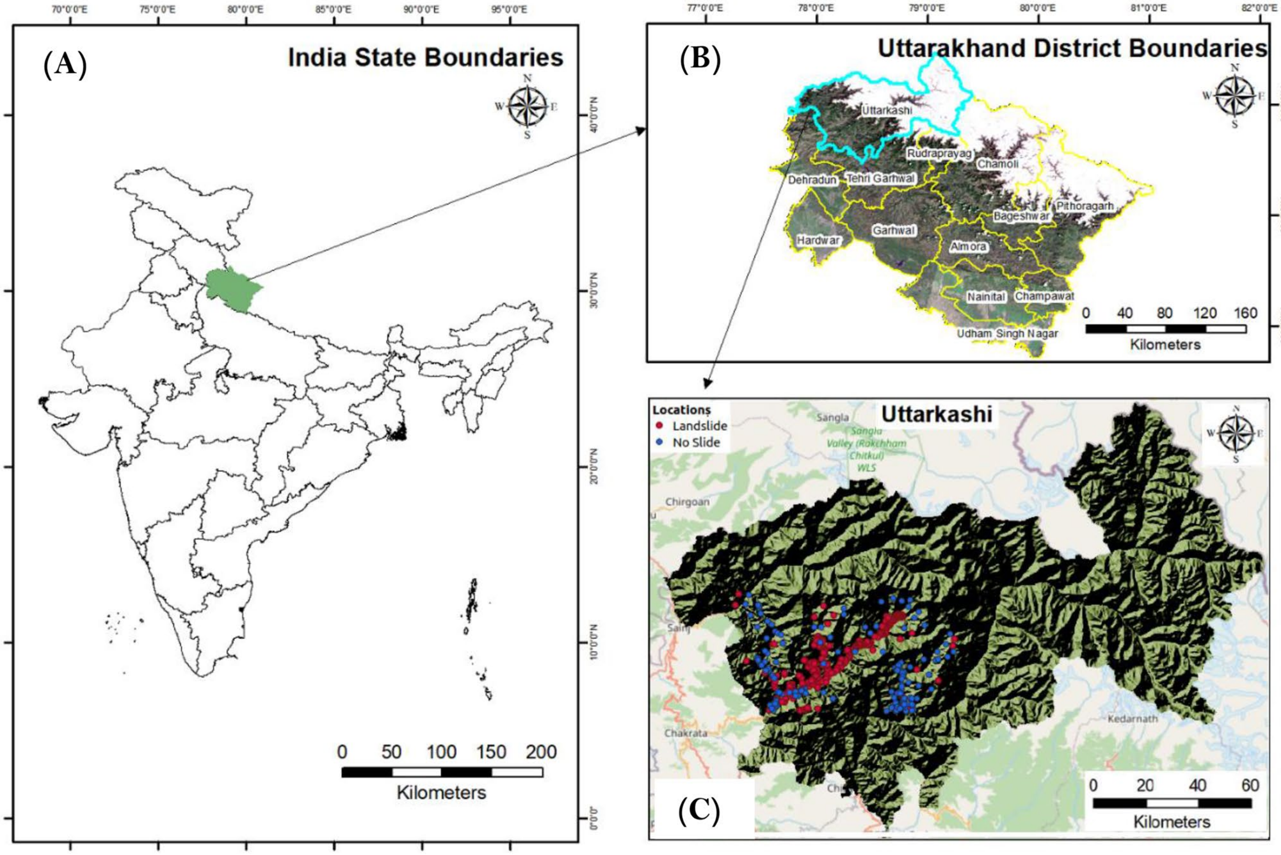


Figure 1. Location map of the study area (**A**) state boundaries of India and study area (Uttarakhand state) (**B**) district boundaries of Uttarakhand state with Uttarkashi district highlighted and (**C**) showing the landslide (red) and non-landslide(blue) locations of Uttarkashi district.

conditions. By evaluating and comparing models using multiple statistical indicators[97], the best prediction model will be found. The primary distinction between this study and earlier ones is that it is the first time that advanced XGBoost ensemble optimization techniques combined with Rough Set are used to investigate the possibility of a more accurate landslide model in the Indian Himalayan region. The main goal of the research is to find an appropriate machine learning algorithm for identifying possible landslide regions in the Uttarkashi region.

Thus, the primary objective of this investigation is to assess the impact of the following on the accuracy of landslide prediction modelling: (1) pre-processed landslide data inventory; (2) landslide-conditioning factors (LASSO and Multicollinearity); (3) prediction techniques; (4) selective validation methods and statistical measures; and (5) visualising the outcome using interactive maps. A case study of the Uttarkashi district (Fig. 1) is established based on the use of 554 landslide/non-landslide locations and 15 conditioning factors to support the proposed methods.

Study area. The study area as shown in Fig. 1, is in the Uttarkashi district between 30.7268° N latitude and 78.4354° E longitude, with an area of approximately 8016 km^2. The map of the study area is generated using ArcGIS v10.0 software. The study area is famous for pilgrimage that is located in the NW direction of Uttarakhand state. During the monsoon season from July to September the study area receives heavy rainfall and from January to March months it receives moderate rainfall[98]. The average rainfall recorded is 1902 mm. The elevation varies from 920 to 3830 m and the average altitude calculated is 1158 m. Failure of slopes in these regions is very frequent during a rainstorm. The state northern portions are characterized by the Greater Himalaya highlands, which are dominated by towering Himalayan peaks and glaciers, with lush vegetation covering the lower foothills. The Indo-Gangetic plain, with a maximum elevation of fewer than 1200 m, is home to the southernmost districts. Uttarakhand is drained by the Ganges river system, with the Ganga, Yamuna, and Kali being the main glacially fed rivers that originate in the province Greater Himalayan areas. Every year, massive landslides occur in the study area as a result of heavy rainfall, steep slopes, and road cutting. Academicians are interested in understanding the landslide activities that occur there and proposing a solution in the field of landslide modeling for those ongoing environmental processes because of the immense diversity and complexity of the environment. Landslides occur regularly throughout the study area, inflicting damage to roads, buildings, and economic infrastructure. Ongoing tectonic activity, extreme rains, and water seepage into the slopes, shear planes, weak lithology, and man-made activities are all contributing to the instability of these Himalayan slopes.

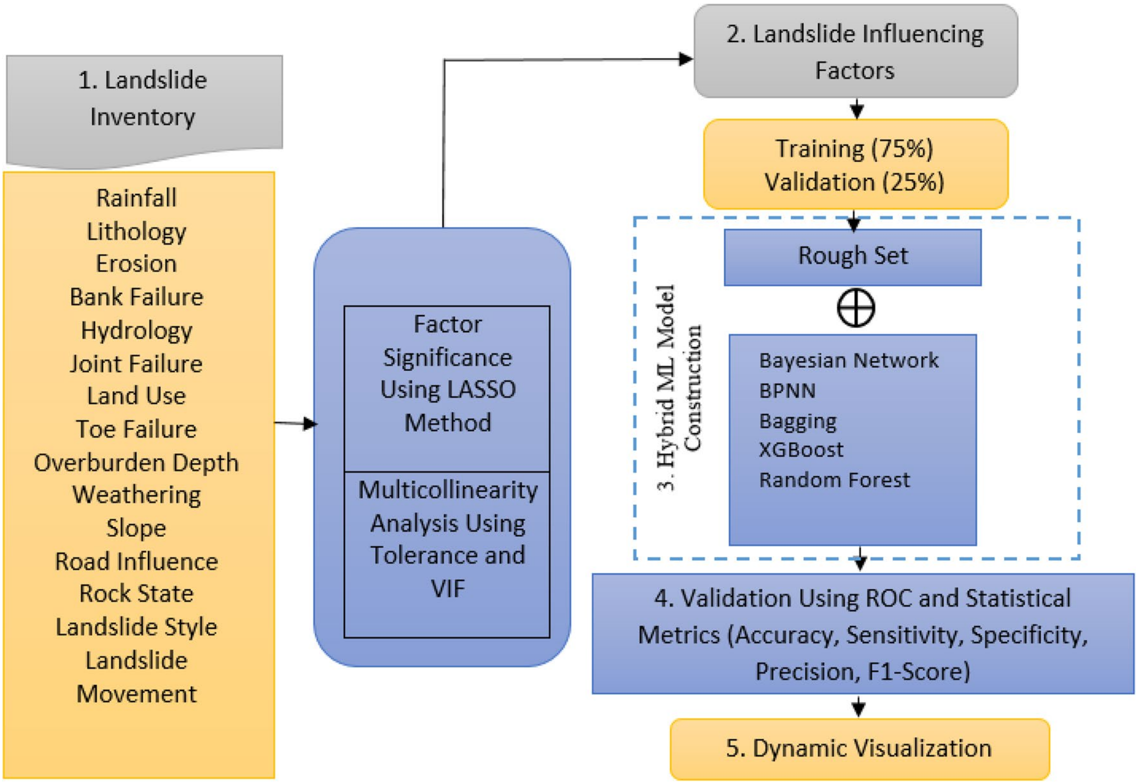


Figure 2. Methodological flowchart showing different steps for the construction of hybrid machine learning models for the prediction of landslide events supported by LASSO method.

The study area was determined by the availability of landslide occurrence causes as well as exact slide locational data. Furthermore, the research field has a substantial challenge that must be addressed using scientific methods.

Methodology

To prepare the landslide prediction model of the study area, the research was divided into five phases: The workflow of the methodology is shown in Fig. 2.

Data collection and pre-processing. The creation of a landslide inventory is required for landslide forecasting. Landslide databases record the locations and conditions of landslides that have migrated in the past, but the mechanism(s) that caused them to move is rarely indicated. This can provide historical and current landslide sites, occurrence periods, landslide types, frequency and severity of occurrence, magnitude and extent, failure mechanisms, causal-related factors, damages, and consequences. This comprises reports based on field studies prepared by the Geological Survey of India (GSI, https://www.gsi.gov.in/webcenter/portal/OCBIS/pageReports/pageGsiReports?_adf.ctrl-state=1gq1usi84_5&_afrLoop=935857055668031#!). Past landslide spots in Uttarkashi District were used as markers (Latitude, Longitude). The coordinates and other details are taken from publicly available landslide records. A total of 554 landslides with precise characteristics were recorded in the study area. Because the dataset contains missing values, a data preparation process was used to handle the data. This phase will aid in the model prediction capability improvement. The quality of the datasets and the models utilized determine the final data quality. Previous research has utilized a variety of ways to clean the data. Statistical approaches including mean and median computations, Multivariate Imputation by Chained Equation (MICE)[99], Imputation Using Deep Learning[100], K-Nearest Neighbor[101], and others are included. In this study, a popular method K-Nearest Neighbor model (KNN) was utilized to the fill missing values to build a high-quality dataset for training. The KNN model, based on the Euclidian distance formula mentioned in Eq. (1), is employed in this work to fill missing values using the KNN Imputer function.

$$Distance(X_i, Y_i) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{1}$$

where X_i and Y_i represent observed and actual data values, respectively, and "n" represents the total number of instances.

As a result, the landslide inventory prepared and listed in Table 1, is valuable for determining the spatial distribution of present landslides as well as the potential for future landslides. The total inventory of landslides was then randomly divided into two data sets with 279 landslides and 136 non landslides for training (75%) and 94 landslides and 45 non landslides for testing (25%). Figure 3 displays the training and testing split of landslide

Parameters	Values			
(Latitude, Longitude)	30° 48′ 36.9" 78° 13′ 10.7"	30° 48′ 35.5", 78° 13′ 8.5"	30° 48′ 40" 78° 13′ 6.2"	30° 48′ 50.9 78° 13′ 18.9"
Rock State	Massive	Sheared	Fractured	Jointed
Hydrology	Dry	Wet	Damp	Dry
Weathering	Low	Moderate	High	low
Overburden Depth(m)	> 5	0–1	2–5	0–1
Erosion	No	No	yes	No
Rainfall	Yes	No	No	Yes
Road Influence	No	No	No	No
Slope	Steep	Moderate	Gentle	Steep
Joint failure	No	No	No	No
River Bank Failure	No	No	No	No
Toe failure	No	No	Yes	Yes
Land Use	Dense	Agriculture	Sparse	Moderately Vegetated
Lithology	Slates	Gneiss	Quartzite	Slates
Landslide Movement	Rotational	Translational	Debris	Debris
Landslide Style	Single	Complex	Multiple	Complex
Landslide Prediction	Yes	No	Yes	No

Table 1. Landslide Inventory sample of the study area (Uttarkashi district) obtained based on field investigation that is publicly available on GSI landslide reports [19–21].

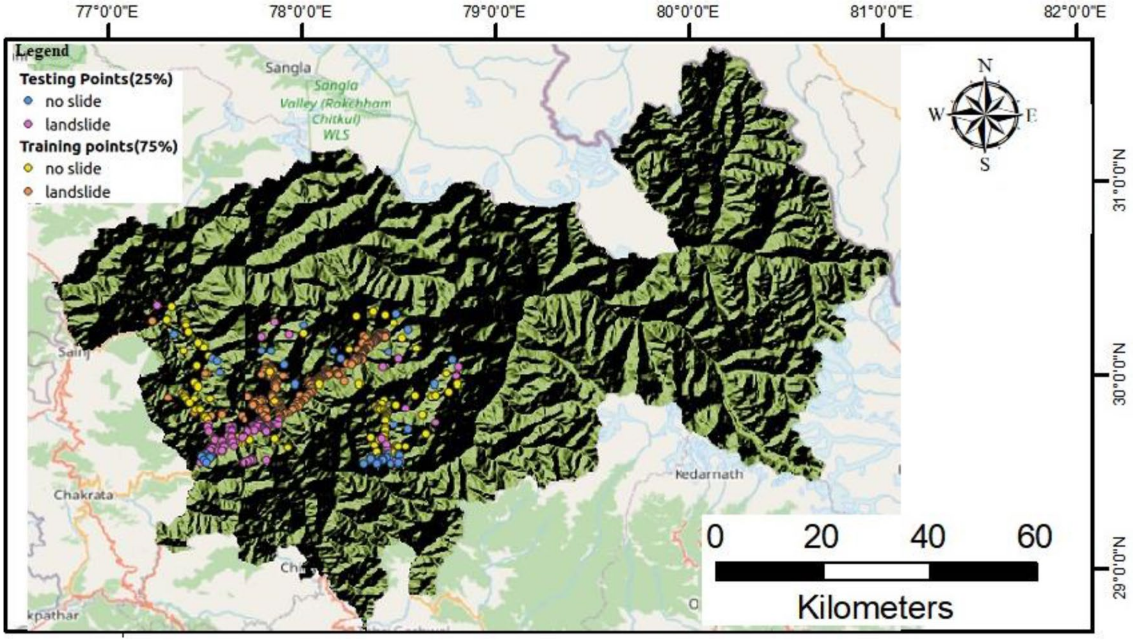


Figure 3. Partition of the landslide and non-landslide locations into training and testing data points with in ratio of 75:25.

locations. Total of fifteen landslide conditioning factors (LCFs) was used in this study for landslide prediction modeling in the Uttarkashi district. The description of each landslide conditioning factor is given below:

Erosion. Erosion is characterized as a physical process with significant global variation in intensity and frequency that is influenced by a variety of social, economic, and political issues in addition to environmental factors. The importance of erosion (channel incision, lateral channel migration, and slope erosion owing to agricultural runoff) as a landslide trigger cannot be underestimated[102]. Massive erosion is a type of erosion that occurs when a large amount of soil or rock mass, or a combination of both, is pushed down a slope by gravity. In fact, this erosion happens when the weight of the material exerts a force greater than the resistance force imposed by soil shear force. Massive erosion in the study area is typically associated with natural erosion, but human activities, such as mine excavation, road construction, and forest vegetation degradation, exacerbate this (deforesting). Massive erosion contribute to landslides, and underground erosion is one of the mechanisms that contributes to soil mass instability. The study area Uttarkashi is prone to landslides due to discussed erosional activities[103].

Lithology. Lithology describes the physical characteristics of a rock unit, such as color, texture, size, and composition. Lithology is one of the most essential parts of landslide study since different lithological units have different geological strength indices, permeability, and susceptibility to failure. As a result, lithology is commonly considered one of the most critical landslide conditioning factors[28].

Bank failure. There are various small and major rivers in the research area. The largest and most revered of these rivers are the Bhagirathi and Yamuna. Cutting the river banks in these locations puts people's lives in jeopardy. Construction, weight on the riverbank, vegetation, tectonic activity, water saturation, and other conditioning factors can all contribute to river cut failure. As a result of these conditioning factors, stress develops along riverbanks, weakening the soil strength and increasing the risk of landslides[104].

Hydrology. Hydrology is the study of the water cycle, water resources, and watershed sustainability on Earth and other worlds, as well as water transportation, distribution, and management. Anthropogenic activities are continually disturbing the Himalayan environment's natural system, with apparent results in the hydrology of streams and springs[105]. The variation of water discharge with the seasons and the increasing amount of sediment load in streams are two of the most important challenges in the study region. The majority of the locations along the stream that runs through the slide zone are dry to damp.

Rainfall. In the studied region, the average annual rainfall is 1902 mm. The most common and widespread destructive landslides are caused by prolonged or intense rainfall. Extreme rainfall events in landslide-prone areas can be disastrous, causing property, infrastructure, and human life to be lost. At several spatiotemporal scales, research on the rainfall characteristics known to produce landslides has been done, and it frequently relies on local knowledge of landslides and rainfall. Rainfall has increased the Himalayan mountains fragility, resulting in a rise in the number of landslides in the region. The majority of rain-caused landslides are shallow, small, and move swiftly. Many rainfall-induced landslides become debris flows as they descend steep slopes, especially those that enter stream systems, where they may mix with extra water and sediment[106].

Joint failure. Landslide distribution is influenced by faults/fractures and other structural lineaments. Local stress fields are assumed to be reflected by different-trend faults/fractures, which have an unequal influence on landslides. As we all know, joints govern the slope and are associated to rockfall. A cross joint positioned NE-SW and perpendicular to the fold axis, a longitudinal, sub-vertical joint oriented WNW-ESE to NW–SE, and a shear joint oriented diagonally from NNW-SSE to ENE-WSW are among the joints in the research area. These joints can sometimes lose control due to the fractured and massive rock, resulting in a landslide[107]. As a result, joint failures are viewed as a key triggering factor in the research.

Land use. Land cover is the physical material that covers the earth's surface. Examples include grass, asphalt, trees, bare ground, water, and other land covers. Land cover/land use has an impact on landslide occurrences because the roots of trees in the forest region play an important role as anchors in retaining the stability of soils and rocks on the slope. Natural vegetation cover on a slope (Landcover), as well as anthropogenic activities such as agriculture, plantation, and large excavation for roads and other civic constructions (Landuse), have a direct impact on hillslope stability. Extensive Cutting is locations along the road and river where extensive toe removal is found in the research region, either as a result of anthropogenic or natural conditions. Wasteland refers to both cultivable lands that have been abandoned and areas where materials (trash) have been deposited. Landslides are associated with barren, massive slope cuttings, and thinly vegetated places. Understanding how different land coverings affect landslides is increasingly required for developing landslide prediction models[107].

Toe failure. In the investigated area, slope toe cutting is sometimes manmade and most of the time natural. When the toe of a slope is removed, the friction to movement is reduced. Frequent rains, on the other hand, eroded the toe of the slope, deforming the hill. Furthermore, building or widening roads without properly understanding the nature of the slope might lead to toe collapse, which can increase landslide occurrences[108].

Overburden depth. Overburden refers to the material (in-situ or transported) that exist on the surface of young rock formation, such as soil and weathered mass[109]. The thickness of the overburden is determined by a variety of characteristics, including terrain shape, slope forming material, weathering rate, slope degree, and so on. Shallow translational debris landslides have been linked to the depth of the overburden to the bedrock. It is also affected by slope and erosion. The overburden depth could be beneficial for recognizing landslide-prone regions and constructing prediction models because the research area is prone to erosion and includes steep slopes. In this study, the overburden depth factor spans from 0–1 m to > 5 m.

Weathering. The weathering and landslides in this mountainous tract in India's northern region have been influenced by the meteorological conditions[110] in the Uttarakhand region. Landslides are widespread in the study area during the monsoon season, and their severity is determined by the thickness of the loose, unconsolidated soil created by weathering. Strong weathering and tectonic activity can be seen near the thrust and fault in the research region. Other conditioning factors speeding up the weathering process and causing landslides in the region include deforestation, land-use practices, and soil erosion. The power of weathering might be low, moderate, or high, depending on the geology and climatic circumstances. When the power is really high, landslides are more likely.

Slope. The steepest direction is always used to calculate the angle of all slopes. When the slope angle of a viscous material increases, the shear strength of the material generally falls[111]. There are several types of slope angles, ranging from 0 to > 35 degrees. The range of slopes is divided into three categories: gentle (0–8.5), moderate (8.5–35), and steep (> 35).

Road influence. The anthropogenic component of road construction (cutting and excavation) has a major impact on natural slope morphology[112]. Road construction activities degrade the mountain's structural stability along the corridor, increasing the risk of landslides. Where roads and highways cross steep terrain, maintaining the stability of road cuts and road terracing is vital. Furthermore, water must be diverted away from slopes so that the road does not block water flow and the groundwater level beneath the road's surface is not too shallow. Natural landslides may occur, causing road instability, road cuttings, and earthworks. Road cuttings and earthworks that are unstable can cause the road to disintegrate, and become dangerous, and inaccessible.

Rock state. The landslide research area has phyllite, quartzite, argillaceous, dolomite, and other rocks. The existence of structural discontinuities in the stones, such as faults, joints, bedding, and thrusts to slope inclination and direction, has a significant impact on slope stability. According to the recorded data, the rock status for the specified study location is classed as massive, sheared, fractured, and jointed. As a result, it could be involved in both the initial rockslide and the secondary rock avalanche. Bonding determines the cohesiveness and internal friction angle of a rock's shear strength. Mechanical fractures in rocks cause displacement discontinuities across surfaces or narrow zones, which are known as fractures[113].

Landslide movement. A landslide is defined as the movement of a rigid body of earth or land along a shear surface. A slide is defined as a series of mass motions that remove the slide mass from more secure earth material in weak zones. Inside the research region, there are a variety of landslide types, including translational, rotational, or combination rock-debris slides, as well as ordinary debris slides. A rotational occurs when the surface is curved and concave to the sky. Moreover, the upper side of the slide is inclined in the backward direction towards the original slope and the lower surface leaves the slope. Whereas translational slide moves in a downward and outward direction inclined on upper planar surface developing dangerous situation by flowing debris on steep ground[6].

Landslide style. Various landslide activity styles have been seen in the Uttarakhand study region. Single, complex, and multiple styles were the most commonly recorded. A single landslide is the flow of displaced material in a single direction. At least two forms of movement (falling, toppling, sliding, spreading, and flowing) occur in a complex landslide. Multiple landslides are characterized by the emergence of the same type of movement multiple times[114].

Selection of landslide-influencing factors.

Landslides are a very complicated geo-environmental event that can be caused by a number of different factors. However, it is also true that not all factors have an equal influence on the initiation of a landslide in any given location. As a result, determining appropriate and accountable factors for the occurrence of a landslide is an important task that demands special consideration in landslide prediction modeling. As a result, careful evaluation of landslide conditioning factor (LCF) influence and removal of the less significant ones from the model is critical. LASSO and Multicollinearity tests are two effective ways to determine the influencing elements for this purpose.

Independent test of causative factors.

The least absolute shrinkage and selection operator (LASSO)[115] is a widely used and recommended technique used to test the capability of predicting factors[116]. When selecting factors, a larger value implies that the component is more important in the occurrence of a landslide. This is a flexible technique for identifying and regularising features. The LASSO improves model interpretability and avoids overfitting by removing extraneous factors that are unrelated to the response factors. Shrinkage is used in the Lasso regression model. The data values are shrunk towards a center point with this technique, which is comparable to the mean approach. Mathematically, the LASSO method can be expressed as:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{2}$$

Which is representing the same as minimizing the sum of squares error that is the first component of formula shown in Eq. (2) with constraint component $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ dependent on the value of λ and β_j. The value of β_j can be shrunk to exactly zero.

Where, λ is the size of the shrinkage. If λ = 0, it signifies that all features are considered, and it is now the same as linear regression. On the other hand, if λ = ∞ it signifies that no feature is used; as the number grows larger, more features are discarded, and feature selection gets more precise. Additionally, the value of rises as the bias increases. The value of falls as the variation grows.

Multicollinearity test of causative factors.

Using several machine learning methods, the study incorporated fifteen landslide conditioning factors to measure landslide prediction. When two or more predictors are associated, the standard error of the coefficients increases. This is known as multicollinearity [117]. In landslide prediction modeling, checking the multicollinearity of multiple conditioning factors is critical. This analysis can help with attribute selection as well as understanding which conditioning factor have the most impact on the target features to predict future landslides. The tolerance and variance inflation factor (VIF) was used to determine the proper assessment of multicollinearity among landslide causative parameters. The tolerance and VIF were calculated using Eq. 3.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2} = \frac{1}{\mathrm{Tolerance}} \tag{3}$$

where R is the coefficient of regression of independent conditioning factor "i ". When the TOL number is > 0.2, there is no evidence of multicollinearity; on the other hand, a TOL value less than 0.1 suggests significant multicollinearity. The presence of strong multicollinearity among the independent conditioning factor is indicated by a value of greater than 10 in VIF.

Methods for landslide prediction modeling.

Bayesian network. A Bayesian Network[70] is a statistical classifier that shows a probabilistic link between several conditioning factors. BN is a powerful technique for modeling a complicated causal network system. BN preserves the joint probability among causal components in the form of an acyclic directed graph (DAG)[118]. When multiple causal elements are coupled, they might have a lot of power. Each node in the graph keeps a conditional probability table to illustrate cause-and-effect interactions (CPT). The BN model can learn cause-and-effect networks quickly and can tolerate partial datasets. The CPT computation for each causal element of a landslide is expressed as Eq. (4).

$$P(X_i|Y) = P(Y|X_i) * \frac{P(X_i)}{P(X)} \tag{4}$$

The combined probability of predicting landslide based on numerous parameters is expressed as Eq. (5)

$$P(X_i) = \prod_{i=1}^{n} P(X_i|predecessor(Y_i)) \tag{5}$$

where X_i the number of cases, Y_i is a set of causative factors, $P(X_i|predecessor(Y_i)$ denotes that X_i is the cause of Y_i, and $P(X_i)$ is the probability of all possible X_i values.

BN is thought to be a potential strategy for predicting landslides[119]. It is, however, only seldom used to predict landslides. Landslide stability is influenced by a variety of elements, which are divided into internal and external influence components. Landform, stratum structure, rock and soil characteristics, and hydrological elements are among the internal components.

Artificial neural network. The artificial neural networks (ANN)[120] is a form of machine learning algorithm that learns by finding valuable patterns from data on its own. Artificial neural networks, like the human brain, have neurons in multiple layers that are coupled to one another. Three layers make up Artificial Neural Networks: (1) Input Layer, (2) Hidden Layers, and (3) Output Layer[30]. In prediction and regression applications, the backpropagation neural network (BPNN)[121] learning technique is the most often employed ANN model. Because of the nature of the backpropagation learning mechanism, backpropagation neural networks can be utilized to tackle issues in a wide range of domains. This method is employed in any sector where neural networks are used to solve problems involving a set of inputs and a set of output targets[121]. The backpropagation algorithm's training is divided into three parts. The training input is fed forward in the first stage. The calculation and backpropagation of linked mistakes is the second stage. The sigmoid function is used to calculate mistakes, and Eq. (6) is the mathematical expression utilized to calculate the error.

$$\mathrm{Error_{BP}} = \mathrm{Layer_{Outputs}}\left(1 - \mathrm{Layer_{Outputs}}\right)\left(\mathrm{Output_{Target}} - \mathrm{Layer_{Outputs}}\right) \tag{6}$$

where

$\mathrm{Error_{BP}}$ represents the calculated back propagated error,

$\mathrm{Layer_{Outputs}}$ is the actual output of 'Layer,' $\mathrm{Output_{Target}}$ is known target value of the training tuple.

Using Eq. (7), modify the relevant weight until it has a minimum error in the third stage.

$$\mathrm{W}_{\mathrm{ij}_{new}} = \mathrm{W}_{ij_{old}} + \left(\mathrm{Error_{BP}} {}^{*} \mathrm{Layer_{Outputs}}\right) \tag{7}$$

The updated weight is $\mathrm{W}_{\mathrm{ij}_{\mathrm{new}}}$ while the original or previous weight is $\mathrm{W}_{ij_{old}}$. The weight is adjusted until the smallest variance is reached if the output does not match the target. Finally, BPNN[68] may be used to forecast the probability of a landslide.

Bagging. A Bagging[122] is a meta-estimator that fits base classifiers to random subsets of the original dataset, then combines their predictions (through voting or averaging) to get a final prediction[73]. Each base classifier is trained in parallel with a training set formed by replacing 'n' samples (or data) from the original training dataset with new data at random. Where 'n' is the original training set's size. The training sets for each base classifier are distinct from one another. Many of the original data points may be reproduced in the final training set, while others may be eliminated. By averaging or voting, the Bagging approach lowers overfitting; nevertheless, this increases bias, which is countered by the loss in invariance. For a given training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, sample T sets of n elements from dataset $D_i = (D_1, D_2, \ldots, D_T)$ are chosen using the replacement procedure. On each D_i, $(i = 1, \ldots, T)$ train a landslide model and acquire a sequence of T outputs $f_1(x), \ldots, f_T(x)$. The final aggregate classifier for the majority of votes is expressed in Eq. (8):

$$\bar{f}(X) = \text{sign}\left(\sum_{i=1}^{T} \text{sign}\left(f_i(X)\right)\right) \tag{8}$$

XGBoost. Boosting[122] is an ensemble modeling strategy that aims to create a strong classifier out of a large number of weak ones. It's done by putting together weak models to create a strong model. To begin, a model is created using the training data. The second model is then created, which attempts to correct the faults in the first model. This approach is repeated until either the entire training data set is properly predicted or the maximum number of models has been added. Gradient Boosting is a boosting approach in which a new predictor is built each iteration to fit the preceding predictor's pseudo-residuals[26]. Various studies show XGBoost is the better model in comparison to base models[123]. XGBoost[124] stands for Extreme Gradient Boosting and is a special implementation of Gradient Boosting. XGBoost employs second-order gradients and improved regularisation to get more accurate approximations.

The objective function of XGBoost is the total of all loss functions applied to all predictions, as well as a regularisation function for all predictors (j trees) is shown in Eq. (9).

$$obj(\theta) = \sum_{i=1}^{n} l\left(y_i - \widehat{y_i}\right) + \sum_{j=1}^{J} \Omega\left(f_j\right) \tag{9}$$

The first component of Eq. (9) represents the training loss that measures how well the model fit on training data and the second component is regularization which measures the complexity of trees. Optimizing training loss encourages predictive models and the same for regularization encourages simple models.

Where the first term denotes the loss function, the second term denotes the regularisation function, and f_j denotes a prediction from the j$^{\text{th}}$ tree. The final prediction will be achieved through Eq. (10).

$$\widehat{y_i} = \sum_{t=1}^{m} f_t(x_i) \tag{10}$$

where $f_t \in$Ϝ,whereϜ$= \{f_1, f_2, f_3, f_4, \ldots f_m\}$ is a set of base learners and x_i is representing the feature vector of i^{th} data point.

Random forest. The Random Forest (RF) model[81] is a data mining system that classifies large amounts of data accurately using an ensemble of decision trees. Decision trees are predictive models that use a collection of binary rules to select a target class[125]. A set of predictor factors, as well as the class to be predicted, are included in the data used to train the model. The RF model is made up of an ensemble learning approach that connects various landslide decision trees to estimate landslide possibility in a specific area spatially[75]. The RF model divides each node based on the best split in a subset of factors chosen at random by the node. As a result, the smaller the number, the better the split for the node in landslide prediction modeling. For landslide prediction mapping, each node of a normal tree can be split using the ideal split for all landslide prediction parameters. A decision tree built from a specific data set shows anomalies in many of the training data set's branches. One of the most common causes of these anomalies is over-fitting[71], which can be mitigated utilizing random forest approaches. These statistical techniques are used to classify a set of data. Based on the majority of votes obtained by each model, the final prediction is accepted.

Let $M_1, M_2, M_3, \ldots, M_i$ denotes a collection of classifiers used to build a composite model M^* on a given dataset D_{st_i}, where i denotes the total number of trees used to build a random forest.

D_{st_1}, D_2 .....D_{st_i} denotes the various training sets, and n denotes the total number of prediction classes for the models $M_1, M_2, M_3, \ldots, M_i$.

The Eq. (11) expresses the entropy or information of the data set D_{st} as:

$$Entropy(dataset) = -\sum_{i=1}^{n} \left(\frac{\left|D_{st_i}\right|}{\left|D_{st}\right|}\right) log_2 \left(\frac{\left|D_{st_i}\right|}{\left|D_{st}\right|}\right) \tag{11}$$

The Eq. (12) aids in obtaining the expected information for a single attribute

$$Entropy(attribute) = -\sum_{i=1}^{m}\left(\frac{\left|D_{st_k}\right|}{\left|D_{st}\right|}\right)*info\left(D_{st_k}\right) \tag{12}$$

where k denotes the number of attributes in total.

The Eq. (13) computes the information gain for each attribute and chooses the information gain with the highest value.

$$inf^{Gn}(Attr_k) = Entropy(D_{st}) - Entropy(Attr_k) \tag{13}$$

where $inf^{Gn}(Attr_k)$ denotes the information gain computed for each attribute $Attr_k$. This method, however, is biased toward tests with multiple outcomes and prefers to select attributes with a large number of values. The Gain Ratio method addresses the Information Gain method's limitation by normalizing the results using Eq. (14):

$$splitinfo = -\sum_{k=1}^{m}\left(\frac{\left|D_{st_k}\right|}{\left|D_{st}\right|}\right)log_2\left(\frac{\left|D_{st_k}\right|}{\left|D_{st}\right|}\right) \tag{14}$$

Rough set. The rough sets[126] are used in classification to uncover structural links in noisy and imprecise data. Rough sets naively fit the data before computing membership function values[71]. The preliminary set's categorization is based on the establishment of equivalence classes using the available training data. An equivalence class is made up of tuples that are indistinguishable. There are two parts to the supplied class C: a lower approximation and a higher approximation[127]. The lower approximation of C is made up of all tuples that are based on attribute information and are certain to belong to class C without ambiguity. The upper approximation of C is made up of all tuples that are based on attribute knowledge and cannot be defined as not belonging to class C. Each class has its own set of decision rules, which are displayed in a table.

The lower and upper approximations are expressed in Eqs. (15) and (16):

Upper Approximation:

$$\overline{R}X = \cup\{Y \in U/R : Y \cap X \neq \emptyset\} \tag{15}$$

Lower Approximation:

$$\underline{R}X = \cup\{Y \in U/R : Y \subseteq X\} \tag{16}$$

The basic goal of the rough set analysis is to infer (learn) approximations of concepts. It basically explains how to use arithmetic to find hidden patterns in data[128]. Data model information is maintained in a table in Rough Set. Each row represents a fact or an object (tuple). The facts often contradict each other. In Rough Set, a data table is referred to as an Information System. As a result, the information table can represent the input data for any domain. An information system is made up of a non-empty finite set of objects (U) and a non-empty finite set of attributes (A) (U, A). The elements of A are conditional characteristics. A decision table is defined as a table with one or more decision attributes. A decision system is defined as (U, A union d), where d is the decision attribute.

On tables, there are many things with similar qualities. One strategy for reducing table size is to store only one representative item for each set of objects with the same attributes. Indecipherable objects are often known as tuples. Any P subset A is related by an equivalence relation IND (P), where IND (P) stands for relation indiscernibility where x and y are indistinguishable from one another due to attribute P's nature. Indiscernibility is defined as follows in Eq. (17):

$$IND(P) = \{(x, y) \in \mathbb{U}^2 | \forall_a \in P, a(x) = a(y)\} \tag{17}$$

Following the successful discovery of hidden patterns, the next step is to feed these patterns into the landslide prediction model. Random Forest, Bagging, XGBoost, BPNN, and BN methods are used for classification and prediction. Random Forest, Bagging, XGBoost, BPNN, and BN classifiers receive the rules derived using the rough set technique as input for training purposes as shown in Fig. 4.

Validation and comparison methods. A final model is not acceptable unless it is validated, as proven by landslide studies. To evaluate the landslide models in this work, the authors employed the receiver operating characteristic (ROC) and other statistical estimators such as accuracy and precision. The most widely used performance indicators were computed using the confusion matrix[127] provided by the classifier. The entries in the confusion matrix have true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

Receiver operating characteristics (ROC) curve. ROC illustrates the percentages of true positive over false negative percentages to rank previous landslides cumulatively in decreasing order[129]. This includes determining success rates by utilizing the areas under ROC curves (AUROC). The AUROC is used to determine the prediction ability of a model. The AUROC values get smaller or equal to 50 when the prediction is poor, but AUC values closer to 100 indicate a reliable estimate. Furthermore, the greater the AUROC number, the better the model performance, and an AUROC value of 100 indicates exceptional performance. In recent studies, the AUROC curves have been recognized as one of the best and most often used tools for validating and comparing models. As a result, it has been widely used as a standard technique to analyze the model performance capability. The AUROC for the present study is computed using Eq. (18).

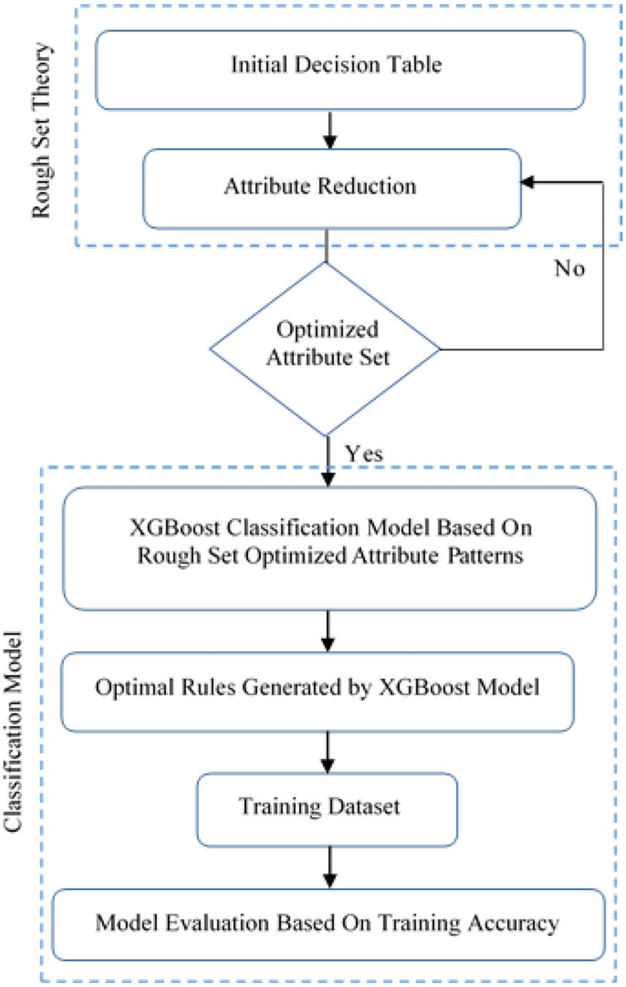

Rough Set Theory
Initial Decision Table
Attribute Reduction
No
Optimized Attribute Set
Yes
Classification Model
XGBoost Classification Model Based On Rough Set Optimized Attribute Patterns
Optimal Rules Generated by XGBoost Model
Training Dataset
Model Evaluation Based On Training Accuracy

Figure 4. Fusion of rough set theory and XGBoost for landslide prediction modeling.

$$AUC = \left(\frac{\Sigma TP + \Sigma TN}{P + N}\right) \tag{18}$$

Sensitivity, specificity, precision, accuracy, F1-score. Other than the ROC, popular and widely used statistical indices such as sensitivity or recall, specificity, precision, accuracy, and F1-score have also been used to assess the overall accurateness of the final results generated by various machine learning algorithms. Equations (19)–(23) establish the performance metrics under examination, for sensitivity, specificity, precision, accuracy, and F1 score. Sensitivity or recall is a metric that measures a model ability to predict real positive outcomes for each class label. In this study, sensitivity describes the right classification of landslide occurrence.

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) \tag{19}$$

The proportion of genuine negatives correctly detected is determined by the specificity metric. When it comes to categorizing negative samples, a model with a high specificity performs better. It indicates how many real negative cases have been identified.

$$Specificity = \left(\frac{TN}{TN + FP}\right) \tag{20}$$

In comparison to all anticipated positive samples, the precision represents the fraction of genuine positive samples. The model high precision means that it has a good chance of successfully categorizing positive samples.

$$Precision = \left(\frac{TP}{TP + FP}\right) \tag{21}$$

Factor	Quantitative measures
Erosion	0.19
Lithology	0.16
Bank failure	0.14
Hydrology	0.13
Rainfall	0.12
Joint failure	0.09
Land use	0.06
Toe failure	0.05
Overburden depth	0.04
Weathering	0.03
Slope	0.03
Road influence	0.01
Rock State	0.01
Landslide movement	0.00
Landslide style	0.00

Table 2. Significance of landslide conditioning factors (LCFs) using LASSO technique.

The accuracy of a database is defined as the percentage of samples accurately predicted. Out of all potential classifications, this phrase represents how many right classifications were made. It denotes the proportion of “True” to the total number of "True" and "False".

$$Accuracy = \left(\frac{TP + TN}{P + N}\right) \tag{22}$$

F1 score is the harmonic mean of precision and recall with a maximum value of 1 and the minimum value of 0. Precision and Recall are weighted in the F1 score, meaning that FP and FN are equally relevant. This is a considerably more useful metric when compared to "Accuracy." The problem with utilizing accuracy is that if we train the model on a severely imbalanced dataset, the model will learn how to properly forecast the positive class but not how to identify the negative class.

$$F1score = 2*(Precision*Recall)/(Precision + Recall) \tag{23}$$

Results

Independent test of causative factors.

The authors used the LASSO technique to choose conditioning factors for the landslide prediction model for the Uttarkashi region. The factors were chosen based on their weights in order to fit with the landslide prediction models. For the landslide study, causes with weights greater than zero were evaluated. Factors with a weight below or equal to zero were, on the other hand, removed from landslide modeling. According to the calculated LASSO values listed in Table 2 and represented in Fig. 5, 13 out of 15 factors are significant in landslide prediction (Quantitative Measures > 0).

The results from Table 2 and Fig. 5, it is disclosed that erosion has the highest value of 0.19, followed by Lithology (0.16), Bank Failure (0.14), Hydrology (0.13), Rainfall (0.12), Joint Failure (0.09), Land use (0.06), Toe failure (0.05), Overburden Depth (0.04), Weathering (0.03), Slope (0.03), Road Influence (0.01), Rock State (0.01), Landslide Movement (0.00), and Landslide Style (0.00). However, because the two conditioning factors, landslide movement and landslide style, were assessed as null prediction capability (measure = 0), they have no positive relation. In order to increase the accuracy of the final output, both elements were not incorporated in the current landslide modeling.

Multicollinearity test of causative factors.

A good selection of predictive conditioning factors is required in this type of study to ensure that these conditioning are independent of one another. The linear relationship between distinct independent conditioning factors determines a models overall accuracy, and excessive multicollinearity might reduce the model’s predictive capacity. Python programming was used to perform multicollinearity analysis on the selected dominating factors in this investigation. To identify the predictive conditioning factor and assess their multicollinearity, tolerance and VIF were utilized in this study. Table 3 reveals that the TOL value of all factors is greater than 0.1, and the VIF is less than 10, indicating that there are no severe multicollinearity issues. For this study, the factors with tolerance (≤ 0.1) and VIF (> 5) are eliminated. Therefore, the factors of landslide style and landslide movements are not considered for landslide prediction modeling.

Additionally, a heat map is generated to visualize the correlation matrix between the conditioning factors shown in Fig. 6. The colour value of the right band that ranges from low to high [-0.2, 1] indicates how closely the factors are correlated.

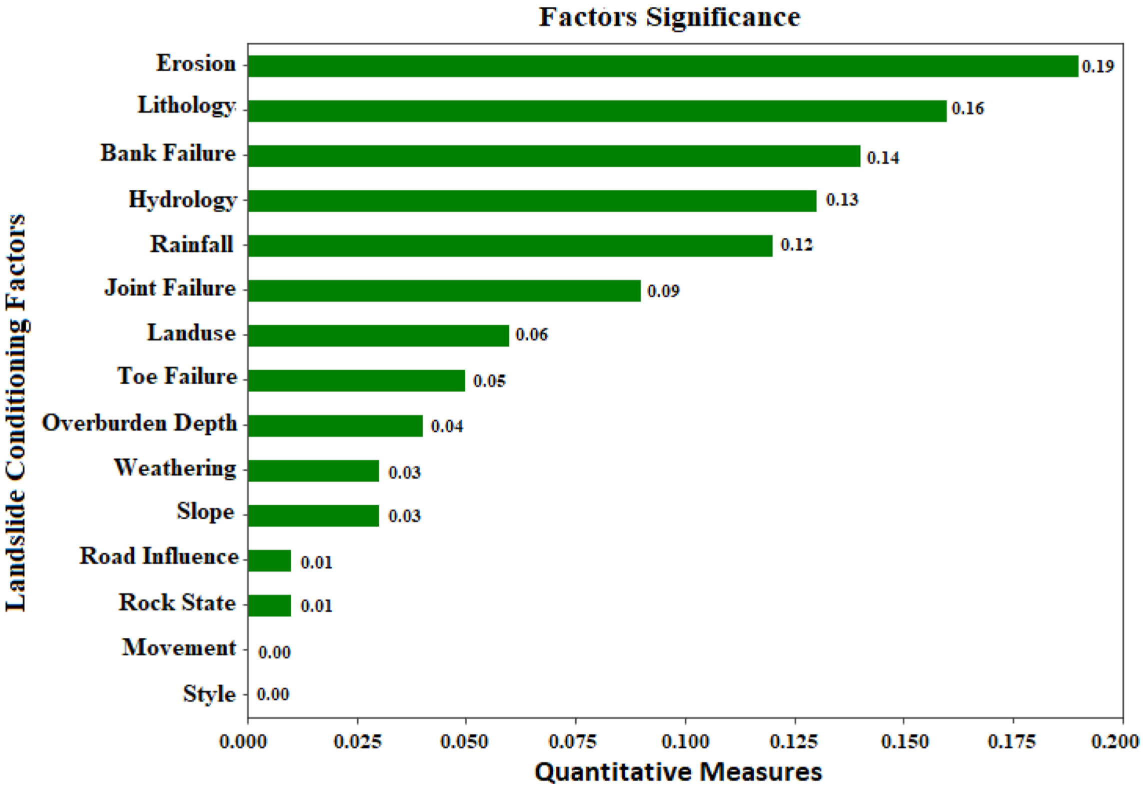


Figure 5. Significance of total 15 landslide conditioning factors using LASSO feature selection technique.

Factor	Tolerance	VIF
Bank failure	0.882102	1.133656
Joint failure	0.786793	1.270982
Erosion	0.736451	1.357864
Toe failure	0.665457	1.502726
Overburden depth	0.664323	1.505292
Road influence	0.436506	2.290920
Rainfall	0.387546	2.580337
Rock state	0.383740	2.605933
Lithology	0.360629	2.772933
Hydrology	0.332577	3.006826
Weathering	0.277825	3.599388
Slope	0.275084	3.635255
Land use	0.271750	3.679853
Landslide movement	0.144255	6.932168
Landslide style	0.103840	9.630224

Table 3. Multicollinearity analysis using tolerance and VIF for landslide conditioning factors used in landslide prediction modeling.

Model hyper-parameters tuning. The optimum hyperparameters for the ML models were discovered applying a grid search. The scikit-learn is used to implement all of the techniques in Python. HXGBRS and is suitable for building baseline models in landslide prediction analysis for contrastive ML models due to the fixed limit of parameters. For our proposed hybrid models, we trained and tested the model for a total of 554 instances. In the training process, different hyperparameters were tuned to improve the accuracy of the models. For HBNRS model, a grid search is performed tuning learning rate using values [0.1 to 0.9]. From the results, it was observed that = 0.1 was useful in yielding optimal accuracy. For HBPNNRS model, a grid search was performed to tune learning rate using values [0.01, 0.02, 0.03, 0.08, 0.09, 0.1, 0.2, 0.8, 1]. The learning rate at 0.2 produced the highest accuracy for the model. For HBRS, HXGBRS, and HRFRS models, a grid search was performed to tune a number of estimators from 10 to 1000. The HBRS model achieved the highest accuracy

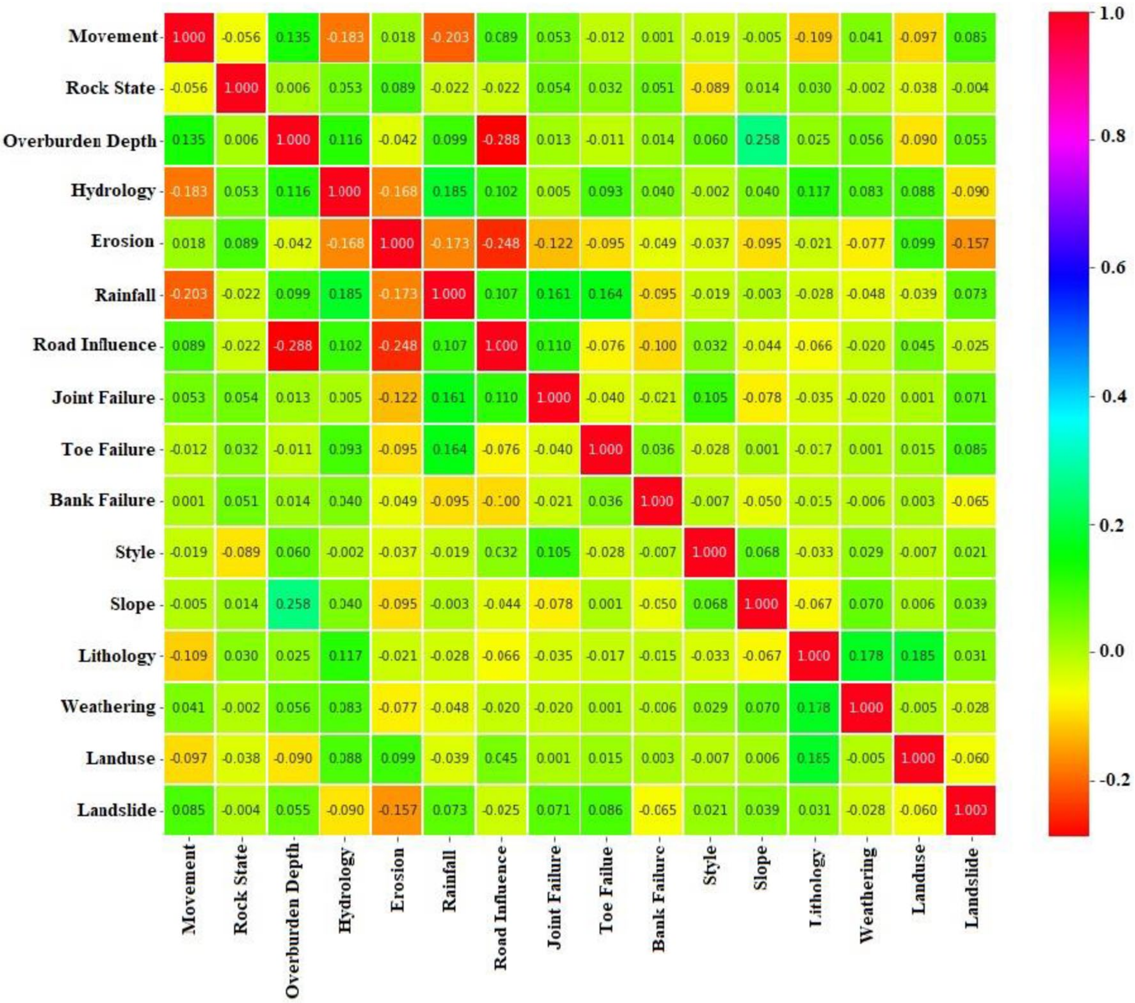


Figure 6. Heat map showing correlation matrix of landslide conditioning factors (LCFs). The color bar value of the right bar from high to low represent how closely the factors are correlated.

using 150 estimators, HXGBRS achieved the highest accuracy using 60 estimators and HRFRS achieved highest accuracy using 110 estimators.

Models validation and comparison. To improve the prediction capability of the landslide models, five hybrid ML approaches were implemented: Random Forest based Rough Set (HRFRS), Bagging based Rough Set (HBRS), XGBoost based Rough Set (HXGBRS), BPNN based Rough Set (HBPNNRS), and Bayesian Network based Rough Set (HBNRS). These hybrid models were constructed by combining the Rough set model with individual empirical methods. The rough set model is useful for identifying the optimized landslide patterns. These optimized patterns were utilized by individual empirical models for the training purpose of the model. The final model for landslide prediction in this work will be a hybrid method with the best prediction capabilities. The confusion matrices for the developed models using training and testing dataset is shown in Fig. 7.

The overall performance of the five approaches was evaluated using five statistical indicators (sensitivity, specificity, precision, accuracy, and F1-Score) and the AUROC curves as shown in Fig. 8. ROC curves have been employed in landslide hazard evaluation by a large number of researchers[130–133.] The prediction accuracy of all models was measured using ROC in this study. The AUC of the ROC curve was calculated using training and testing datasets. The performance of the models showed acceptable outcomes as shown in Fig. 8. In both the training and testing datasets, the HXGBRS model obtained the highest AUC values. The AUC values for the HXGBRS model were determined to be (0.937), respectively, for the testing dataset followed by HBPNNRS (0.924), HRFRS (0.904), HBRS (0.894), and HBNRS (0.883) respectively listed in Table 5. All of the models, however, produced reasonable results and showed to be promising solutions for predictive modeling in the region under investigation.

Additionally, the models were compared and evaluated using five statistical estimators (accuracy, sensitivity, specificity, precision, and F1-score) listed in Tables 4 and 5. The results of the study revealed that the HXGBRS model had the highest accuracy, precision, and F1-Score values (89.92%, 0.946, and 0.926), followed by HRFRS (88.48%, 0.936, and 0.916), HBRS (88.48%, 0.925 and 0.915), HBPNNRS (87.00%, 0.905 and 0.895), and HBNRS (75.50%, 0.851 and 0.824) respectively. Furthermore, the HXGBRS model gained the highest specificity (0.878)

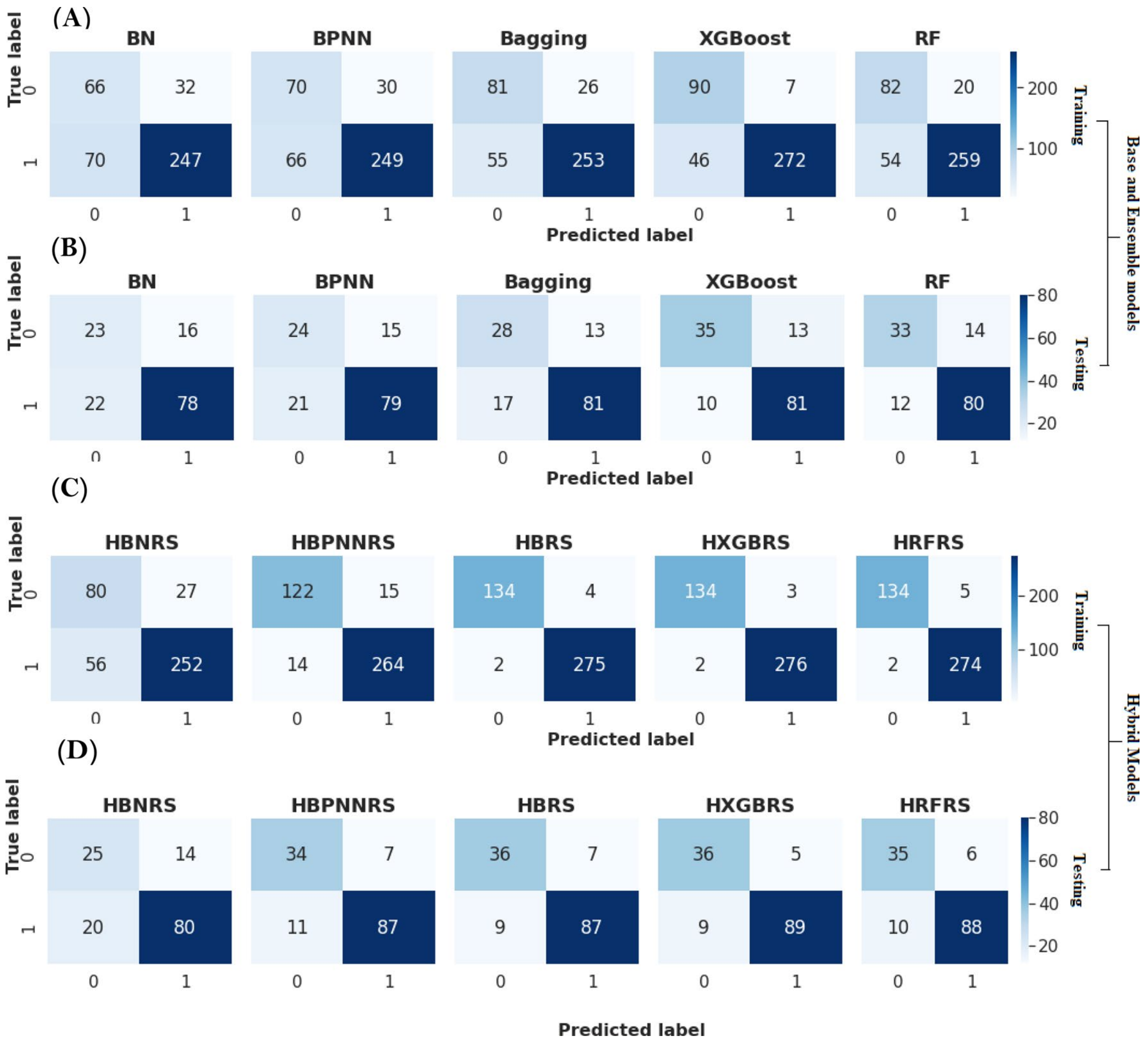


Figure 7. Confusion matrices showing classification results of landslides and non-landslides using base, ensemble and hybrid models based on training and testing dataset. It depicts the TN (top left), FP (top right), FN (bottom left) and TP (bottom right). (**A**) Represents the confusion matrix using base and ensemble models for training data. (**B**) Represents the confusion matrix using base and ensemble models for testing data. (**C**) Represents the confusion matrix using hybrid models for training data. (**D**) Represents the confusion matrix using hybrid models for testing data.

followed by HRFRS (0.853), HBPNNRS (0.829), HBRS (0.837) and HBNRS (0.641) respectively. Finally, utilizing testing data, all of the landslide prediction models used in this study show an acceptable goodness of fit. Moreover, it is also observed that the individual predictive capability of the landslide models is improved when integrated with the rough set model listed in Tables 4 and 5.

Dynamic visualization.

The ability to visualize spatial data through dynamic mapping has enormous potential[134]. Multiple, short-term, significantly different views of a data set, each constructed with a specific query in mind, are a key component of the analytical process, and dynamic maps that display observer-related behaviour are especially ideal for data exploration[135]. This method allows for the enrichment of traditional cartography and statistical representations of spatial data with dynamic visuals and transient symbolism, which provide on-demand additional details about a symbol or statistic[134]. Maps that provide more than a single static view of a spatial data set, or those that change over time, are becoming more popular and easier to create[136]. Each of these observer-related strategies includes interacting with dynamic views of data by identifying entities in single or multiple views and using temporary symbolism to show some attribute of the entity in the graphics that depict them[137]. Map design for visualization may now be more concerned with establishing and providing an acceptable flexible framework for exploratory dynamic mapping than with appropriate representation[136]. The dynamic displays consist of an interactive interface customized to a certain set of users, which enables experimenting with different combinations of conditioning factors in landslide study[138].

Visualizations are necessary for quickly identifying data issues that require more examination and analysis. For this study, a GIS-based user interface for dynamic interaction and visualization is built utilizing OpenStreet Map[139] and study area shapefile. To begin, the research area landslide locations were merged into OpenStreet

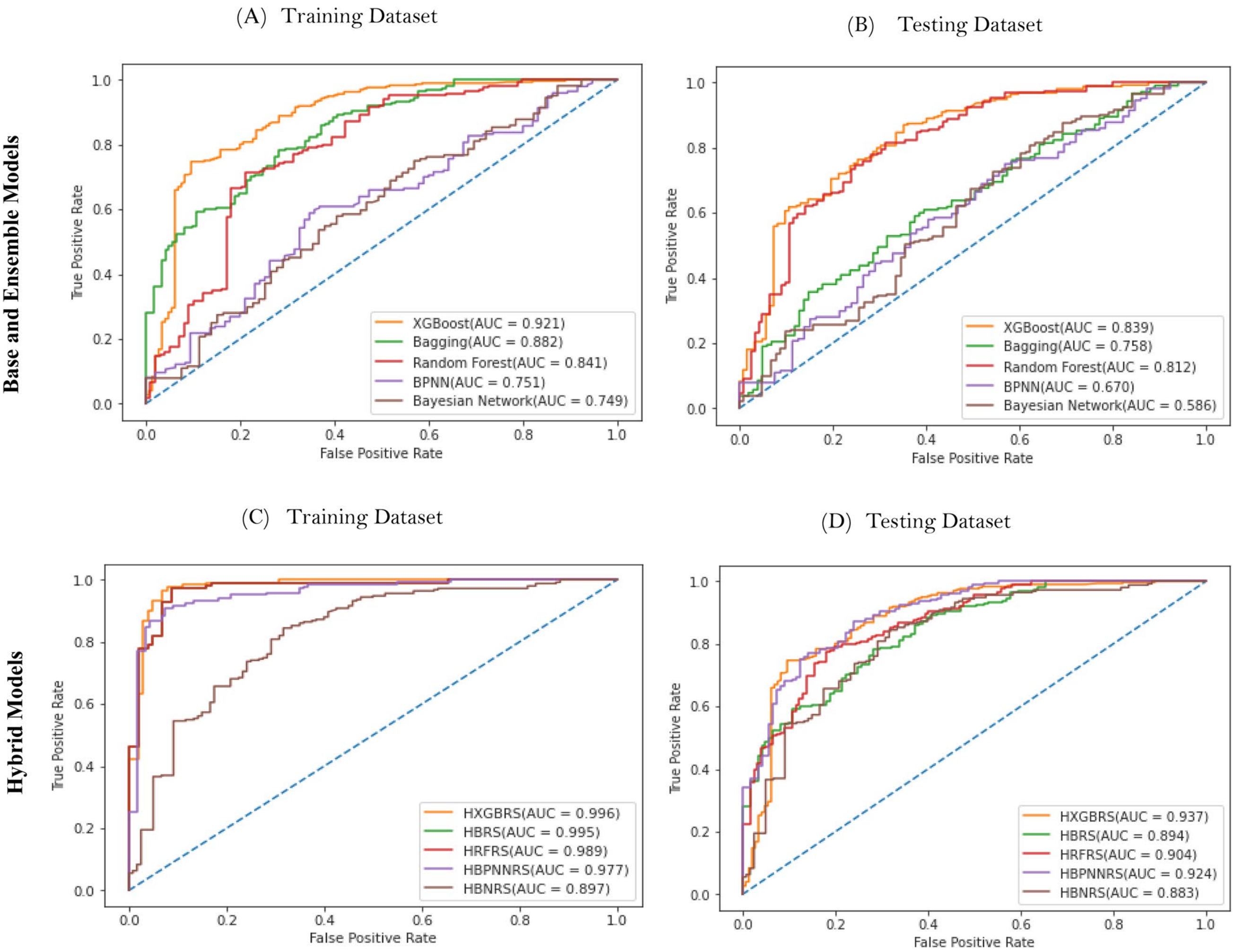


Figure 8. ROC curves used to analyze the prediction capability of various base, ensemble and hybrid landslide prediction models. (**A**) and (**C**) represents the performance for training dataset, (**B**) and (**D**) represents the performance for testing dataset.

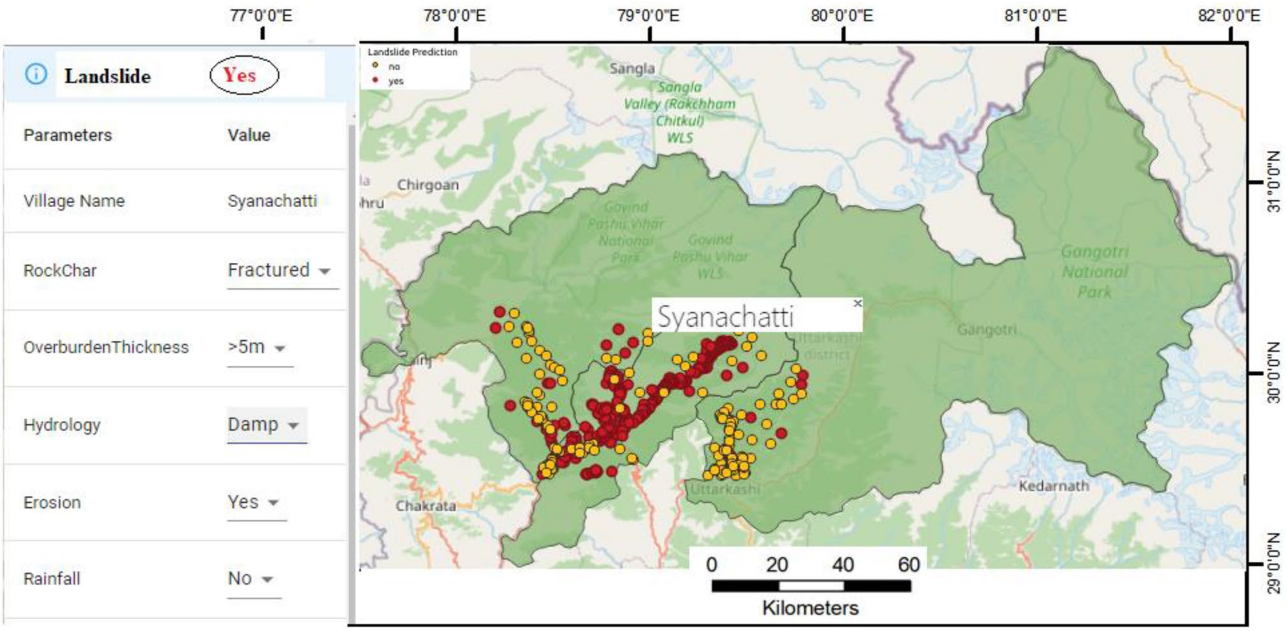
	Results on Training Dataset									
Performance Measures	**BN**	**HBNRS**	**BPNN**	**HBPNNRS**	**RF**	**HRFRS**	**Bagging**	**HBRS**	**XGBoost**	**HXGBRS**
Sensitivity	0.779	0.792	0.790	0.949	0.827	0.992	0.821	0.992	0.855	0.992
Specificity	0.673	0.747	0.700	0.890	0.803	0.964	0.757	0.971	0.927	0.927
Precision	0.885	0.903	0.892	0.946	0.928	0.982	0.906	0.985	0.974	0.989
F1-Score	0.861	0.850	0.837	0.947	0.874	0.986	0.861	0.988	0.881	0.990
AUC	0.749	0.897	0.751	0.977	0.841	0.989	0.882	0.995	0.921	0.996
Accuracy (%)	75.42	80.00	76.86	93.00	82.00	98.31	80.48	98.55	87.71	98.79

Table 4. Performance comparison of various base, ensemble and hybrid landslide prediction models on training dataset using statistical parameters based on confusion matrix.

maps by using the PostgreSQL database management system to connect the map to the landslide inventory. Second, the landslide triggering patterns[140–142] that had been identified using the HXGBRS method were used. The HXGBRS model provides the strongest prediction ability, according to the modeling findings. Thirteen landslide conditioning factors were considered in the current study to predict the landslide possibility in a non-landslide area. Landslides occur when each factor interacts with the others in some way. Users or analysts can adjust the values of the conditioning factors using the dropdowns option provided by the GIS-based user interface demonstrated in Fig. 9. The landslide class [yes, no] will vary if the user changes the value of any feature, according to the patterns detected. Analysts will be able to quickly detect the hidden combinations of elements that cause landslides with this approach.

Performance Measures	Results on Testing Dataset									
	BN	HBNRS	BPNN	HBPNNRS	RF	HRFRS	Bagging	HBRS	XGBoost	HXGBRS
Sensitivity	0.780	0.800	0.790	0.887	0.869	0.897	0.826	0.906	0.890	0.908
Specificity	0.589	0.641	0.615	0.829	0.702	0.853	0.682	0.837	0.729	0.878
Precision	0.829	0.851	0.840	0.905	0.851	0.936	0.861	0.925	0.861	0.946
F1-Score	0.803	0.824	0.814	0.895	0.859	0.916	0.843	0.915	0.875	0.926
AUC	0.586	0.883	0.670	0.924	0.812	0.904	0.758	0.894	0.839	0.937
Accuracy (%)	72.66	75.50	74.00	87.00	81.00	88.48	78.41	88.48	83.45	89.92

Table 5. Performance comparison of various base, ensemble and hybrid landslide prediction models on testing dataset using statistical parameters based on confusion matrix.



Landslide [✓ Yes]

Parameters	Value
Village Name	Syanachatti
RockChar	Fractured
OverburdenThickness	>5m
Hydrology	Damp
Erosion	Yes
Rainfall	No

Figure 9. GIS based user interface to predict the possibility of landslide and non-landslide events on changing the values of landslide conditioning factors considered for the study area.

Discussion

Landslide prediction study is vital for land management, planning, and development in hilly mountainous locations[143]. This benefit makes this method a useful tool for improving the accuracy of predicting landslides dynamic geological and geomorphological conditions[68]. It's difficult to create a precise landslide prediction model. Many academicians throughout the world have employed a variety of approaches and procedures to develop realistic landslide prediction models, but the subject has prompted disagreement among researchers[12,27,30]. As a result, innovative methods for developing and implementing landslide prediction models must be created and applied. The discovery of innovative methodologies for landslide prediction analysis has resulted from ongoing research. The goal of this research is to compare hybrid landslide prediction models developed utilizing the rough set theory approach.

The parameters examined in this study have a major impact on landslide disasters. As a result, the relationship between the outcomes and the factor distribution must be examined. Although the role of LCFs varies by location, there is no question that a combination of geo-environmental elements acts as a landslide regulator[52]. Selecting the right landslide conditioning factors in landslide hazard modeling leads to more accurate results and less noise, improving the model's predictive abilities[144]. However, it is understood that there is no set procedure or standard for selecting LCFs[117]. As a result, in landslide modeling studies, selecting proper LCFs is considered an important task. In order to analyze the landslide vulnerability in Uttarkashi, Uttarakhand, fifteen LCFs were chosen as an independent factor. The LASSO model was used to quantify and analyse the involvement of LCFs in landslides in the research region. The crucial importance of the triggering component is perfectly captured by LASSO[116]. Erosional activities, which are the triggering condition used in this study, have the greatest LASSO value. The significance of land use factor demonstrates the landslide influence of anthropogenic activities. Because all landslide conditioning factors had LASSO values greater than zero, it may be concluded that all of the identified elements have continued to play a role in landslide occurrence. Landslides are caused by a complex collection of circumstances, and there is no single factor that causes them. Calculating multicollinearity is crucial for multivariable landslide modeling. Because the model overall precision is influenced by a clear correlation between the LCFs. To test for multicollinearity, among the independent conditioning factors, TOL and VIF were used. The results reveal that the identified landslide influencing factors are not multicollinear listed in Table 3. After the identification of the most significant influencing factors, the dataset was randomly split into training set (75%) and testing set (25%) for the 554 landslide and non-landslide locations Empirical studies frequently show that using 20–30% of the data for testing and the remaining 70–80% for training produces the best results[13,145–147]. To overcome the issue of underfitting and overfitting due to the size of the dataset, a series of runs with varied numbers of data and testing data [70:30, 75:30, and 80:20] were performed for this study. For this investigation, a 75:25 ratio was observed suitable. The accuracy obtained applying 75:25 split yield better results than 80:20 and 70:30 splits.

The main goal of the study is to create a reliable landslide prediction model that can forecast future landslides using regional conditioning factors. This would also aid in the identification of landslide-prone locations in the study region. For this study, five hybrid models namely HBNRS, HBPNNRS, HRFRS, HBRS, and HXGBRS models were trained for landslide prediction modeling. Basically, the predictive capability of two single models and three ensemble models was improved through rough set fusion. The advantage of using rough set techniques with single and ensemble models is rough sets can be used to generate optimized patterns of landslides that improve the prediction capability of constructed models. From the study, it is revealed that the HXGBRS model outperformed in comparison with other hybrid models.

The AUC of the ROC curve was used to validate and compare landslide models, as well as numerous statistical indicators (sensitivity, specificity, Precision, Accuracy, and F1-score) to reveal the models' prediction capacity. Both the training and testing datasets were taken into account while validating models. The findings reveal that all of the models performed well. HXGBRS has outperformed other hybrid models in terms of accuracy and prediction, as evidenced by the highest values of AUC and Precision (for both training and validation datasets) listed in Tables 4 and 5 and shown in Fig. 8. Additionally, the study revealed that the generated hybrid models and common single machine learning techniques were successfully contrasted with training and testing datasets. In the Uttarkashi district of Uttarakhand, hybrid XGBoost with a rough set has shown a higher landslide prediction capacity. However, the interpretation of the results of such methods requires considerable attention because the performance of hybrid models is determined by the models design, which includes the structure of the training data, and the size of the input data.

After comparing the findings of this study to those of previous studies, the following conclusions were reached. Models with similar performance do not however have the same prediction capabilities. The results of these hybrid machine learning methods differ from those of other research conducted in different parts of the world[84,148–150] but they all indicate a high level of landslide prediction (AUC > 0.850). The fact that each expert used different dataset sources in his investigation illustrates the difference. The emergence of this variation in results was also influenced by differences in regional conditions[61]. As a result, the data selection and type, as well as the machine learning method, are critical.

Different modeling methodologies may produce different outcomes. The predicted performance of these hybrid machine learning models was shown to be more accurate than traditional statistical models for landslide prediction modeling in the majority of cases[64,86]. Because machine learning approaches are designed to automatically detect correlations between effective factors[151]. Comparing the results of the current study to those of earlier research using single or hybrid models, conducted in different, study regions and which had almost identical topographical and geological circumstances, it can be concluded that there is a difference in terms of the acquired AUC values and accuracy. In comparison to the approaches, namely Rotation Forest based Radial Basis Function (RFRBF) neural network (AUC = 0.891 and Accuracy = 82%)[152], Bagging based Reduced Error Pruning Trees (BREPT) accuracy (AUC = 0.872 and Accuracy = 80%)[153], and Support vector machine with cuckoo optimization algorithm(AUC = 0.738)[148], the hybrid in the current study produced high performances for Himalayan regions in terms of AUC and overall accuracy AUC = 0.937 and Accuracy = 89.92%). Normally, because of the HXGBRS technique, the model can correctly detect the impact of specific predictors even when there is a lot of additive noise in the data. Moreover, the XGBoost ensemble alone is a powerful method that is capable of handling missing values and supports regularization[154]. Additionally, the rough set method helped in identifying the optimized landslide patterns that were later utilized by XGBoost classifier[128].

Finally, a GIS based user interface is designed by the authors to predict the probability of landslides in the locations that are less or not probable of landslides. The reason behind this is the study area underlying geological characteristics are complex, with massive, sheared, fractured rocks occupying the majority of the region[88]. The rock structure is weak and unstable in shear. Because it is simple to vary due to atmospheric precipitation, surface water, groundwater, and land use, due to which the majority of slopes are expected to fall[92]. Human activities, particularly road development[94], are primarily concentrated in the middle and low elevation range, which, when combined with slope toe excavation, accumulation, blasting, mining, and other activities, will increase the slope's instability. Around the same time, deforestation, vegetation transformation to cultivated land, and the building will create many open places, causing landslides to occur[155]. Therefore, the user interface Fig. 9 designed for this study will be helpful to the analysis to identify the hidden patterns that trigger landslides[134]. Landslide prediction modeling could be a useful visualization tool[135] for preventing landslides, and enhancing the accuracy of the landslide prediction models is critical.

Conclusion

Landslides have become an extremely sensitive concern in the Uttarkashi region in recent years. Landslides have been a nightmare for residents in this area due to relatively young, fresh, and fragile geological formations. During the study phase of this type of research, a detailed evaluation of the future likelihood of landslides is an essential way to proceed. As a result, demarcating possible landslide occurrence zones in steep terrain is critical for development, urban planning, and land management. A landslide prediction modeling can be considered a beneficial tool in this context. Machine learning approaches have already produced a highly accurate outcome and, as a result, have become well-known in recent years. For the examination of slope instability, different machine learning techniques were chosen. The main goal of this study is to compare the results of five hybrid machine learning techniques in order to determine which method is best for assessing landslide prediction in the study area. The AUC results show that all of the machine learning methods worked well. However, the XGBoost-based rough set outperformed other hybrid machine learning methods in terms of accuracy (89.92%) and predictive capability. Moreover, the prediction capability of individual models (BN, BPNN, RF, Bagging, and XGBoost) was improved using the optimized patterns generated by the rough set method. The results generated by the HXGBRS approach had the highest accuracy in our research; the reason for this could be that multiple types of single classifiers are utilized, increasing the diversity of the models, and the same type of base models may induce overfitting. The methodologies and predicting factors utilized in this study were chosen from data collected, the study's aims, and the environmental circumstances of the study area. Visualization has proven to be a beneficial tool for swiftly examining prediction outputs, making it a trustworthy tool for dealing with the landslide prediction model. Authors will continue to investigate the use of hybrid methods in the field of landslide research in the future, attempting to use more dynamic factors and optimized ML methods to improve model performance and provide powerful visualization techniques by including dynamic and detailed map layers useful for decision-makers and managers in landslide disaster prevention.

Data availability

The data used in this research is taken from the open-access platform of geological survey of India (GSI). This platform comprises publicly available reports based on field studies prepared by the Geological Survey of India (GSI, https://www.gsi.gov.in/webcenter/portal/OCBIS/pageReports/pageGsiReports?_adf.ctrl-state=1gq1usi84_5&_afrLoop=935857055668031#!).

Received: 23 January 2022; Accepted: 19 October 2022
Published online: 22 November 2022

References

1. Simon, T., Goldberg, A. & Adini, B. Socializing in emergencies—A review of the use of social media in emergency situations. *Int. J. Inf. Manag.* **35**, 609–619 (2015).
2. Gariano, S. L. & Guzzetti, F. Landslides in a changing climate. *Earth Sci. Rev.* **162**, 227–252 (2016).
3. Huang, A.-B., Lee, J.-T., Ho, Y.-T., Chiu, Y.-F. & Cheng, S.-Y. Stability monitoring of rainfall-induced deep landslides through pore pressure profile measurements. *Soils Found.* **52**, 737–747 (2012).
4. Tao, Z. *et al.* Physical model test study on shear strength characteristics of slope sliding surface in Nanfen open-pit mine. *Int. J. Min. Sci. Technol.* **30**, 421–429 (2020).
5. Shanmugam, G. & Wang, Y. The landslide problem. *J. Palaeogeogr.* **4**, 109–166 (2015).
6. McColl, S. T. Chapter 2—Landslide causes and triggers. In *Landslide Hazards, Risks, and Disasters* 2nd edn (eds Davies, T. *et al.*) 13–41 (Elsevier, 2022). https://doi.org/10.1016/B978-0-12-818464-6.00011-1.
7. Gutiérrez, F., Parise, M., De Waele, J. & Jourde, H. A review on natural and human-induced geohazards and impacts in karst. *Earth-Sci. Rev.* **138**, 61–88 (2014).
8. Sidle, R. C., Gallina, J. & Gomi, T. The continuum of chronic to episodic natural hazards: Implications and strategies for community and landscape planning. *Landsc. Urban Plan.* **167**, 189–197 (2017).
9. Guzzetti, F. *et al.* Geographical landslide early warning systems. *Earth-Sci. Rev.* **200**, 102973 (2020).
10. Lv, L., Chen, T., Dou, J. & Plaza, A. A hybrid ensemble-based deep-learning framework for landslide susceptibility mapping. *Int. J. Appl. Earth Obs. Geoinf.* **108**, 102713 (2022).
11. Choi, K. Y. & Cheung, R. W. M. Landslide disaster prevention and mitigation through works in Hong Kong. *J. Rock Mech. Geotech. Eng.* **5**, 354–365 (2013).
12. Chen, W. & Zhang, S. GIS-based comparative study of Bayes network, Hoeffding tree and logistic model tree for landslide susceptibility modeling. *CATENA* **203**, 105344 (2021).
13. Chalkias, C., Ferentinou, M. & Polykretis, C. GIS-based landslide susceptibility mapping on the Peloponnese Peninsula, Greece. *Geosciences* **4**, 176–190 (2014).
14. Sarkar, S., Kanungo, D. P., Patra, A. K. & Kumar, P. GIS based spatial data analysis for landslide susceptibility mapping. *J. Mt. Sci.* **5**, 52–62 (2008).
15. Merghadi, A. *et al.* Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci. Rev.* **207**, 103225 (2020).
16. Fustos, I., Abarca-del-Río, R., Mardones, M., González, L. & Araya, L. R. Rainfall-induced landslide identification using numerical modelling: A southern Chile case. *J. South Am. Earth Sci.* **101**, 102587 (2020).
17. Mondini, A. C. *et al.* Landslide failures detection and mapping using Synthetic Aperture Radar: Past, present and future. *Earth-Sci. Rev.* **216**, 103574 (2021).
18. Chen, F., Yu, B., Xu, C. & Li, B. Landslide detection using probability regression, a case study of Wenchuan, northwest of Chengdu. *Appl. Geogr.* **89**, 32–40 (2017).
19. Wankhade, H. L. & Kumar, M. R. *Macro-Scale (1:50,000) Landslide, Susceptibility Mapping in Parts of Toposheet NOS. 53F/9, 53F/10, 53F/11, 53F/13, 53F/14, 53F/15, 53F/16 AND 53I/8, Uttarkashi, Tehri, Dehradun and Haridwar Districts, Uttarakhand.* www.gsi.gov.in (2016).
20. R, S. & Vinay. *Macro-Scale (1:50,000) Landslide, Susceptibility Mapping in Parts of Toposheet NOS. 53J/1 and 53J/5, Dehradun AND Uttarkashi Districts, Uttarakhand.* gov.gsi.in.
21. Shukla, K. & Gogoi, D. *Macro-Scale (1:50,000) Landslide, Susceptibility Mapping in Parts of Toposheet NOS. 53J/2 AND 53J/3, Tehri Garhwal, Dehradun and Uttarkashi Districts, Uttarakhand.*

22. Pham, B. T. *et al.* Ensemble machine learning models based on Reduced Error Pruning Tree for prediction of rainfall-induced landslides. *Int. J. Digit. Earth* **14**, 575–596 (2021).
23. Hong, H., Liu, J. & Zhu, A.-X. Modeling landslide susceptibility using LogitBoost alternating decision trees and forest by penalizing attributes with the bagging ensemble. *Sci. Total Environ.* **718**, 137231 (2020).
24. Stanley, T. A. *et al.* Building a landslide hazard indicator with machine learning and land surface models. *Environ. Model. Softw.* **129**, 104692 (2020).
25. Zhang, H. *et al.* Combining a class-weighted algorithm and machine learning models in landslide susceptibility mapping: A case study of Wanzhou section of the Three Gorges Reservoir, China. *Comput. Geosci.* **158**, 104966 (2022).
26. Sahin, E. K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* **2**, 1308 (2020).
27. Berhane, G. *et al.* Landslide susceptibility zonation mapping using GIS-based frequency ratio model with multi-class spatial data-sets in the Adwa-Adigrat mountain chains, northern Ethiopia. *J. Afr. Earth Sci.* **164**, 103795 (2020).
28. Reichenbach, P., Rossi, M., Malamud, B. D., Mihir, M. & Guzzetti, F. A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **180**, 60–91 (2018).
29. Chuang, Y.-C. & Shiu, Y.-S. Relationship between landslides and mountain development—Integrating geospatial statistics and a new long-term database. *Sci. Total Environ.* **622–623**, 1265–1276 (2018).
30. Amato, G., Palombi, L. & Raimondi, V. Data–driven classification of landslide types at a national scale by using Artificial Neural Networks. *Int. J. Appl. Earth Obs. Geoinf.* **104**, 102549 (2021).
31. Lombardo, L., Tanyas, H., Huser, R., Guzzetti, F. & Castro-Camilo, D. Landslide size matters: A new data-driven, spatial prototype. *Eng. Geol.* **293**, 106288 (2021).
32. Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R. & Kutz, J. N. Data-driven modeling and learning in science and engineering. *C. R. Méc.* **347**, 845–855 (2019).
33. Shano, L., Raghuvanshi, T. K. & Meten, M. Landslide susceptibility evaluation and hazard zonation techniques—A review. *Geoenviron. Disasters* **7**, 18 (2020).
34. Sonker, I., Tripathi, J. N. & Singh, A. K. Landslide susceptibility zonation using geospatial technique and analytical hierarchy process in Sikkim Himalaya. *Quat. Sci. Adv.* **4**, 100039 (2021).
35. Stamelos, I., Vlahavas, I., Refanidis, I. & Tsoukiàs, A. Knowledge based evaluation of software systems: A case study. *Inf. Softw. Technol.* **42**, 333–345 (2000).
36. Deng, C., Ji, X., Rainey, C., Zhang, J. & Lu, W. Integrating machine learning with human knowledge. *iScience* **23**, 101656 (2020).
37. Ghosh, P. & Lepcha, K. Weighted linear combination method versus grid based overlay operation method—A study for potential soil erosion susceptibility analysis of Malda district (West Bengal) in India. *Egypt. J. Remote Sens. Sp. Sci.* **22**, 95–115 (2019).
38. Panchal, S. & Shrivastava, A. K. Landslide hazard assessment using analytic hierarchy process (AHP): A case study of National Highway 5 in India. *Ain Shams Eng. J.* **13**, 101626 (2022).
39. Sharifi Teshnizi, E., Golian, M., Sadeghi, S. & Rastegarnia, A. Chapter 4—Application of analytical hierarchy process (AHP) in landslide susceptibility mapping for Qazvin province, N Iran. In *Computers in Earth and Environmental Sciences* (ed. Pourghasemi, H. R.) 55–95 (Elsevier, 2022). https://doi.org/10.1016/B978-0-323-89861-4.00041-5.
40. Kayastha, P., Dhital, M. R. & De Smedt, F. Application of the analytical hierarchy process (AHP) for landslide susceptibility mapping: A case study from the Tinau watershed, west Nepal. *Comput. Geosci.* **52**, 398–408 (2013).
41. Mandal, B. & Mandal, S. Analytical hierarchy process (AHP) based landslide susceptibility mapping of Lish river basin of eastern Darjeeling Himalaya, India. *Adv. Sp. Res.* **62**, 3114–3132 (2018).
42. Kaur, H., Gupta, S., Parkash, S. & Thapa, R. Knowledge-driven method: A tool for landslide susceptibility zonation (LSZ). *Geol. Ecol. Landsc.* **0**, 1–15 (2018).
43. Huang, F. *et al.* Uncertainty pattern in landslide susceptibility prediction modelling: Effects of different landslide boundaries and spatial shape expressions. *Geosci. Front.* **13**, 101317 (2022).
44. Asmare, D. Landslide hazard zonation and evaluation around Debre Markos town, NW Ethiopia—A GIS-based bivariate statistical approach. *Sci. Afr.* **15**, e01129 (2022).
45. Morgan, G. A., Gliner, J. A. & Harmon, R. J. Quantitative research approaches. *J. Am. Acad. Child Adolesc. Psychiatry* **38**, 1595–1597 (1999).
46. Mandaglio, M. C., Gioffrè, D., Pitasi, A. & Moraci, N. Qualitative landslide susceptibility assessment in small areas. *Procedia Eng.* **158**, 440–445 (2016).
47. Taşoğlu, E. & Abujayyab, S. K. M. Chapter 36—Comparison of the frequency ratio, index of entropy, and artificial neural networks methods for landslide susceptibility mapping: A case study in Pınarbaşı/Kastamonu (North of Turkey). In *Computers in Earth and Environmental Sciences* (ed. Pourghasemi, H. R.) 491–508 (Elsevier, 2022). https://doi.org/10.1016/B978-0-323-89861-4.00042-7.
48. Lombardo, L. & Mai, P. M. Presenting logistic regression-based landslide susceptibility results. *Eng. Geol.* **244**, 14–24 (2018).
49. Berhane, G. & Tadesse, K. Landslide susceptibility zonation mapping using statistical index and landslide susceptibility analysis methods: A case study from Gindeberet district, Oromia Regional State, Central Ethiopia. *J. Afr. Earth Sci.* **180**, 104240 (2021).
50. Neuhäuser, B. & Terhorst, B. Landslide susceptibility assessment using “weights-of-evidence” applied to a study area at the Jurassic escarpment (SW-Germany). *Geomorphology* **86**, 12–24 (2007).
51. Althuwaynee, O. F., Pradhan, B. & Lee, S. Application of an evidential belief function model in landslide susceptibility mapping. *Comput. Geosci.* **44**, 120–135 (2012).
52. Alsabhan, A. H. *et al.* Landslide susceptibility assessment in the Himalayan range based along Kasauli–Parwanoo road corridor using weight of evidence, information value, and frequency ratio. *J. King Saud Univ. Sci.* **34**, 101759 (2022).
53. Fan, W., Wei, X., Cao, Y. & Zheng, B. Landslide susceptibility assessment using the certainty factor and analytic hierarchy process. *J. Mt. Sci.* **14**, 906–925 (2017).
54. Pradhan, B. Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia. *Adv. Sp. Res.* **45**, 1244–1256 (2010).
55. Si, Y. *et al.* Predicting individual decision-making responses based on single-trial EEG. *Neuroimage* **206**, 116333 (2020).
56. Hemasinghe, H., Rangali, R. S. S., Deshapriya, N. L. & Samarakoon, L. Landslide susceptibility mapping using logistic regression model (a case study in Badulla District, Sri Lanka). *Procedia Eng.* **212**, 1046–1053 (2018).
57. Chen, W., Han, H., Huang, B., Huang, Q. & Fu, X. A data-driven approach for landslide susceptibility mapping: A case study of Shennongjia Forestry District, China. *Geomat. Nat. Hazards Risk* **9**, 720–736 (2018).
58. Choi, J., Oh, H.-J., Lee, H.-J., Lee, C. & Lee, S. Combining landslide susceptibility maps obtained from frequency ratio, logistic regression, and artificial neural network models using ASTER images and GIS. *Eng. Geol.* **124**, 12–23 (2012).
59. Corominas, J. *et al.* Recommendations for the quantitative analysis of landslide risk. *Bull. Eng. Geol. Environ.* **73**, 209–263 (2014).
60. Isinkaye, F. O., Folajimi, Y. O. & Ojokoh, B. A. Recommendation systems: Principles, methods and evaluation. *Egypt. Inform. J.* **16**, 261–273 (2015).
61. Gaidzik, K. & Ramírez-Herrera, M. T. The importance of input data on landslide susceptibility mapping. *Sci. Rep.* **11**, 19334 (2021).
62. Yaseen, Z. M. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. *Chemosphere* **277**, 130126 (2021).

63. Guo, Z., Shi, Y., Huang, F., Fan, X. & Huang, J. Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management. *Geosci. Front.* **12**, 101249 (2021).
64. Tiyasha, Tung, T. M. & Yaseen, Z. M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **585**, 124670 (2020).
65. Bhagat, S. K., Tung, T. M. & Yaseen, Z. M. Development of artificial intelligence for modeling wastewater heavy metal removal: State of the art, application assessment and possible future research. *J. Clean. Prod.* **250**, 119473 (2020).
66. Wang, H., Zhang, L., Yin, K., Luo, H. & Li, J. Landslide identification using machine learning. *Geosci. Front.* **12**, 351–364 (2021).
67. Palamakumbura, R. *et al.* Geological and geomorphological influences on a recent debris flow event in the Ice-scoured Mountain Quaternary domain, western Scotland. *Proc. Geol. Assoc.* **132**, 456–468 (2021).
68. Kainthura, P. & Sharma, N. Machine learning driven landslide susceptibility prediction for the Uttarkashi region of Uttarakhand in India. *Georisk Assess Manag. Risk Eng. Syst. Geohazards* **0**, 1–14 (2021).
69. Kumar, D., Thakur, M., Dubey, C. S. & Shukla, D. P. Landslide susceptibility mapping & prediction using Support Vector Machine for Mandakini River Basin, Garhwal Himalaya, India. *Geomorphology* **295**, 115–125 (2017).
70. Bangert, P. Chapter 3—Machine learning. In *Machine Learning and Data Science in the Oil and Gas Industry* (ed. Bangert, P.) 37–67 (Gulf Professional Publishing, 2021). https://doi.org/10.1016/B978-0-12-820714-7.00003-0.
71. Mehta, P. *et al.* A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **810**, 1–124 (2019).
72. Hong, H. *et al.* Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *CATENA* **163**, 399–413 (2018).
73. Wu, Y. *et al.* Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. *CATENA* **187**, 104396 (2020).
74. Shi, X., Wong, Y. D., Li, M.Z.-F., Palanisamy, C. & Chai, C. A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accid. Anal. Prev.* **129**, 170–179 (2019).
75. Sun, D., Wen, H., Wang, D. & Xu, J. A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. *Geomorphology* **362**, 107201 (2020).
76. McCoy, D., Mgbara, W., Horvitz, N., Getz, W. M. & Hubbard, A. Ensemble machine learning of factors influencing COVID-19 across US counties. *Sci. Rep.* **11**, 11777 (2021).
77. Aleryani, A., Wang, W. & de la Iglesia, B. Multiple imputation ensembles (MIE) for dealing with missing data. *SN Comput. Sci.* **1**, 134 (2020).
78. Can, R., Kocaman, S. & Gokceoglu, C. A comprehensive assessment of XGBoost algorithm for landslide susceptibility mapping in the upper basin of Ataturk dam, Turkey. *Appl. Sci.* **11**, 4993 (2021).
79. Montomoli, J. *et al.* Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *J. Intensive Med.* **1**, 110–116 (2021).
80. Martínez-Muñoz, G. & Suárez, A. Using boosting to prune bagging ensembles. *Pattern Recognit. Lett.* **28**, 156–165 (2007).
81. Aria, M., Cuccurullo, C. & Gnasso, A. A comparison among interpretative proposals for Random Forests. *Mach. Learn. Appl.* **6**, 100094 (2021).
82. Saha, S., Roy, J., Pradhan, B. & Hembram, T. K. Hybrid ensemble machine learning approaches for landslide susceptibility mapping using different sampling ratios at East Sikkim Himalayan, India. *Adv. Sp. Res.* https://doi.org/10.1016/j.asr.2021.05.018 (2021).
83. Paryani, S., Neshat, A. & Pradhan, B. Improvement of landslide spatial modeling using machine learning methods and two Harris hawks and bat algorithms. *Egypt. J. Remote Sens. Sp. Sci.* **24**, 845–855 (2021).
84. Panahi, M., Gayen, A., Pourghasemi, H. R., Rezaie, F. & Lee, S. Spatial prediction of landslide susceptibility using hybrid support vector regression (SVR) and the adaptive neuro-fuzzy inference system (ANFIS) with various metaheuristic algorithms. *Sci. Total Environ.* **741**, 139937 (2020).
85. Zhou, X., Wen, H., Zhang, Y., Xu, J. & Zhang, W. Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. *Geosci. Front.* **12**, 101211 (2021).
86. Wei, R. *et al.* Combining spatial response features and machine learning classifiers for landslide susceptibility mapping. *Int. J. Appl. Earth Obs. Geoinf.* **107**, 102681 (2022).
87. Ritu, J. Living with and responding to risk in the Uttarakhand Himalayas: A call for prioritizing lived experiences in research policy praxis. *Int. J. Disaster Risk Reduct.* **48**, 101499 (2020).
88. Joshi, L., Kotlia, B. & Singh, A. Geomorphic characteristics of landscape development and formation of lakes in the zone of Munsiari Thrust, Garhwal Himalaya, Uttarakhand, India. *Quat. Int.* **507**, 233–248 (2018).
89. Haigh, M. & Rawat, J. Landslide Disasters: Seeking Causes – A Case Study from Uttarakhand, India. In *Management of Mountain Watersheds.* (eds Krecek, J. et al.) 218–253. https://doi.org/10.1007/978-94-007-2476-1_18 (2012).
90. NASA. NASA. Global Landslide Catalog. https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog/h9d8-neg4#About (2019).
91. Bose, N. & Mukherjee, S. Estimation of deformation temperatures, flow stresses and strain rates from an intra-continental shear zone: The Main Boundary Thrust, NW Himalaya (Uttarakhand, India). *Mar. Pet. Geol.* **112**, 104094 (2020).
92. Sarkar, S., Ghosh, A., Kanungo, D. & Ahmad, Z. Slope stability assessment and monitoring of a vulnerable site on Rishikesh-Uttarkashi Highway, Uttarakhand, India. 2nd World Landslide Forum At: Rome (Italy). https://doi.org/10.1007/978-3-642-31445-2_8 (2011).
93. Pham, B. T., Pradhan, B., Tien Bui, D., Prakash, I. & Dholakia, M. B. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). *Environ. Model. Softw.* **84**, 240–250 (2016).
94. Komadja, G. C. *et al.* Geotechnical and geological investigation of slope stability of a section of road cut debris-slopes along NH-7, Uttarakhand, India. *Results Eng.* **10**, 100227 (2021).
95. Climate, Uttarkashi. https://en.climate-data.org/asia/india/uttarakhand/uttarkashi-33837/.
96. Jokar Arsanjani, J., Zipf, A., Mooney, P. & Helbich, M. An introduction to OpenStreetMap in geographic information science: Experiences, research, and applications. *Lecture Notes in Geoinformation and Cartography.* 1–15 https://doi.org/10.1007/978-3-319-14280-7_1 (2015).
97. Bolboacă, S. D. & Jäntschi, L. Sensitivity, specificity, and accuracy of predictive models on phenols toxicity. *J. Comput. Sci.* **5**, 345–350 (2014).
98. Kumar, A. *et al.* Assessment and review of hydrometeorological aspects for cloudburst and flash flood events in the third pole region (Indian Himalaya). *Polar Sci.* **18**, 5–20 (2018).
99. Novotny, P. J. *et al.* Do missing values influence outcomes in a cross-sectional mail survey?. *Mayo Clin. Proc. Innov. Qual. Outcomes* **5**, 84–93 (2021).
100. Duan, Y., Lv, Y., Liu, Y.-L. & Wang, F.-Y. An efficient realization of deep learning for traffic data imputation. *Transp. Res. Part C Emerg. Technol.* **72**, 168–181 (2016).
101. Zhang, S. Nearest neighbor selection for iteratively kNN imputation. *J. Syst. Softw.* **85**, 2541–2552 (2012).
102. Lévy, S., Jaboyedoff, M., Locat, J. & Demers, D. Erosion and channel change as factors of landslides and valley formation in Champlain Sea Clays: The Chacoura River, Quebec, Canada. *Geomorphology* **145–146**, 12–18 (2012).
103. George, K. J., Kumar, S. & Hole, R. M. Geospatial modelling of soil erosion and risk assessment in Indian Himalayan region—A study of Uttarakhand state. *Environ. Adv.* **4**, 100039 (2021).

104. Kumar, M., Rana, S., Pant, P. D. & Patel, R. C. Slope stability analysis of Balia Nala landslide, Kumaun Lesser Himalaya, Nainital, Uttarakhand, India. *J. Rock Mech. Geotech. Eng.* **9**, 150–158 (2017).
105. Andualem, T. G. & Demeke, G. G. Groundwater potential assessment using GIS and remote sensing: A case study of Guna tana landscape, upper blue Nile Basin, Ethiopia. *J. Hydrol. Reg. Stud.* **24**, 100610 (2019).
106. Kumar, V., Shanu, & Jahangeer,. Statistical distribution of rainfall in Uttarakhand, India. *Appl. Water Sci.* **7**, 4765–4776 (2017).
107. Haigh, M. J., Rawat, J. S. & Bartarya, S. K. Environmental correlations of landslide frequency along new highways in the Himalaya: Preliminary results. *CATENA* **15**, 539–553 (1988).
108. Beddoe, R. A. & Take, W. A. Loss of slope support due to base liquefaction: Comparison of 1g and centrifuge landslide flume experiments. *Soils Found.* **56**, 251–264 (2016).
109. Sah, N., Kumar, M., Upadhyay, R. & Dutt, S. Hill slope instability of Nainital City, Kumaun Lesser Himalaya, Uttarakhand, India. *J. Rock Mech. Geotech. Eng.* **10**, 280–289 (2018).
110. Miščević, P. & Vlastelica, G. Impact of weathering on slope stability in soft rock mass. *J. Rock Mech. Geotech. Eng.* **6**, 240–250 (2014).
111. Ghosh, A. *et al.* Slope instability and risk assessment of an unstable slope at Agrakhal, Uttarakhand. in *Proceedings of the India Geotechnical Conference, Guntur, India* (2009).
112. Komadja, G. C. *et al.* Assessment of stability of a Himalayan road cut slope with varying degrees of weathering: A finite-element-model-based approach. *Heliyon* **6**, e05297 (2020).
113. Gerrard, J. The landslide hazard in the Himalayas: Geological control and human action. *Geomorphology* **10**, 221–230 (1994).
114. Pollock, W., Grant, A., Wartman, J. & Abou-Jaoude, G. Multimodal method for landslide risk analysis. *MethodsX* **6**, 827–836 (2019).
115. Lee, C.-Y. & Cai, J.-Y. LASSO variable selection in data envelopment analysis with small datasets. *Omega* **91**, 102019 (2020).
116. Zhang, Z., Tian, Y., Bai, L., Xiahou, J. & Hancock, E. High-order covariate interacted Lasso for feature selection. *Pattern Recognit. Lett.* **87**, 139–146 (2017).
117. Chen, X. & Chen, W. GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. *CATENA* **196**, 104833 (2021).
118. Marcot, B. G. & Penman, T. D. Advances in Bayesian network modelling: Integration of modelling technologies. *Environ. Model. Softw.* **111**, 386–393 (2019).
119. Lan, M., Zhu, J. & Lo, S. Hybrid Bayesian network-based landslide risk assessment method for modeling risk for industrial facilities subjected to landslides. *Reliab. Eng. Syst. Saf.* **215**, 107851 (2021).
120. Bhagat, S. K. *et al.* Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models. *Environ. Pollut.* **268**, 115663 (2021).
121. Wythoff, B. J. Backpropagation neural networks: A tutorial. *Chemom. Intell. Lab. Syst.* **18**, 115–155 (1993).
122. González, S., García, S., Del Ser, J., Rokach, L. & Herrera, F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Inf. Fusion* **64**, 205–237 (2020).
123. Bhagat, S. K., Tung, T. M. & Yaseen, Z. M. Heavy metal contamination prediction using ensemble model: Case study of Bay sedimentation, Australia. *J. Hazard. Mater.* **403**, 123492 (2021).
124. Chen, T. & Guestrin, C. XGBoost. in *Proceedings of the 22nd {ACM} {SIGKDD} International Conference on Knowledge Discovery and Data Mining* (ACM, 2016). https://doi.org/10.1145/2939672.2939785.
125. Franco-Arcega, A., Carrasco-Ochoa, J. A., Sánchez-Díaz, G. & Martínez-Trinidad, J. F. Decision tree induction using a fast splitting attribute selection for large datasets. *Expert Syst. Appl.* **38**, 14290–14300 (2011).
126. Zhang, Q., Xie, Q. & Wang, G. A survey on rough set theory and its applications. *CAAI Trans. Intell. Technol.* **1**, 323–333 (2016).
127. Düntsch, I. & Gediga, G. Indices for rough set approximation and the application to confusion matrices. *Int. J. Approx. Reason.* **118**, 155–172 (2020).
128. Othman, M. L., Aris, I., Othman, M. R. & Osman, H. Rough-Set-and-Genetic-Algorithm based data mining and Rule Quality Measure to hypothesize distance protective relay operation characteristics from relay event report. *Int. J. Electr. Power Energy Syst.* **33**, 1437–1456 (2011).
129. Yuvaraj, R. M. & Dolui, B. Statistical and machine intelligence based model for landslide susceptibility mapping of Nilgiri district in India. *Environ. Chall* **5**, 100211 (2021).
130. Marjanović, M., Kovačević, M., Bajat, B. & Voženílek, V. Landslide susceptibility assessment using SVM machine learning algorithm. *Eng. Geol.* **123**, 225–234 (2011).
131. Wang, L.-J., Guo, M., Sawada, K., Lin, J. & Zhang, J. Landslide susceptibility mapping in Mizunami City, Japan: A comparison between logistic regression, bivariate statistical analysis and multivariate adaptive regression spline models. *CATENA* **135**, 271–282 (2015).
132. Conforti, M., Pascale, S., Robustelli, G. & Sdao, F. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). *CATENA* **113**, 236–250 (2014).
133. Wubalem, A. Landslide susceptibility mapping using statistical methods in Uatzau catchment area, northwestern Ethiopia. *Geoenviron. Disasters* **8**, 1 (2021).
134. Goodchild, M. F. Spatial thinking and the GIS user interface. *Procedia Soc. Behav. Sci.* **21**, 3–9 (2011).
135. *Geographic Information Systems for Geoscientists*. vol. 13 (Pergamon, 1994).
136. Dykes, J. A. Exploring spatial data representation with dynamic graphics. *Comput. Geosci.* **23**, 345–370 (1997).
137. Cook, D., Symanzik, J., Majure, J. J. & Cressie, N. Dynamic graphics in a GIS: More examples using linked software. *Comput. Geosci.* **23**, 371–385 (1997).
138. Stumvoll, M. J., Schmaltz, E. M. & Glade, T. Dynamic characterization of a slow-moving landslide system: Assessing the challenges of small process scales utilizing multi-temporal TLS data. *Geomorphology* **389**, 107803 (2021).
139. Alhamwi, A., Medjroubi, W., Vogt, T. & Agert, C. OpenStreetMap data in modelling the urban energy infrastructure: A first assessment and analysis. *Energy Procedia* **142**, 1968–1976 (2017).
140. Mi, J. *et al.* Vegetation patterns on a landslide after five years of natural restoration in the Loess Plateau mining area in China. *Ecol. Eng.* **136**, 46–54 (2019).
141. Massey, C. I., Petley, D. N. & McSaveney, M. J. Patterns of movement in reactivated landslides. *Eng. Geol.* **159**, 1–19 (2013).
142. Broothaerts, N. *et al.* Spatial patterns, causes and consequences of landslides in the Gilgel Gibe catchment, SW Ethiopia. *CATENA* **97**, 127–136 (2012).
143. Dai, F. C., Lee, C. F. & Ngai, Y. Y. Landslide risk assessment and management: An overview. *Eng. Geol.* **64**, 65–87 (2002).
144. Odhiambo Omuya, E., Onyango Okeyo, G. & Waema Kimwele, M. Feature selection for classification using principal component analysis and information gain. *Expert Syst. Appl.* **174**, 114765 (2021).
145. Huang, F. *et al.* Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. *CATENA* **191**, 104580 (2020).
146. Saito, H., Nakayama, D. & Matsuyama, H. Comparison of landslide susceptibility based on a decision-tree model and actual landslide occurrence: The Akaishi Mountains, Japan. *Geomorphology* **109**, 108–121 (2009).
147. Utomo, D., Chen, S.-F. & Hsiung, P.-A. Landslide prediction with model switching. *Appl. Sci.* **9**, 1839 (2019).

148. Balogun, A.-L. *et al.* Spatial prediction of landslide susceptibility in western Serbia using hybrid support vector regression (SVR) with GWO, BAT and COA algorithms. *Geosci. Front.* **12**, 101104 (2021).
149. Teja, T. S., Dikshit, A. & Satyam, N. Determination of rainfall thresholds for landslide prediction using an algorithm-based approach: Case study in the Darjeeling Himalayas, India. *Geosciences* **9**, 302 (2019).
150. Tien Bui, D., Hoang, N.-D., Nguyen, H. & Tran, X.-L. Spatial prediction of shallow landslide using Bat algorithm optimized machine learning approach: A case study in Lang Son Province, Vietnam. *Adv. Eng. Inform.* **42**, 100978 (2019).
151. Haigh, M. J., Rawat, J. S., Rawat, M. S., Bartarya, S. K. & Rai, S. P. Interactions between forest and landslide activity along new highways in the Kumaun Himalaya. *For. Ecol. Manag.* **78**, 173–189 (1995).
152. Pham, B. T., Shirzadi, A., Tien Bui, D., Prakash, I. & Dholakia, M. B. A hybrid machine learning ensemble approach based on a Radial Basis Function neural network and Rotation Forest for landslide susceptibility modeling: A case study in the Himalayan area, India. *Int. J. Sediment Res.* **33**, 157–170 (2018).
153. Pham, B. T. *et al.* Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: Hybrid machine learning approaches. *CATENA* **175**, 203–218 (2019).
154. Tiyasha, T. *et al.* Functionalization of remote sensing and on-site data for simulating surface water dissolved oxygen: Development of hybrid tree-based artificial intelligence models. *Mar. Pollut. Bull.* **170**, 112639 (2021).
155. Rawat, J. S. & Kumar, M. Monitoring land use/cover change using remote sensing and GIS techniques: A case study of Hawalbagh block, district Almora, Uttarakhand, India. *Egypt. J. Remote Sens. Sp. Sci.* **18**, 77–84 (2015).

Acknowledgements

The authors are grateful to Geological Survey of India for providing valuable landslide reports to carry out this study. https://www.gsi.gov.in/.

Author contributions

P.K. and N.S. carried out the whole experiment and prepared the manuscript with equal contribution.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.