scientific reports

Check for updates

OPEN

"Ensembled transfer learning approach for error reduction in landslide susceptibility mapping of the data scare region"

Ankit Singh[1], Nitesh Dhiman[1], K. C. Niraj[2] & Dericks Praise Shukla[1]✉

Landslide susceptibility map (LSM) plays an important role in providing the knowledge of slopes prone to future landslides. However, the applicability of LSM is often hindered due to high cost of data collection especially in mountainous region such as Himalayas. Therefore, this study proposes transfer learning approach (TL) to improve the performance of LSM by transferring the information from the data rich region (source) to data scare region (target). Two machine learning based model are trained in source area which is then transferred for prediction in target area and a source trained model combined with the knowledge of target is used for the prediction in the target area. The applicability of the proposed approach is tested considering Mandi district as source area and Kullu and Rudraprayag districts as target areas. Independent variables which included 11 landslide conducive factors from the source and target area, the independence was critically analysed using multi-collinearity test, while the geomorphic similarity was assessed using KL divergence method. Efficient machine learning method such as random forest (RF) and multi-layer perceptron (MLP) was used to train the models in both areas, statistical measure such as AUC-ROC, precision, recall, F-score, and accuracy were used to evaluate the performance of the LSMs. The results demonstrated the proposed approach for target area 1, the AUC value increased from 0.908 (Target trained on itself), to 0.942 (Target-Transfer Learned (TL)) and 0.959 (Target Combined) for RF and 0.896, 0.907 and 0.946 for MLP. Additionally, an increase in the precision was observed in RF while all other statistical measures increased to 0.023 for precision, 0.022 for recall, 0.023 for F-score and 0.066 for accuracy in the case of MLP. While on the other hand, in the case of target area 2, the AUC value for RF, ranged from 0.95 (Target area trained on itself), 0.80 (target TL) and 0.98 (target combined) and for MLP the AUC ranged from 0.82 for the source trained model while 0.84 for LSM obtained when both target and source data was used. Thus, the results revealed that proposed TL approach can be effective in improving the performance of LSM in data scare region in Himalayan region, thereby providing a promising approach in overcoming the issue of data limitation.

Landslides and their related hazard occur predominantly in the mountainous regions of the world, thereby affecting the societal and economic fabric of such areas. According to the data published by the World Health Organisation (WHO), between the years 1998–2017, landslides affected 4.7 million people and caused 18,000 deaths. The statistics in terms of the number of landslides and the losses are higher in developing economies primarily due to lack of susceptibility maps and adequate risk assessment. Landslides involve movement of soil, rocks, debris, and earth material due to several geo-hydrological and anthropogenic factors which in turn moves downslopes under the influence of gravity[1]. The geo-hydrological factors encompass rock types, curvature, inclination, aspect of the slope, presence of lineament and drainage networks, while anthropogenic factors include presence of road and settlement. In recent years the mountainous regions have experienced rapid urbanisation and heavy rainfall, particularly in the monsoon season, which leads to an increased frequency of landslides. Thus, effective landslide mitigation is required to minimize the losses stemming from it.

The mitigation mainly involves the identification of susceptible slopes (LSM), slope stabilization, and installation of an early warning system[2]. In view of this, LSM plays a key role in mitigation, thus quantifying the spatial likelihood of future landslides, providing critical information for informed decision making. The

[1]School of Civil and Environmental Engineering, Indian Institute of Technology, Mandi, Himachal Pradesh, India. [2]Department of Geomatics Engineering, Tribhuvan University, Pashchimanchal Campus, Pokhara, Nepal. ✉email: dericks@iitmandi.ac.in

preparation of LSM modelling saw a great advancement from the early 70's till today. The chronological evolution of the LSM method follows the order: deterministic, heuristic, statistical and machine learning methods. The early 70's witnessed immense use of deterministic approach which was subjective and had limited prediction area. With the increasing availability of remote sensing datasets, heuristic[3,4] and statistical methods[5–7], took over deterministic approach. However, these methods failed to capture the complexity and non-linearity involved in the landslide mechanism. Therefore, in today's time, machine learning methods such as logistic regression, support vector machine, random forest, multi-layer perceptron and deep neural network are employed extensively in landslide predictions. The workflow of these models involves dependent variables in the form of landslide conditioning factors and landslide inventory as independent variables, resulting in landslide probability of the location.

Previous studies such as Pourghasemi et al.[8] identified landslide prone areas in Iran employing fuzzy logic and analytical hierarcgy process. The result obtained a good performance providing general land use planning. Broeckx et al.[9] used logistic regression models to map of landslide susceptibility of Africa using 10,800 landslides. Ngo et al.[10] used recurrent neural network (RNN) and convolutional neural network (CNN) to obtain a national scale landslide susceptibility map for Iran using 4069 landslide obtained from government organisations and google earth. Similarly Azarafza et al.[11] used 112 landslide points to develop a deep convolutional neural network (CNN-DNN) to model the state of landslide susceptibility. Although state-of-the-art machine learning and deep learning models manifest the non-linearity of the landslide mechanism, these data-driven methods are prone to estimation error of susceptible areas due to insufficient landslide datasets. Additionally, this issue further exacerbates because of remoteness in complex mountainous regions and the high cost involved in data collection[12], thus exposing the most vulnerable populations to landslides. Current practices involve data acquisition through field visits, satellite image interpretation, and application within the same study area for prediction. However, data acquisition becomes practically challenging when the cost of field visits is high and time-consuming due to remote locations and complex mountainous terrains. Recent studies, Wang et al.[13], Defang et al.[14]suggest that landslide information obtained from data-rich regions can be transferred to data-scarce regions in landslide susceptibility mapping (LSM) using transfer learning (TL)[12].TL involves learning from a one task and reusing on a new task, therby reducing the need for data recollection. Recently, TL have been applied in land use mapping, water resource, vegetation, crop yield and natural hazard management. In landslides where the prediction results often encounters data incompleteness, transfer learning can provide a promising solution.

Previously, Ai et al.[15] used TL to overcome the requirement of big data in traditional machine learning. They considered Wenchuan and Jiuzhaigou as there source and target data respectively, artificial neural network was used to train the dataset from Wenchuan and transferred to Jiuzhaigou. The results demonstrated the effectiveness of TL in landslide susceptibility evaluation model. Similarly in Wang et al.[12] performed transfer learning using domain adaption and case based reasoning. Case based reasoning obtains knowledge from past events to train on new location, while domain adoption select data from the source area having similarity with the target area. They used generalized additive models for 10 study area located in Austria, Ecuador and Italy. The results revealed that TL approach successfully increase the model performance. Additionally, Fu et al.[16] applied TL approach in Three Gorges Reservoir. The results showed that when TL approach was applied the results were more refined compared to traditional machine learninug. Simialrly,[13,17,18] have applied TL approach in various parts of the world. Notwithstanding with the attempts made in the field of TL, its applicability on other parts is still an open topic of research where the source trained model was applied on the target area to prepare the LSM of the target area and its accuracy and superiority over other models was studied. Such type of TL approach has not been explored in the Himalayan or Indian context.

In the Indian context, critically understudied is the relevance of TL in the Himalayas especially in the northwestern region comprising Jammu and Kashmir, Himachal Pradesh, and Uttarakhand. This region which spreads to 33 million ha and habituated to 25 million people, accounted an increase in the number of landslides from 107 to 703 between 1971 and 2009. Geologically, the area comprises weak and fragile rocks predominantly sedimentary rocks with thrusts traversing 3 major mountainous ranges Main frontal thrust (MFT), Main boundary thrust (MBT) and Main central thrust (MCT). Adding to it, the increasing population, infrastructure and shifting climate pattern have led to an unprecedented rise in the number of landslides.

Thus, in view of the current limitation, this study proposes TL approach coupled with state-of-the-art machine learning and deep learning methods such as random forest (RF) and multi-layer perceptron (MLP) to identify areas probable to landslides. To do so, we have taken Mandi district as the source region since this district have rich data which was used to train the ML model for LSM preparation. While Kullu and Rudraprayag districts were considered as target regions where the availability of data is scarce. This study provides a more detailed methodology by transferring the knowledge learned from source area, applied to target areas. The LSMs were prepared using RF and MLP: (a) LSM of source area (b) LSM of target areas using training itself (c) LSM of target area, knowledge transferred from source area and (d) LSM of target area using the knowledge obtained from source as well as target areas. Moreover, the performance was analysed using statistical measures such as area under curve and receiver operator characteristics (AUC-ROC), precision, recall, and F-score, which was further evaluated using the Seed cell area index.

Therefore, this study provide meaningful insights into the reliability and application of the TL approach in the mountainous terrains of the Himalayas, overcoming the challenges of limited data availability. Consequently, it aids in the preparation of enhanced LSMs, providing policymakers and planners with valuable information for decision-making.

Study area and geology

Mandi district (source area)

Mandi district is located at the centre of Himachal Pradesh state and is bounded by Kangra on the northwest, whereas Kullu is in the northeast, sharing the longest boundary with the district (Fig. 1b). The capital, Shimla district, is southeast of Mandi, while Hamirpur and Bilaspur are on the western side, and Solan shares the shortest boundary in the south direction . The whole Mandi district drains into the Beas River basin, with only a small portion of the district pouring into the Satluj basin. The population of this district is predominantly agrarian; about 79% of the people are directly dependent on agriculture and activities related to agriculture and animal husbandry for their livelihood[19]. The district has a total area of 3950 km^2, and most areas of the Mandi district lie in the lesser Himalayan region (Fig. 1c), which represents rugged mountainous terrain and is composed of crystalline, fossiliferous sedimentary, and majorly metamorphic rocks[7]. The rocks here range from the Proterozoic period to the youngest quaternary, but they do not continuously lie upon one another. There are some gaps in the deposition in the area, and thrusting in the Himalayan region shows some underlying formations from different ages while covering most of the conformal continuous series of rocks [7]. The lithology of the Mandi district is oriented in the same parallel arrangement of the major thrust zones, which is the NNW-SSE direction (Fig. 2a. These thrust zones are the Main Boundary Thrust (MBT) and Main Frontal Thrust (MFT), which are tectonically active after the formation of the Himalayan range. The Proterozoic-aged rocks formation consisting of of three lithotypes namely Jutogh, Shali and Shimla formation. The oldest among them is the Jutogh Group, while the Shimla Group is the youngest in Proterozoic age (Fig. 2a). Between these two extremes,

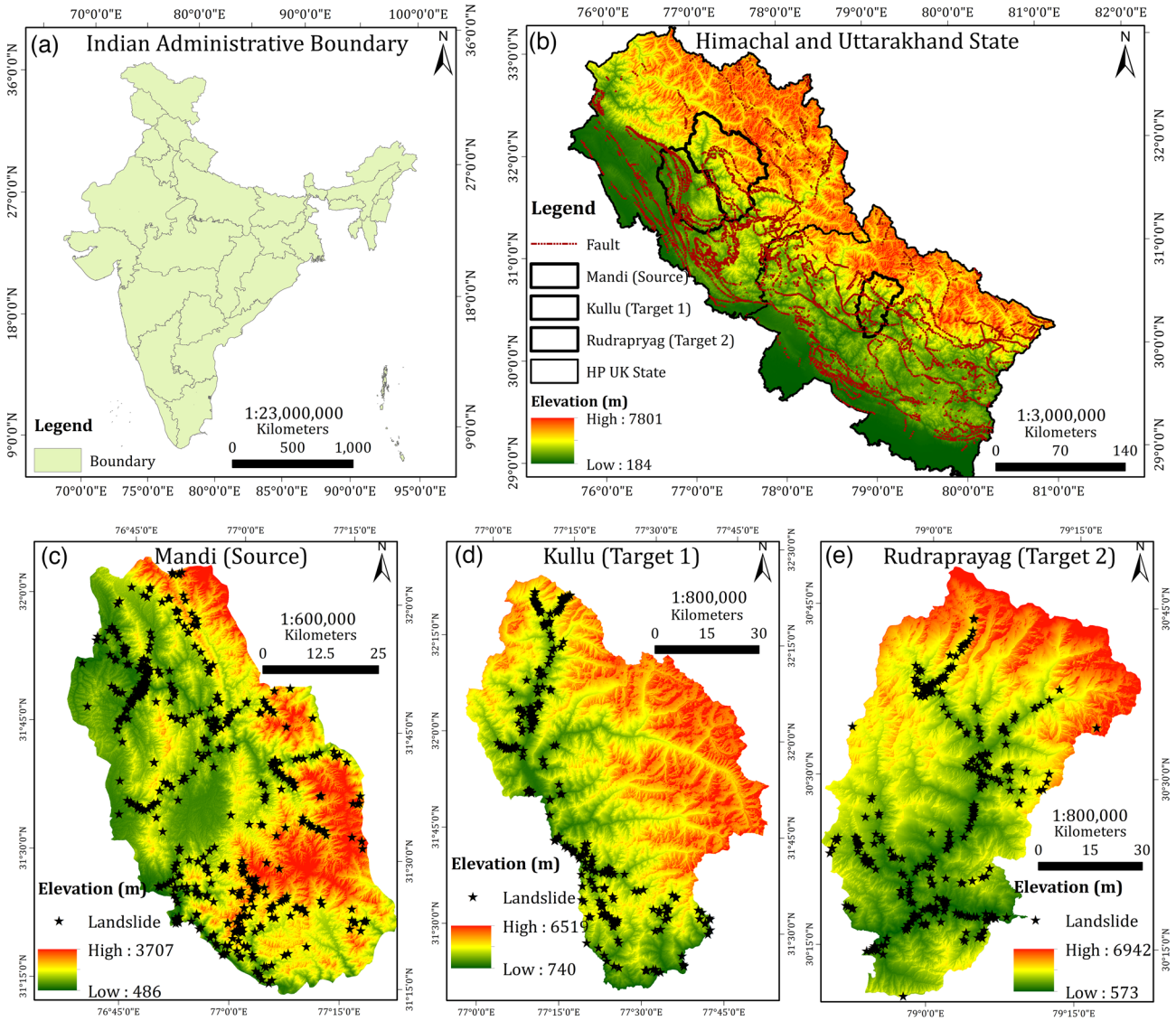


Fig. 1. Study area map **a** Himachal Pradesh administrative boundary demonstrating the major faults in the region, **b** Map showing the elevation and landslide points in the source area (Mandi), **c** Map showing the elevation and landslide points in target area 1 (Kullu). Finally **d** Map showing the elevation range of landslide points in target area 2 (Rudraprayag).

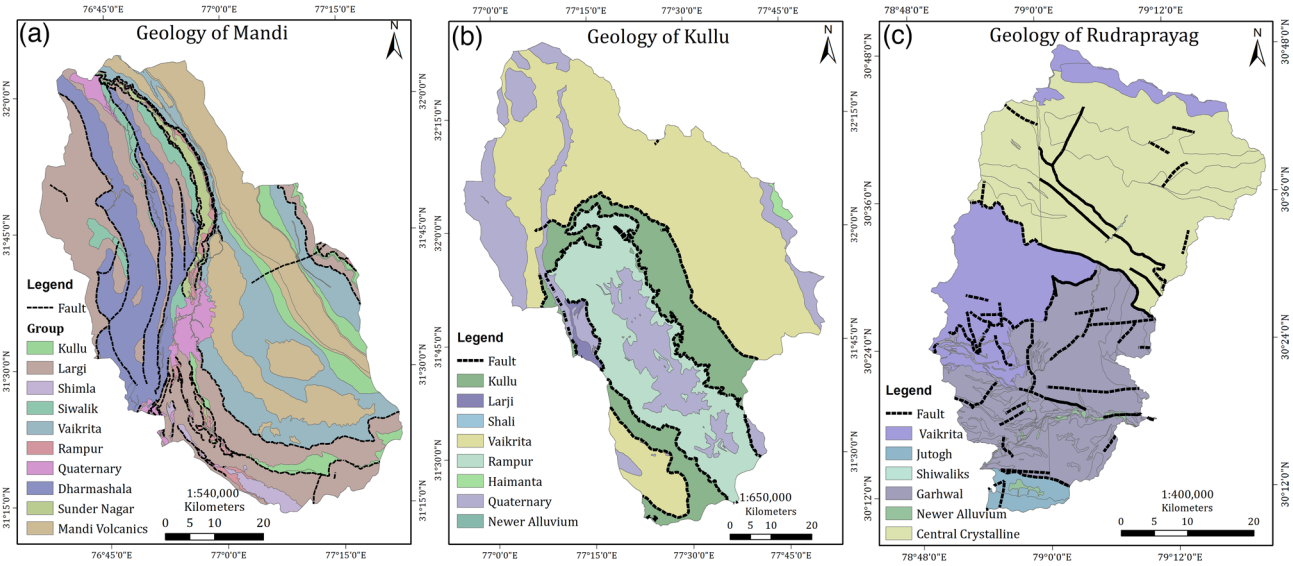


Fig. 2. **a** Geology of source (Mandi), **b** Geology of target area 1(Kullu) and **c** Geology of target area 2 (Rudraprayag).

the Shali formation is present, consisting of Phyllites, Quartzites, Dolomites, conglomerate, and Limestone (Shali/Sundernagar/Kullu/Largi formation)[20]. In the Paleozoic era has only basic volcanic intrusive, which were deposited during the Permian age. The whole Mesozoic era is missing from any outcrop from the Mandi district, while most of the area is covered by the Siwalik group of lower Pleistocene-Middle Miocene age (Fig. 2a). The youngest deposits of the quaternary age are formed along the alluvium, terraces, and fluvial deposits, mainly along the river and small channels of the Beas basin[21]. These young deposits mainly consist of sediments of fine clay, sand, pebbles-gravels and as large as cobbles and boulders of river origin. The metamorphic rocks are mainly exposed in upper ranges beyond the MBT, which traverses from almost the middle of Mandi and consists of phyllite, quartzite, limestone, shale, and dolomite[22]. These oldest rocks of the Proterozoic age are exposed to higher temperatures and pressures, making them foliated and having immense discontinuities like joints and bedding planes.

MBT marks the boundary between crystalline-metamorphic rocks and sedimentary rocks, however unconsolidated alluvium sediments are concentrated in the small plane valley known to be Sundernagar formation (Fig. 2a). The Siwaliks lie below the MBT and above MFT in the stratigraphic column, covering almost one-third of the extent of the Mandi district. The Siwaliks are reverse-graded tertiary deposits of outer Himalayan ranges, chiefly composed of Sandstone and conglomerate rock formations. The Siwaliks are divided into three categories, upper Siwaliks, middle Siwaliks, and Lower Siwaliks, in which grading ranges from higher to lower grade while going upper Siwaliks towards lower Siwaliks. The areas adjacent to this thrust zone mark many landslide zones, as the rocks are brittle and contain many discontinuities. Shali formation, which is located throughout the Mandi district, has grey-coloured phyllite- slates in their lower regime, while purple-coloured quartzite, and olive-coloured shales in their upper regime (Fig. 2a). Shali formations are deposited in continental settings during the Miocene ages, Chail thrust marks the upper margin of Shali formation. Above these, the rocks are mainly highly metamorphosed and crystalline containing almost no fossils, the rocks here are intercalated by quartzite layers of varied thickness in many parts of the district (Fig. 2a). These rocks are overlain by the Jutogh group, consisting of Biotite-Chlorite schist, Mica-Schist, and Gneisses.

Kullu (target area 1)

Kullu district is a part of the North-western Himalayas with a population of 437,903. The elevation ranges from 6430 to 751 m, with an average elevation of 3530.5 m (Fig. 1b). The district is bordered by Lahul Spiti, Kinnaur, Shimla, Mandi, and Kangra in northeast, southeast, south, southwest, and northwest directions respectively. It is bordered by the Rampur district in the south while the Mandi district is in the north (Fig. 1b). The climatic conditions of the district are cool and dry, the winters are severe from December to February. The study area experiences a temperate highland tropical climate with a dry winter climate. The average precipitation of the region is 1405.7 mm mostly occurring in June and September, fed by Beas, Parvati, and Tirthan rivers. The Beas River drains a major part of the district, which originates Pir-Panjal ranges, along the Beas River Kullu valley is present that forms the consists of majority of settlements[23]. The study area comprises mainly mountains, valleys, and terrace areas. Dhauladhar and Pir Panjal ranges which are mountainous peaks in the district trending NW-SE, the district acts as a connecting route to the Leh-Manali Highway from National Highway No 21, thus acting as a major transit route.

The geological sequence of Kullu district ranges from Precambrian to Quaternary. The geology of Kullu district consists of metamorphic and sedimentary rocks. The Precambrian rocks form the basement consisting of granite, schist, and gneisses which is surrounded by the Batal formation which consists of dark grey

carbonates, slates, and Phyllite[24]. Further, the Larji formation occurs in the western parts of the Kullu district, predominately consisting of slate, phyllite, and quartzite with dolomites and conglomerates[25] (Fig. 2b). The rocks of Batal formation are separated from Kullu group by the Main central thrust (MCT) which is defined to be the tectonic boundary between the greater Himalayan complex and the lesser Himalayan. The Kullu group consists of slate, phyllite, quartzite, and schist of the Mesoproterozoic age (Misra and Tewari 1988) (Fig. 2b). Moreover, the alluvium of the traverse along the Beas River striking N-S consists of clay, boulders, pebbles, and cobbles (Fig. 2b). The Neoproterozoic Rampur group consists of slate, phyllite, siltstone, and greywacke and occurs in southern and western parts of the Kullu district (Fig. 2b). The south Tibet detachment forms the boundary between the Vaikrita and Kullu groups which consists predominantly of quartzite, slate, and phyllite.

Rudraprayag (target area 2)

Rudraprayag district falls in the Indian state of Uttarakhand. Lies between the lesser and higher Himalayas (Fig. 1c), the district is located on the confluence of Mandakini and Alaknanda River. The elevation ranges from 800 to 8000 m, with altitude altitude-dependent climate. Orographic precipitation mainly occurs the period of June to September[26]. Geomorphologically, the area consists of extremely rugged and immature topography characterized by steeper slopes. The target area 2 is exposed to lithological formation of mainly hHigher and lower Himalayas[5,27]. Additionally, the rocks of the target area 2 is are traversed by major tectonic thrust zone namely North Almora thrust (NAT), Main Central Thrust (MCT) and Vaikrita thrust. The lesser Himalayan sequence consists of Dudhatoli group of pre-Cambrian age consisting predominantly of quartzite, and phyllites. Furthermore, the quartzite, phyllites succession is overlain by the Garhwal group of rocks. Notably, NAT thrust separates these two groups of rocks (Fig. 2c). The Garhwal group of rocks ranges from metamorphic (quartzites, phyllite, slate), sedimentary (dolomite) and igneous (volcanic association). Above the Garhwal group, Main central thrust or Ramgarh thrust (MCT-2) marks the boundary between the lesser Himalayan and Higher Himalayan sequences. North of the MCT and increase in the metamorphic grades is observed with major rocks of mica schist, porphyritic gneiss, schist, and calc-silicate zones. Furthermore, MCT popularly known as Vaikrita group separates the Jutogh group with Vaikrita group consisting of high-grade metamorphic rocks consisting of Garnetiferous, Genesis, Quartzite and Migmatites[5]. Therefore, due to the presence of major tectonic discontinuities present in the target area, the area is susceptible to major landslides as evident from past landslides such as Okhimath (2017), Phata-Byung area in 2001, 2005 and 2017 (Fig. 2c).

Methodology

Theory of transfer learning

The performance of machine learning models is often hindered by a lack of labelled datasets. Transfer learning (TL) refines the predictive capability of conventional machine learning results[28]. In a TL approach, knowledge obtained from a pre-trained model is applied to a new task through the transfer of knowledge from a related task that has already been learned.

TL requires two areas: the source and target areas. In the source area, it is assumed that data (landslide) are complete and adequate for training a machine learning model, while in the target area, it is assumed that the landslide data are small and inadequate (Fig. 3). Based on the relationship between the source and target areas, the TL approach is classified into four types: instance transfer, relational-knowledge transfer, feature representation transfer, and parameter transfer. Instance based transfer learning assigns weight from the source to target domain[29]. Relational-knowledge transfer approach focusses mainly upon the relational knowledge obtained from source and target areas. Further, a feature-based approach performs the knowledge transferability by creating a new feature and thereby matching with the target areas. Finally, a parameter transfer is based on transferring knowledge at the parameter level, where a well-trained model in the source domain, having learned

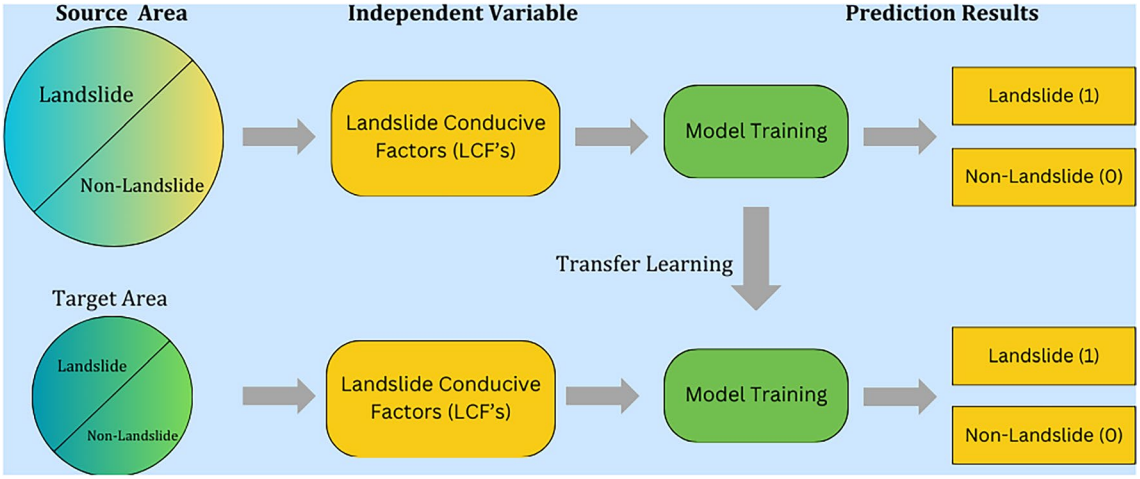


Fig. 3. Theory of transfer learning, the knowledge from the source area is transferred to the target area for prediction.

a well-defined structure, is transferred to the target model if the two tasks are related. Studies such as Zhiyong et al.[30] used instance transfer for landslide susceptibility Zigui basin and Wenchuan-Yiu area. However, in this study parameter transfer is used to transfer the knowledge from source (Mandi) to target (Kullu and Rudraprayag district) in LSM preparation (Fig. 3).

Preparation of landslide and non-landslide datasets

The landslide inventory of a region records the spatial location of previously occurred landslides in that region, which is used to train predictive models[31]. Preparation of the datasets are based on field surveys, satellite images and visual interpretation, recent LSM studies have used multi-temporal landslide inventory in point or polygon format[32]. However, studies have also indicated better performance of point data over polygon landslide data[33–35]. Point data was used in the collection of landslides for source as well as target areas, obtained from Geological Survey of India (GSI) bhukosh portal, collected after multiple field visits and image interpretation in Google Earth Pro software. Therefore, a total of 898 landslide points were obtained for Mandi (source) while 297 landslide locations were mapped in Kullu (target area 1) and 210 landslide locations were obtained for target area 2 (Rudraprayag district) (Fig. 4). Furthermore, in data-driven method the presence of non-landslide points is equally important as the landslide points in model training[5,36], Gupta and Shukla 2023). In landslide studies, the non-landslide points are obtained randomly from the study area creating a buffer of some distance from the existing landslide points[37,38]. However, this approach is unscientific way of negative sample selection resulting in the type 2 error wherein the chance of selecting an unstable area as a non-landslide location is higher[39]. Recent studies have shown other ways to objectively select the negative samples such as Mahalanobis distance[40], frequency ratio[16] and landslide buffer[37], among them frequency ratio approach is used in this study to obtain the non-landslide samples from the source and target areas. In this approach, the landslide points and landslide conditioning factors are used to prepare an initial landslide susceptibility map (LSM), classified into five classes namely very less, less, medium, high, and very high susceptibility classes. The region of very less and less susceptibility thus, represented areas with a low probability of any landslide occurrences, hence the non-landslide points were obtained randomly from these areas (Fig. 4).

Additionally, the LSM preparation requires partitioning of the datasets in the form of training and testing datasets. Strategies for sampling include random, slope, watershed etc., while the ratios are 70:30, 80:20 etc. [7,32,41]. In this study, a random sampling method and a 70:30 ratio were used to partition into training and testing datasets (Fig. 4).

Landslide conducive factors

Landslide conducive factors (LCFs) are broadly divided into time-dependent and time-independent factors. While the time-independent factors do not change significantly over time, the time-dependent factors change within a spatial–temporal time scale. The time-independent factors include slope, curvature, plane curvature, topographic wetness index, proximity to lineament and geology, while rainfall and land use land cover are classified as time-independent factors (Figs. 5, 6 and 7). Since, this study considers three different regions with dissimilar patterns of rainfall and land use, the time-dependent factors were not considered in LCFs for further study. Therefore, out of the many time-independent factors eleven factors were considered for LSM preparation. factors such as aspect, curvature, elevation, plane curvature, surface roughness count, slope and topographic wetness index were obtained from the ASTER-DEM with a spatial resolution of 30 m (Figs. 5, 6 and 7). The geological information of the source and target areas which included the lineament and rock type were obtained from the Bhukosh portal of the Geological Survey of India (Figs. 5c, 6c and 7c), road network was obtained from the Ministry of Road and Highways Government of India (Figs. 5d and 6d). Finally, the drainage network was extracted from the DEM using flow accumulation, and flow direction tools in the Arc environment to prepare the distance to stream map (Figs. 5e, 6e and 7e). All other raster were prepared on 30 m resolution, to match the cell size of the acquired cell size of topographical factors.

Independence and importance of LCFs

Machine learning methods requires the variable to be independent for effective model prediction. Dependent factors may affect the model performance and consequently decrease the ability of the model to map the spatial location of future landslides[8]. Methods used to check the independence among the dependent variables are the multicollinearity test, pearson correlation coefficient, chi-square statistics etc.[42]. In this study, multicollinearity test was used to check the independence of the factors considered in LSM preparation. The multicollinearity test was performed by calculating the tolerance and variance inflation factors (VIF) (Eqs. 1 and 2).

$$VIF = \frac{1}{1 - R^2} \tag{1}$$

$$\text{Tolerance} = \frac{1}{VIF} \tag{2}$$

where R represents the correlation coefficient between the causative factors, factors with VIF > 10 and tolerance > 0.1 were considered redundant and removed for further analysis. The VIF and tolerance were calculated in a Python environment.

Additionally, the machine learning models are black box models in nature, providing limited knowledge and understanding of how predictions are made based on the independent variables . Thus, to address the issue of black box nature, feature importance (information gain) was used to calculate the contribution of each LCF in model prediction.

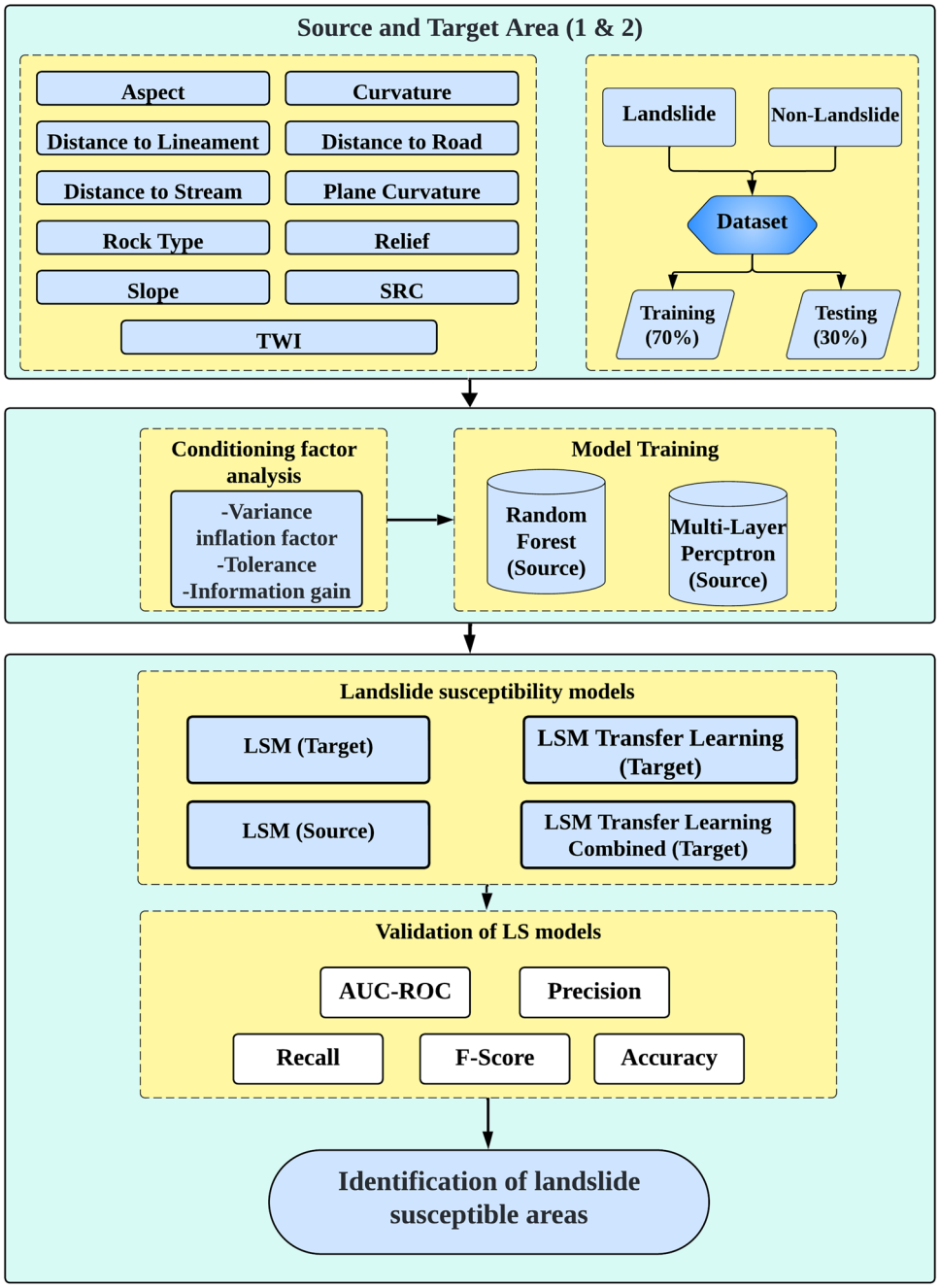


Fig. 4. Flow chart adopted in the study.

Kullback–Leibler (KL) divergence

TL involves the migration of knowledge between the two regions; thus, the two areas should be similar to achieve better knowledge transfer. In transfer learning (TL), a model is trained on the source environment and the pre-trained is applied to a target environment. The similarity test is important in TL for reasons that include: task relevance i.e. if the two regions are similar, the knowledge learned by the pre-trained model is more likely to be useful for the target area. Also, when the two environments are similar the transfer-learned model will be trained more effectively. However, when the source and target domain are dissimilar the performance can be compromised thereby causing a negative transfer. However, if the source and target domains are very different then domain shift between the two environments and weight adjustment methods can be applied[13]. Other TL studies have employed similarity-based approaches such as case-based reasoning and domain-based adoption [18], yet they do not account for the statistical distribution of the data. The Kullback–Leibler (KL) also known as relative entropy, is a statistical measure to estimate the difference between the two probability distributions and was used to establish the similarity relationship between the LCFs of source and target areas (Eq. 3).

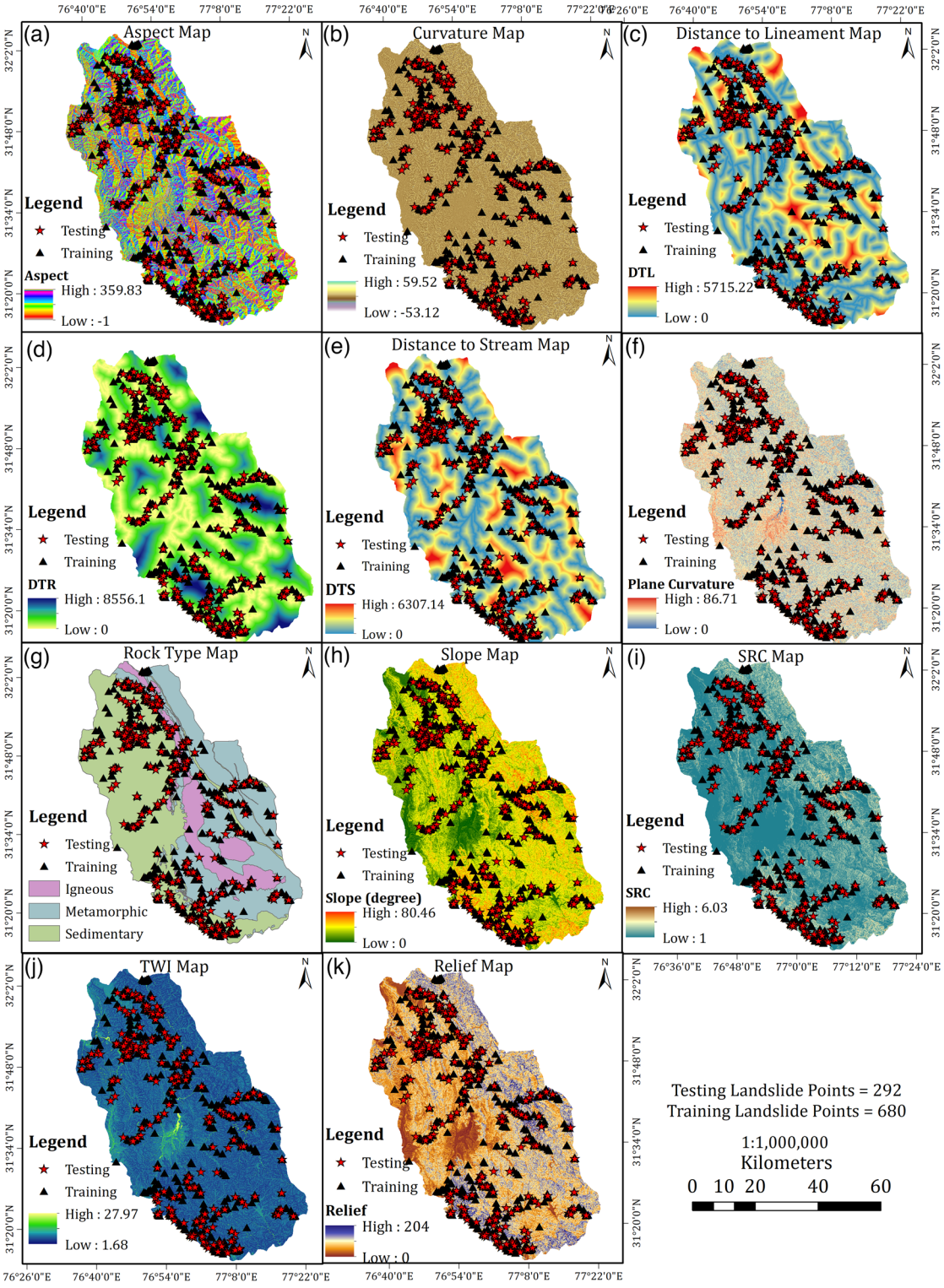


Fig. 5. Landslide conducive factors for source area **a** Aspect, **b** Curvature, **c** Distance to Lineament, **d** Distance to Road, **e** Distance to Stream, **f** Plane Curvature, **g** Rock Type, **h** Slope, **i** Surface Roughness Count (SRC), **j** Topographic Wetness Index (TWI) and **k** Relief.

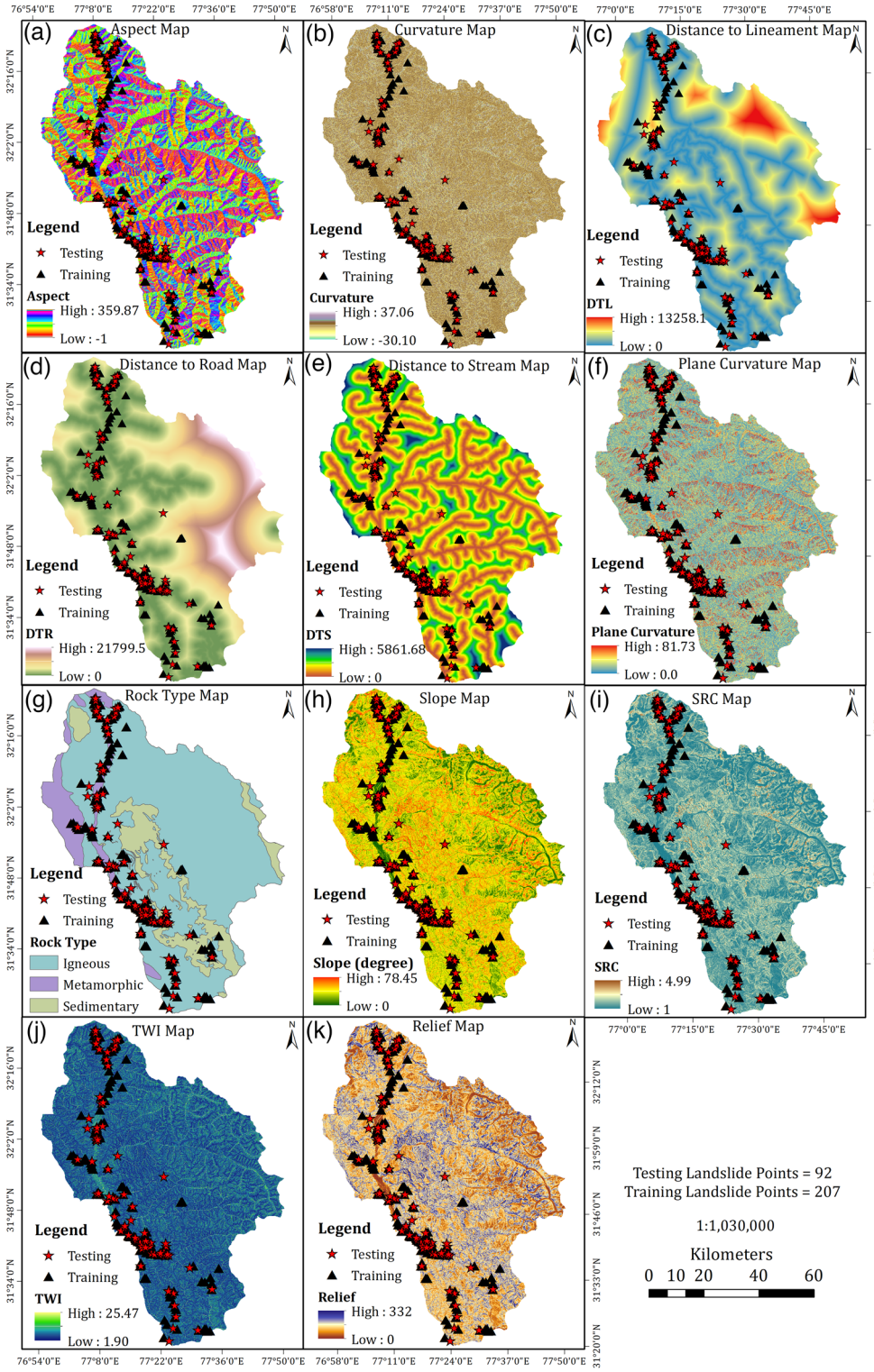


Fig. 6. Landslide conducive factors for target area 1 **a** Aspect, **b** Curvature, **c** Distance to Lineament, **d** Distance to Road, **e** Distance to Stream, **f** Plane Curvature, **g** Rock Type, **h** Slope, **i** Surface Roughness Count (SRC), **j** Topographic Wetness Index (TWI) and **k** Relief.

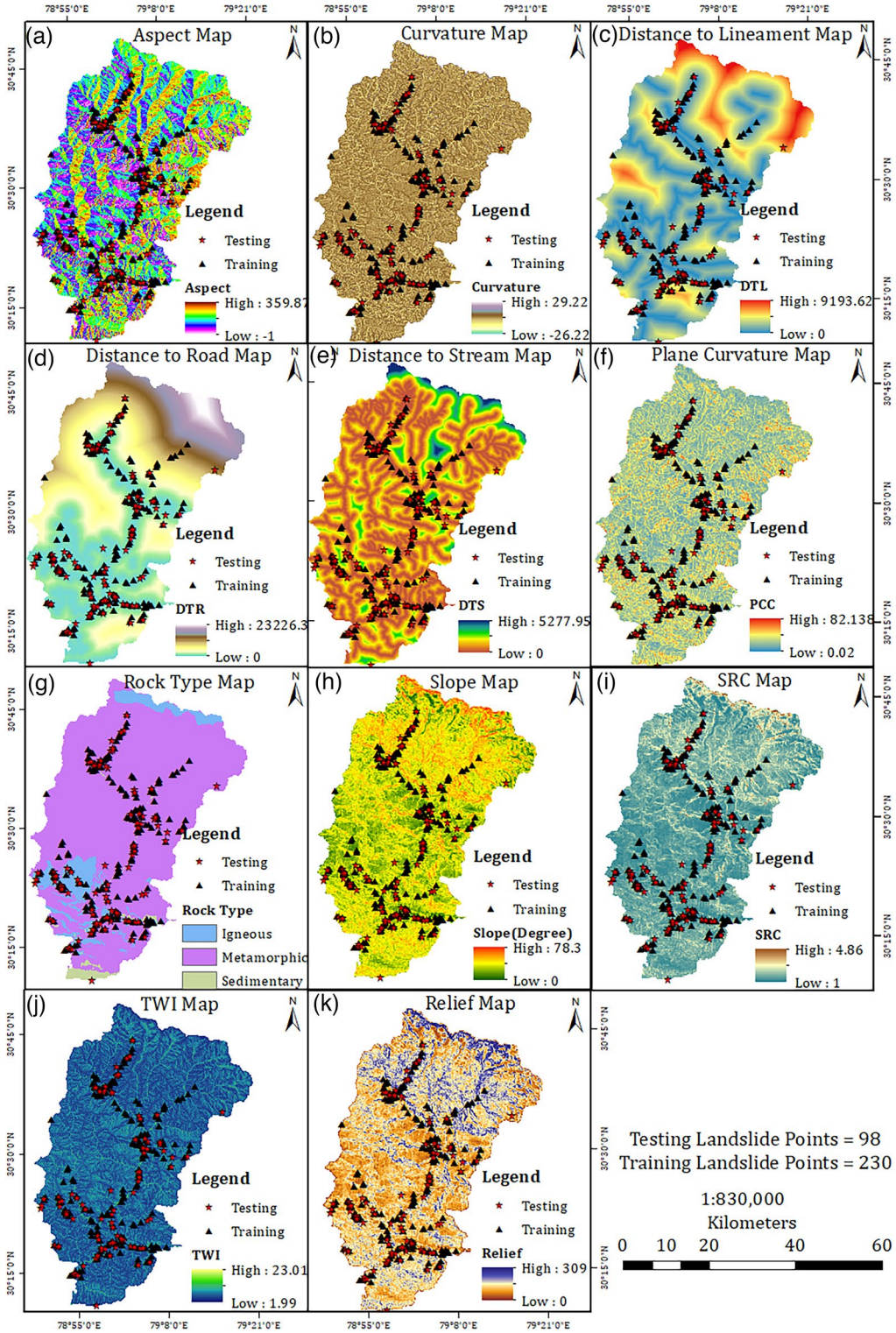


Fig. 7. Landslide conducive factors for target area 2 **a** Aspect, **b** Curvature, **c** Distance to Lineament, **d** Distance to Road, **e** Distance to Stream, **f** Plane Curvature, **g** Rock Type, **h** Slope, **i** Surface Roughness Count (SRC), **j** Topographic Wetness Index (TWI) and **k** Relief.

$$KL\left(P||Q\right) = \int P\left(x\right) \log \frac{P(x)}{Q(x)} dx \tag{3}$$

where P(x) and Q(x) represent the probability distribution of the LCFs for source and target region. The value ranges from 0 to infinity, the KL = 0 value represents identical distribution while KL > 0 is greater dissimilarity.

Random forest

The Random Forest method is an ensemble of machine learning algorithm used commonly to solve classification and characterisation types of problems. The algorithm works on combining multiple decision trees to reach the results[43]. The decision tree (DT) constructs multiple trees where it splits to subset the data. The trees are then trained using classification and regression training algorithms, while the quality of the split is assessed using the information gain ratio[44]. Although a single DT is based upon a supervised machine learning method, they are prone to problems such as bias and overfitting which results in the reduction of performance and accuracy. Furthermore, when multiple DT forms an ensemble in RF algorithms, the model performs better. Therefore, RF uses a collection of DT where each tree is composed of data samples drawn from the training set with replacement popularly known as a bootstrap sample, whereas the remaining is used to obtain the out-of-bag error (OOB). RF method depends upon three hyperparameters, node size, number of trees and number of features (Fig. 8a). There is no thumb rule in the selection of the number of trees and node size, however, the number of trees should be grown until the OOB stabilizes. Finally, the majority votes obtained from each node are used in final predictions.

Multi-layer perceptron (MLP)

The multi-layer perceptron (MLP) is a widely used artificial neural network in machine learning used in the field of pattern recognition and time series prediction etc[45–48]. Mulit-layer is a fully connected dense layer, which transforms the input dimension to a desired dimension[49]. In landslides, the MLP performs better in depicting the non-linear mechanism involved in the landslides. The network works upon the combination of input, hidden and output layers (Fig. 7). The matrix used to construct the nodes between the input layer and hidden layers is defined as an associative matrix, while the decision matrix is used to obtain the nodes between the hidden and output layers. The layers are processed by the weight of each value, obtained from each node. This study uses the backpropagation method for minimisation the error (Fig. 8b). The activation function 'Relu' was used with 3 hidden layers consisting of 32, 32 and 16 nodes[49].

Accuracy Assessment

The LSM prepared using RF, MLP and TL techniques were evaluated using statistical measures such as AUC-ROC, precision, recall, F-score and accuracy. The remaining 30% of the datasets of the source and target areas were used for accuracy assessment. Statistical measures are commonly divided into cut-off independent and cut-off dependent, where a cutoff is defined to superset the classes of table from an unstable one (landslide and non-landslide) in this case. AUC-ROC has been widely used to evaluate the predictive capability of the landslide models. The area under the curve is obtained by plotting a graph between the specificity (Eq. 4) and 1-sensitivity (Eq. 5), while the precision is calculated based on the ability to detect positive predictions (Eq. 6), recall calculates the positive predictions (Eq. 7). The F-score represents the harmonic mean between the precision and recall (Eq. 8). Other measures such as accuracy are the ratio of overall correct predictions from all the predictions (Eq. 9).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{9}$$

Consistency Assessment

Machine learning models are prone to overestimation due to class imbalance, hyperparameter tuning, and complexity of the models[50]. Thus, it is required to assess the consistency of the result obtained. The seed Cell Area Index (SCAI) proposed by Suzen and Doyuran was used to evaluate the consistency of the results, SCAI is the ratio between the area of the susceptibility class and the number of landslide percentages occurring in that susceptibility class[51] (Eq. 10).

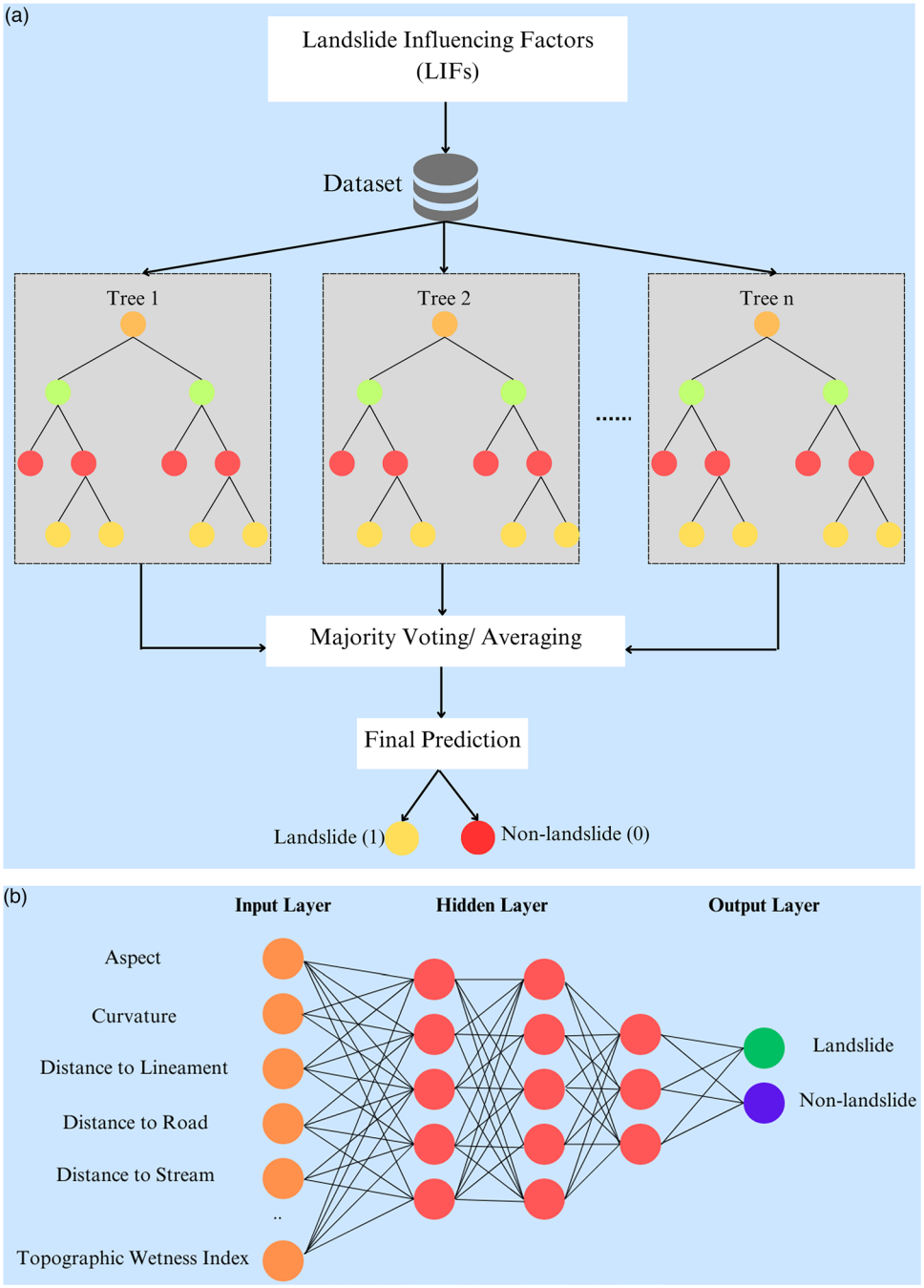


Fig. 8. **a** Flow chart of random forest (RF). **b** Flow chart of multi-layer perceptron (MLP).

$$SCAI = \frac{\text{Percentage area Susceptibility classes}}{\text{Number of landslide in the Susceptibility Classes}} \tag{10}$$

The value of SCAI ranges from 0 to infinity, a consistent model has a gradually decreasing value of SCAI from very low to very high susceptible zone.

Results

Independence and Importance of LCFs

The independence of the LCFs is essential for effective model performance. Multicollinearity was checked by calculating the variance inflation factor (VIF) and tolerance for the LCFs. The results showed that the considered factors had no such issue of multicollinearity for both the source and target areas. In the source area, plane curvature obtained the highest VIF value of 1, in target areas 1 and 2, relief had the highest VIF value of

Variable	Source (Mandi)	Target (Kullu)	Target (Rudraprayag)
Aspect	0.23	1.02	1.02
Curvature	0.01	1.14	1.19
Distance to Lineament	0.43	1.29	1.49
Rock Type	0.45	1.04	1.02
Plane Curvature	1.03	1.46	1.76
Relief	0.22	2.61	1.67
Distance to Road	0.51	1.38	2.34
Slope	0.16	1.91	1.57
Surface Rougness Count	0.48	1.34	1.15
Distance to River	1.00	1.20	1.39
Topographic Wetness Index	0.12	1.36	2.96

Table 1. Multi-collinearity test for factor independence.

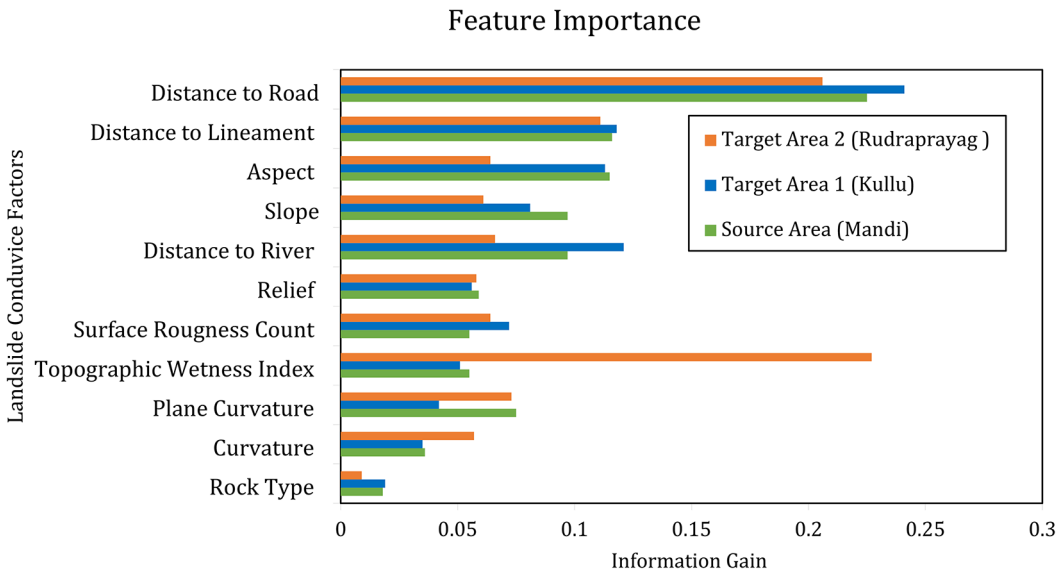


Fig. 9. Feature importance plot showing the relative importance of different features.

2.61, and the topographic wetness index had highest VIF value (2.96) (Table 1). However, the lower values of VIF suggested no such issue of multicollinearity between the dependent and independent variable.

Machine learning and deep learning methods manifest the non-linearity between the variables; however, they are black boxes in nature and do not provide any insight related to the participation of individual factor in final prediction. Thus, feature importance of LCFs was calculated using the information gain ratio (Fig. 9). The IG value obtained for the source area (Mandi) demonstrated the higher contribution of factors such as distance to road (0.225), distance to lineament (0.116), aspect (0.115) and slope (0.097), depicting the occurrence of landslides along the road. Similarly, for distance to road, lineament aspect and slope were among the highest contributors in final prediction for target area 1. Contrastingly, in the case of target area 2, the topographic wetness index was highest contributor in landslides followed by other factors such as distance to road, distance to lineament and plane curvature (Fig. 9).

KL divergence

The Kullback–Leibler (KL) divergence measure was used to estimate the degree of entropy and similarity between the variables of the source and target areas. Although there is no upper bound to the value of KL divergence, it is important to note that a higher value of KL divergence corresponds to the least similarity between two distributions. In this study, KL divergence was calculated between the LCFs of the source and target areas (Table 2). Lithology showed the lowest KL value of 1, with the highest being 2.74 for distance to road, followed by relief (2.66), distance to stream (2.57), and distance to lineament (2.52) (Table 3). Thus, the results depicted the presence of similarity, if not exact, between the two areas in terms of topography, hydrology, and geology (Table 2).

S.no	Landslide conditioning factors	KL divergence source and target area 1	KL divergence (source and target area 2)
1	Aspect	1.63	0.47
2	Curvature	1.66	0.19
3	Distance to Lineament	2.52	0.43
4	Rock Type	1	0.52
5	Plane Curvature	1.65	0.32
6	Relief	2.66	0.26
7	Distance to Road	2.74	0.57
8	Slope	1.64	0.19
9	Surface Roughness Count	1.67	0.73
10	Distance to Stream	2.5	0.41
11	Topographic Wetness Index	1.61	0.62

Table 2. Table showing the KL divergence for LCFs for source and target area.

(a)	Random forest						
	Source	Target 1	Target 1 (TL)	Target TL 1 (Combined)	Target 2	Target 2 (TL)	Target TL 2 (Combined)
Mean	0.39	0.31	0.49	0.41	0.36	0.44	0.35
Max	1	1	1	1	1	1	0.99
Min	0	0	0	0	0	0	0
S.D	0.25	0.26	0.24	0.26	0.14	0.24	0.14
(b)	Multi layer perceptron						
	Source	Target 1	Target 1 (TL)	Target TL 1 (Combined)	Target 2	Target 2 (TL)	Target TL 2 (Combined)
Mean	0.37	0.28	0.46	0.28	0.349	0.44	0.349
Max	1	1	1	1	1	1	1
Min	0	0	0	0	0	0	0
S.D	0.32	0.32	0.34	0.28	0.26	0.342	0.262

Table 3. Statistical details for LSM obtained using random forest and multi-layer perceptron method.

Random forest

The LSMs obtained using random forest (RF), were classified into five classes using the natural break classification method. The minimum, maximum, mean, and standard deviation are shown in the table (Table 3a). In LSM obtained for source area, 12% of the area covered in very high susceptibility, high susceptibility covered 16% area, while low and very low covered 51% area (Fig. 10a). In LSM of target area 1, trained from the data itself, very high susceptibility covered 10% area, high and medium susceptibility covered 28%, and 62% area was classified into very low and low susceptibility classes (Figs. 10a, 11a, 12a). Furthermore, when the source-trained model was applied on target area 1, very high susceptibility covered 16% of the area, 21% with high susceptibility, 22% with medium susceptibility and 40% area with very low and low susceptibility (Figs. 11b, 12ac). An increase of 3% area was observed in very high susceptibility, 4% in high susceptibility followed by a 2% increase in medium susceptibility. The increase in the areas of classes can be attributed to the scarcity of the landslide data in these areas which contrastingly present in the source trained model. Finally, when the landslide information from the source as well as target was used to prepare LSM of target area, 13% area was classified as very high, high 14%, 19% as medium and 52% as very low and low susceptibility (Figs. 11c, 12a). A significant decrease was observed when combined trained model were used, which was also concordant with the percentage values of LSM of the target area trained from the data itself. Furthermore, for target area 2, when random forest was trained, 5% of the area covered very high susceptibility, 15% area covered high susceptibility, 25% medium susceptibility (Figs. 12ba, 13ac). When random forest model was trained from the source area and applied to target area 2, very high susceptibility covered 7%, 18% in high susceptibility, 22% with medium susceptibility and very low and low susceptibility covered 51% of the total study (Figs. 12b and 13b). Finally, when the random forest model was trained using the data from the source as well as the target area , very high susceptibility covered 6% of the area, high and medium susceptibility covered 42% of the area while 49% area was classified as very low and low susceptibility (Figs. 12b, 13c). Notably, in source as well as the target area 1, the very high and high susceptibility were mainly concentrated in the proximity of roads as well as the rivers. Whereas in target area 2, high susceptibility was mainly concentrated along the river Mandakini, depicting the greater influence of river and topographic wetness index on landslide causation. Evidently, the very high and high susceptibility zones from the RF-TL and RF-Combined were nearly same as those obtained by LSM of target areas (Figs. 12b, 13c).

Multi-layer perceptron

Similarly, the LSMs obtained using the multi-layer perceptron (MLP) were classified into very less, less, medium, high and very high susceptibility classes based on natural break classification method (Table 3b). The LSM

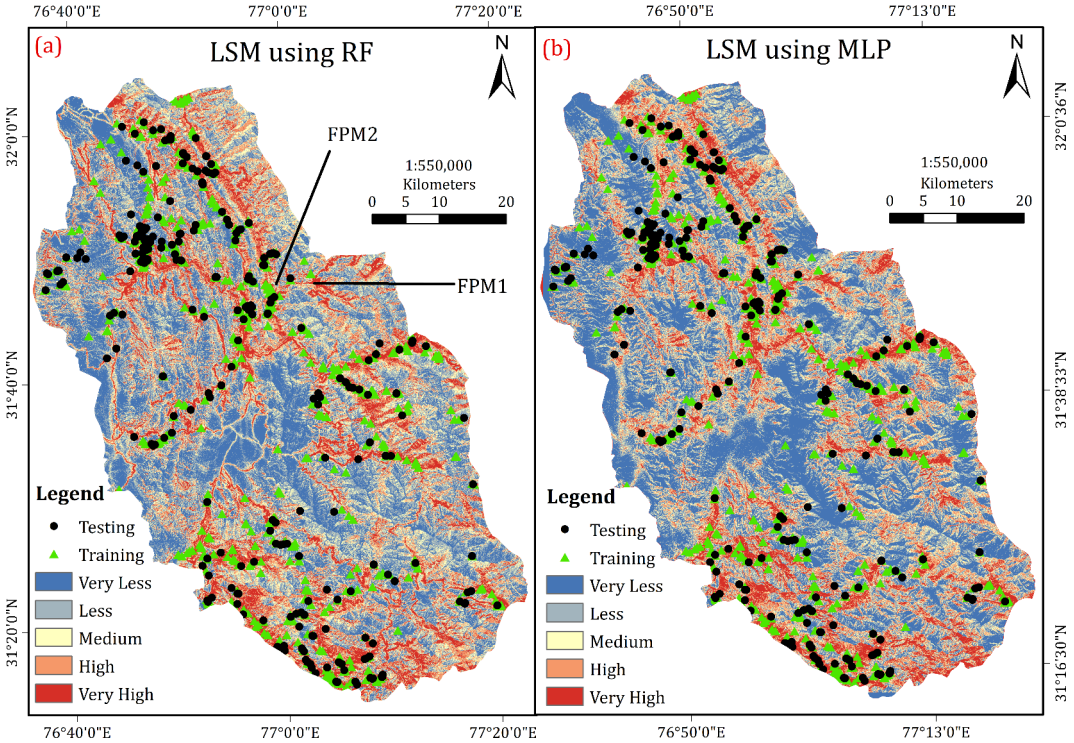


Fig. 10. Landslide susceptibility map of Source (Mandi) **a** Random Forest and **b** Multi-Layer Perceptron (MLP). The details of Field Photograph Mandi 1 (FPM1), FPM2 and FPM3 are provided in Fig. 15.

obtained for the source area, classified 17% area as very high, 13% into high and medium susceptibility, and 39% area as very low susceptibility (Figs. 10b, 12a). When the LSM was prepared for the target area 1 using the training data of itself, very high susceptibility covered 14% of the area while 9% area in high susceptibility, 11%, 16% and 48% area covered medium, high, and very high susceptibility (Figs. 11d, 12a). Further, LSM of target area 1 from the source-trained model, classified 29% area in very high and high susceptibility, while medium and very low and low susceptibility covered 14% and 46% of the target area respectively (Figs. 11e, 12a). Finally, the data from the source and target area were used to prepare the LSM of the target area, very high and high susceptibility constituted 21% of the area, medium covered 13%, low 18% and very low 47% of the target area (Figs. 11f, 12a). Additionally, in the case of target area 2, when LSM was prepared using the training data itself, 13% covered very high susceptibility, followed by 17% high susceptibility and 49% under very low and low susceptibility. When the source-trained model was applied in the area, 27% area classified as very high susceptibility, 15% were in high susceptibility, followed by 14% for medium susceptibility, and 16% and 26% for low and very low susceptible respectively. Also, when the LSM was prepared using the data trained from source and target area 2, 10%, 14%, and 12% area were classified into very high, high and medium susceptibility respectively, while 16% and 46% were classified as low and very low susceptibility. The results from the LSM indicated that when the source source-trained model was transferred in target area 2, the result tended to overestimate the very high and high susceptibility classes, while underestimating the low susceptibility regions (Figs. 12b, 13d–f).

Accuracy assessment and validation

The results from the LSM are utilized by planners and policymakers for future developmental strategies. For this reason, it is imperative to verify the results to ensure their reliability and credibility. Statistical tests, such as AUC-ROC, precision, recall, F-score, and accuracy, were employed to assess the performance of the LSM on unknown datasets during testing. The AUC value for the source area revealed an AUC value of 0.922 for RF and 0.917 for MLP, with precision, recall, F-score, and accuracy values of 0.856 each for RF (Fig. 14a), while MLP obtained a precision value of 0.832, recall of 0.856, F-score of 0.842, and accuracy of 0.844 (Fig. 14b). In the case of LSM prepared for target area 1 using the training data itself, obtained AUC values of 0.908 and 0.896 for RF and MLP, respectively. Precision values of 0.898 were obtained for both RF and MLP, with recall values of 0.851 for RF and 0.879 for MLP. F-score gave values of 0.874 for RF (Fig. 14a) and 0.888 for MLP, and accuracy values obtained were 0.87 and 0.844 for each model (Fig. 14b). Furthermore, when the source-trained model was applied to target area 1, results indicated an increase in the AUC value (0.942), precision (0.955), and F-score (0.876), was observed following up with a decrease in values such as recall (0.809) and accuracy (0.865). Lastly, for target area 1, when the knowledge from the source as well as the target, was used in LSM preparation, the AUC obtained was 0.959 for RF and 0.946 for MLP, whereas other statistical measures such as a precision value of 0.921 for RF and MLP, recall 0.82 and 0.901 for RF and MLP respectively (Fig. 14a, b). A slight decrease in the F-score was observed as 0.876 for RF, while an increase was observed for MLP as 0.911. Similarly, with in the case

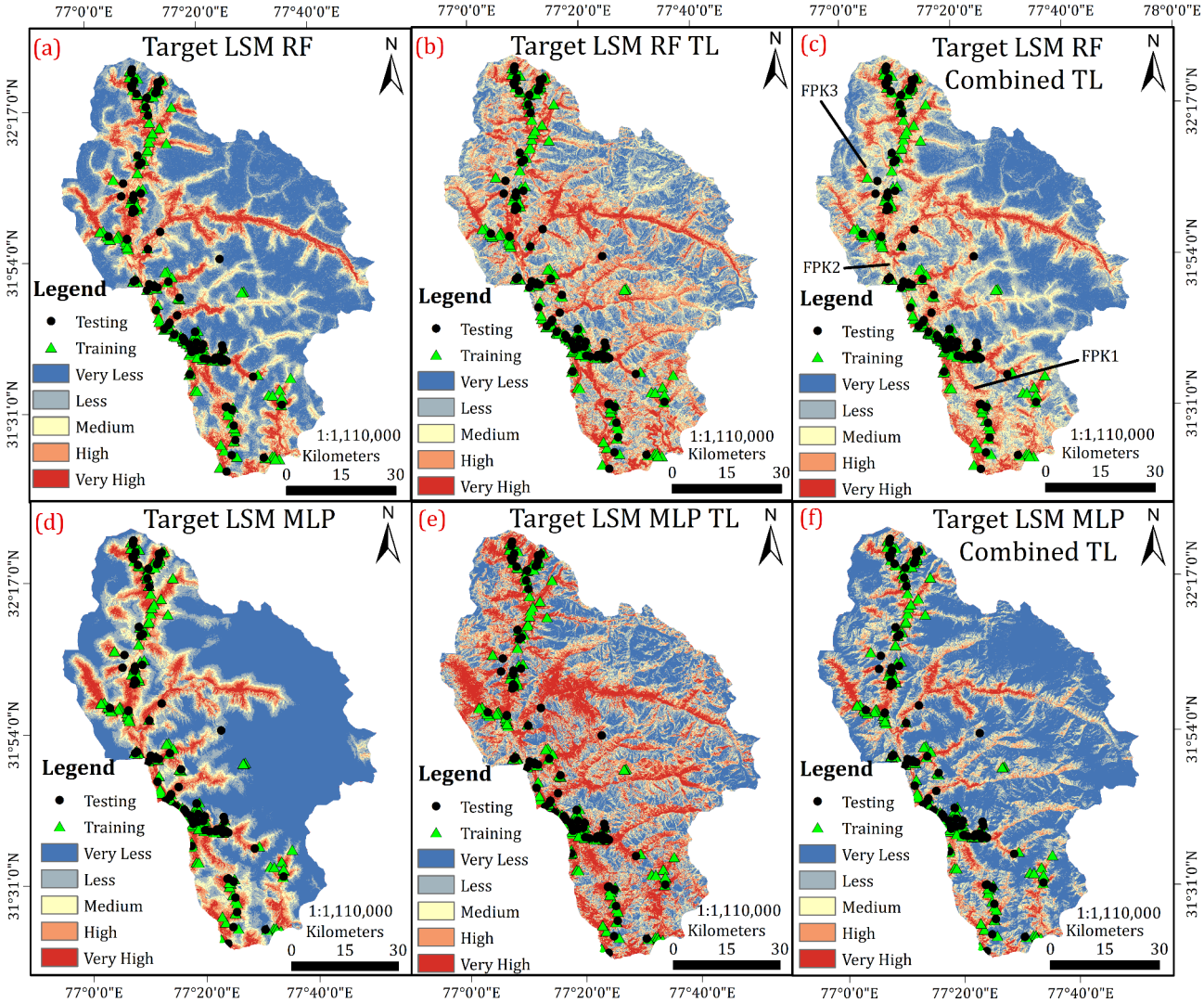


Fig. 11. Landslide susceptibility map of Target area 1 (Kullu) **a** Random Forest (RF), **b** Multi-Layer Perceptron (MLP), **c** RF-Transfer Learning (TL), **d** MLP-TL, **e** RF Combined TL, **f** MLP-Combined TL. The details of Field Photograph Kullu 1 (FPK1), FPK2 and FPK3 are provided in Fig. 15.

of accuracy as for RF accuracy was 0.859 and 0.91 for MLP (Tables 4a, b).Further to test the applicability of the proposed TL approach, the knowledge learned from the source were was used to in other areas. The target area 2 was Rudraprayag district located in the Uttarakhand Himalayas, with physiographic and geological conditions totally different form from that of the source area. The results depicted that when RF and MLP models were trained using the training data itself, RF obtained an AUC value of 0.95 (Fig. 14c)and 0.82 for MLP (Fig. 14d), while in the case of precision RF obtained 0.80 while MLP gave 0.84. Other statistical measures obtained recall values 0.75 and 0.84 for RF and MLP, F-score values 0.79 and 0.82, accuracy 0.82 each. Moving ahead when the source source-trained model was used in LSM preparation for target area 2, the AUC value was 0.80 for RF and 0.75 in the case of MLP., while the precision value for RF and MLP was 0.80 and 0.65, recall 0.89 and 0.71, F-score 0.76 and 0.68, and accuracy 0.72 and 0.66 for RF and MLP respectively (Fig. 14c, d). Moreover, when the knowledge from the source as well as the target area 2 was used, an increase in the AUC value was observed with 0.98 for RF and 0.84 in the case of MLP (Fig. 14c,d). Other statistical measure such as precision provided 0.88 and 0.81 for RF and MLP, whereas in the case of recall value for RF was 0.86 and 0.70 for MLP. While on the other hand, the F-score and accuracy value were 0.87 and 0.75 for RF and MLP each (Table 4a, b). The above result suggested that when the source trained was transferred and supplemented with the existing datasets, the model performance increases when compared with the LSM prepared from the training data itself. Contrastingly, when LSM was prepared for both the target area using the knowledge from the source-trained model, a decrease in the statistical measures was observed in both the case depicting differing geological, topographical, and physiographical conditions between the source and two target areas.

Additionally to validate the performance of the proposed TL we conducted field survey for landslides that were not considered in the model training and testing purposes for Mandi source domain (Fig. 15a and b), target area 1 (Kullu) (Fig. 15c–e) and target area 2 (Rudraprayag) (Fig. 15f–h).

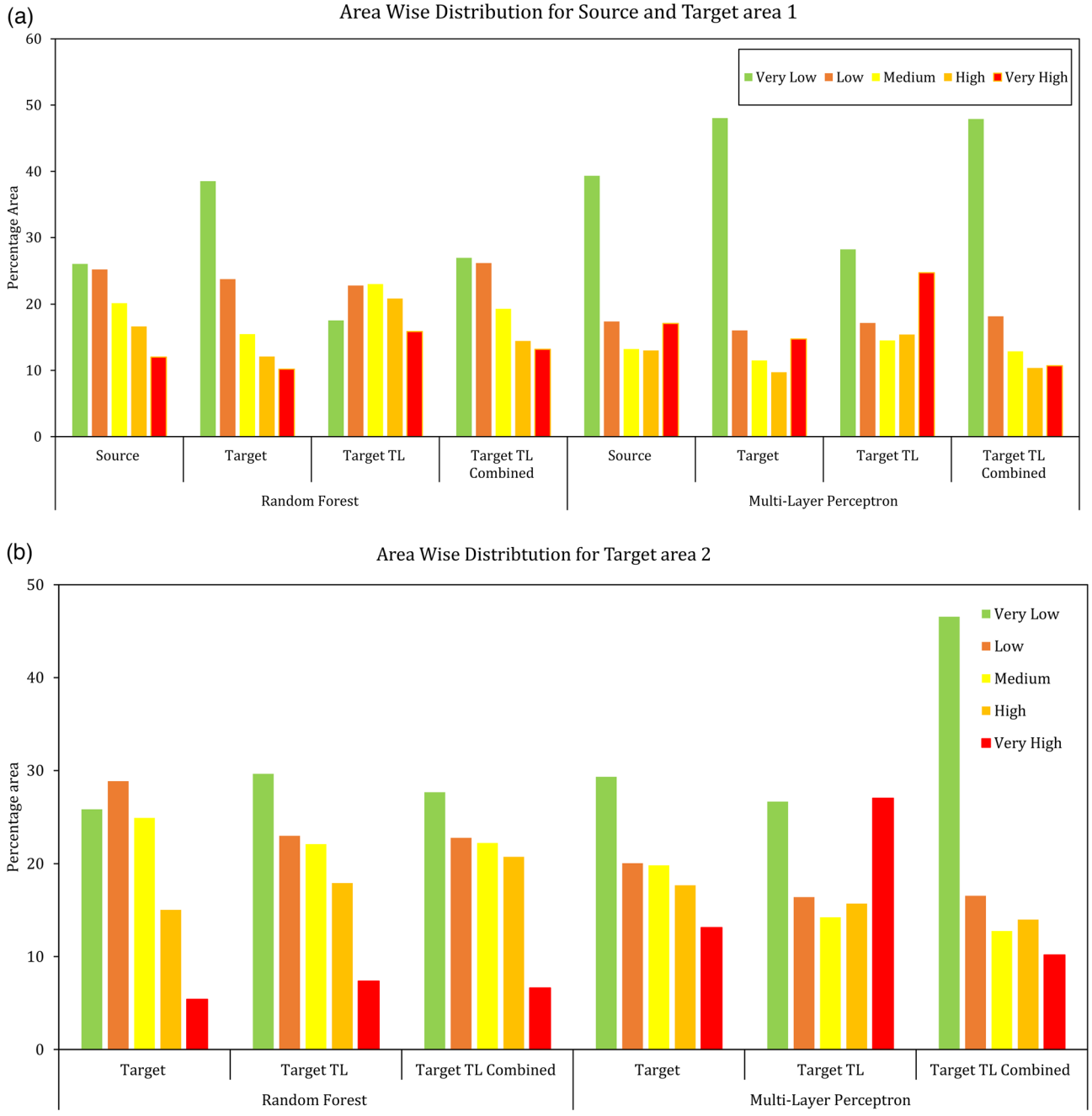


Fig. 12. Area wise distribution of **a** source and target area 1, **b** target area 2.

Consistency assessment

Machine learning models being a data driven method are prone to data overestimation of susceptibility classes; thus, it is important to critically evaluate the results obtained from these models. Seed cell area index (SCAI) was used in this study to evaluate the consistency of the model. The SCAI value generated for the source area, the SCAI value ranged from 6.91 to 0.20, with very high covering 12% of the area and consisting of 60% landslides, while the SCAI value for target area 1, ranged from 34.27 to 0.14 (Fig. 16a). Further, when the source data was used, the SCAI value ranged from 15.61 to 0.20, a slight increase in the SCAI value for low susceptibility indicated slight overestimation in this class. However, the issue of overestimation was successfully removed when both the source data and the target were used to prepare the LSM, where the SCAI value ranged from infinite to 0.17 (Fig. 16b, c). An infinite SCAI value for the very low class in this case was attributed to zero landslide points falling the susceptibility class. Moreover, in the case of target area 2, the SCAI value decreased from very low susceptibility classes to very high susceptibility. Thus, the model gave a consistent result, classifying a greater percentage of low and very low susceptibility classes, and the least area in high and very high susceptibility classes, demonstrating the reliability of the LSM results.

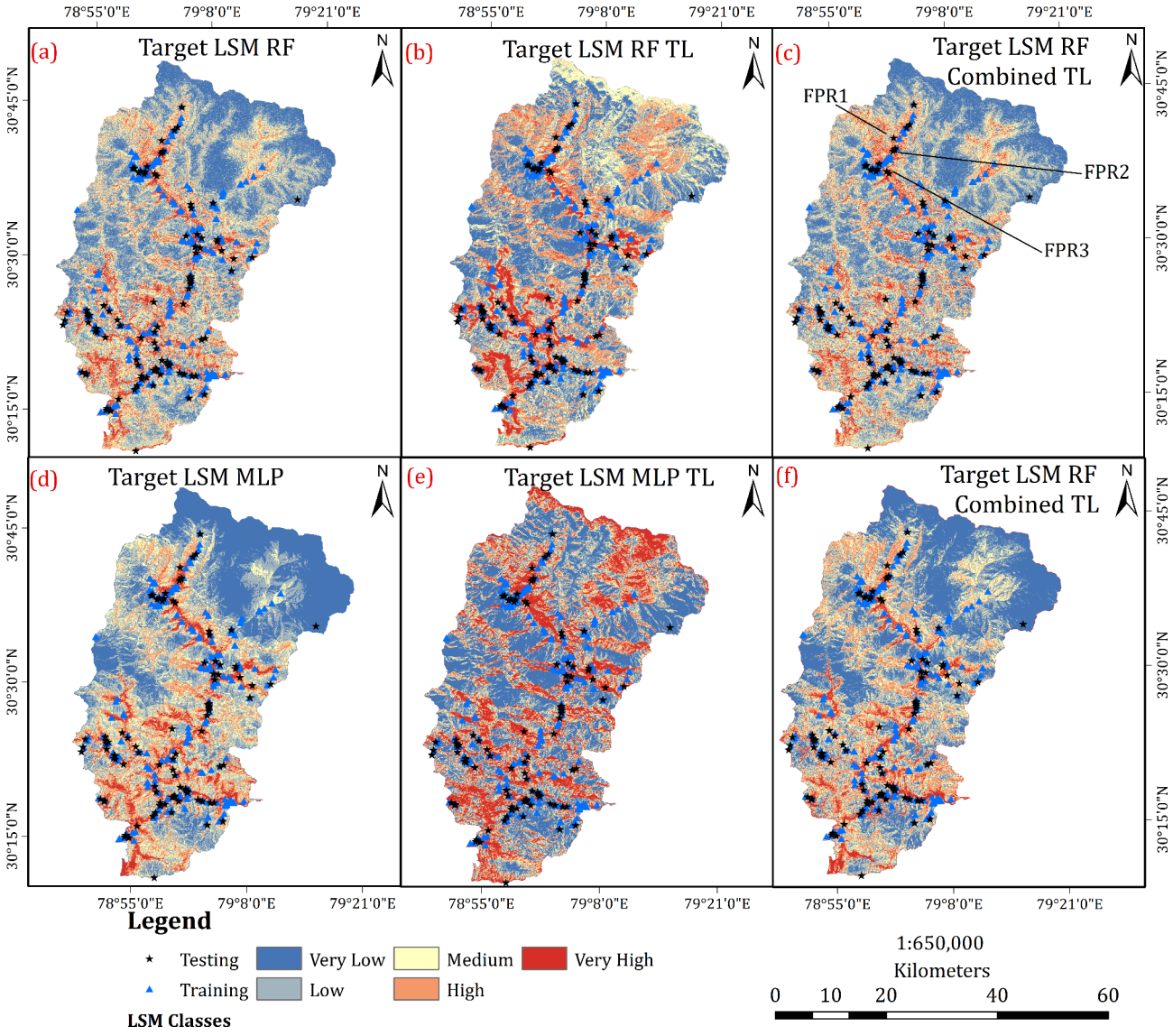


Fig. 13. Landslide susceptibility map of Target area 2 (Rudraprayag) **a** Random Forest (RF), **b** Multi-Layer Perceptron (MLP), **c** RF-Transfer Learning (TL), **d** MLP-TL, **e** RF Combined TL, **f** MLP-Combined TL. The details of Field Photograph Rudraprayag 1 (FPR1), FPR2 and FPR3 are provided in Fig. 15.

Discussion

Landslide causes serious damage to the society and economy of the region. In such conditions, LSM preparation plays a vital role in mitigating loss from landslides, by spatially portraying the location of future landslides in the area. Landslide susceptibility mapping (LSM) has gained meaningful insight in relation to hazard management and risk reduction. Conventional LSM preparation have used landslide dataset from the region itself to train either a machine learning or a deep learning models. Machine learning methods includes random forest (RF), support vector machine (SVM) and logistic regression (LR), while multi-layer perceptron (MLP), convolutional neural network (CNN), long short-term memory (LSTM), and recurrent neural network (RNN) form the deep learning models. Due to their better prediction capability, previous literature in LSM preparation have used these methods extensively. In Yang et al.[52], landslide susceptibility was assessed using various ratio of landslide to non-landslide samples. Three machine learning models such as support vector machine (SVM), random forest (RF) and gradient boost decision tree (GBDT) were used. The results depicted better performance of RF over other algorithms in addressing the sample unbalance issue. Yang et al. 2024 [53] assessed the landslide susceptibility of Wanzhou district, China. 5 factors and landslide point, and polygon inventory were used in random forest model. The results showed that while the prediction performance of the polygon was higher, the random forest obtained an acceptable accuracy of 0.85 in it. Similarly, in another study susceptibility of Chongren and Jiangxi using the random forest (RF), support vector machine (SVM) and logistic regression (LR) manifested that, the RF algorithm performed best among the observed algorithms. However, in studies by Lee et al. 2018[54],and Gupta and Shukla 2023[36] reported greater capability of SVM. Further, studies such as Adnan et al. 2020[55] used

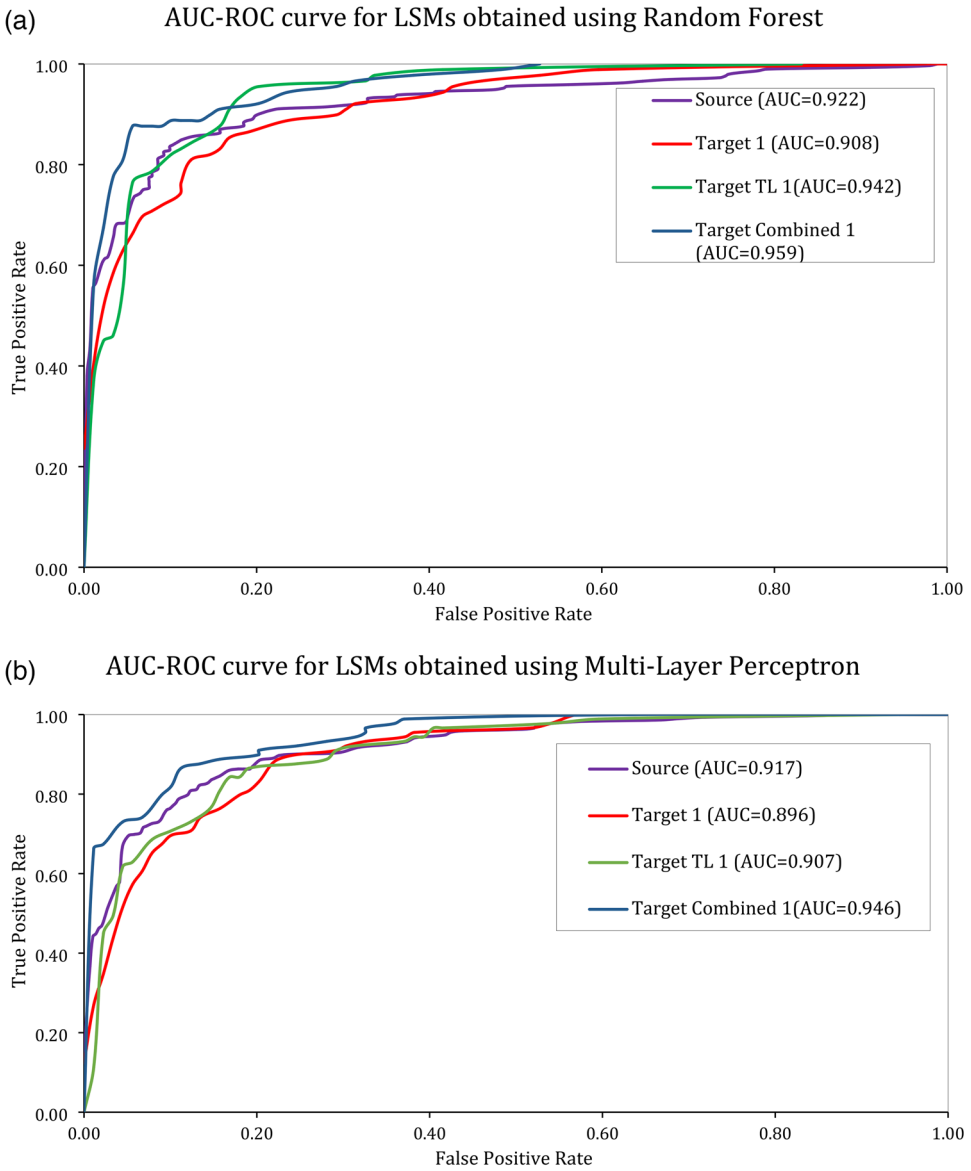


Fig. 14. AUC-ROC curve obtained **a** source and target area 1 using RF, **b** source and target area 1 using MLP, **c** target area 2 using RF and **d** target area 2 using MLP.

(a)	Random forest						
	Source	Target 1	Target 1 (TL)	Target 1 (Combined)	Target 2	Target 2 (TL)	Target TL 2 (Combined)
AUC-ROC	0.922	0.908	0.942	0.959	0.95	0.8	0.98
Precision	0.856	0.898	0.955	0.921	0.8	0.67	0.88
Recall	0.856	0.851	0.809	0.82	0.84	0.89	0.86
F-score	0.856	0.874	0.876	0.867	0.82	0.72	0.87
Accuracy	0.856	0.87	0.865	0.859	0.82	0.8	0.98
(b)	Multi layer perceptron						
	Source	Target 1	Target 1 (TL)	Target 1 (Combined)	Target 2	Target 2 (TL)	Target TL 2 (Combined)
AUC-ROC	0.917	0.896	0.907	0.946	0.82	0.8	0.84
Precision	0.832	0.898	0.9	0.921	0.84	0.65	0.81
Recall	0.856	0.879	0.89	0.901	0.75	0.71	0.7
F-score	0.842	0.888	0.895	0.911	0.79	0.68	0.75
Accuracy	0.844	0.844	0.894	0.91	0.82	0.66	0.75

Table 4. Analysis results of statistical measures precision, recall, F-score and accuracy for models.



Fig. 15. Field photograph of Mandi (source) **a** FPM1 (31° 47' 4" N, 77° 4' 26" E), **b** FPM2 (31° 46' 48" N, 76° 59' 13" E), Kullu (Target area 1), **c** FPK1 (31° 37' 53" N, 77° 25' 47" E), **d** FPK2 (31° 54' 77" N, 77° 07' 44" E), **e** FPK3 (32° 10' 45" N, 77° 7' 26" E) and Rudraprayag (Target area 2), **f** FPR1 (30° 34' 42" N, 79° 2' 39" E), **g** (30° 36' 55" N, 79° 0' 55" E), and **h** (30° 37' 50" N, 78° 59' 56" E).

four machine learning methods: K-nearest neighbour (KNN), RF, MLP, and SVM. The results showed that the performance of RF was best followed by MLP, SVM etc. In the same manner, studies such as[56,57], described the better performance of MLP in susceptibility mapping.

LSM preparation, in recent years is largely based upon data-driven models, where the quality and quantity of the data is important. Data collection is the most tedious task to perform especially in remote mountainous location. . Therefore, previous studies have overcome the data limitation by adopting TL method, in which the knowledge from the data-rich region (source area) is transferred to the data-scarce region via process defined as transfer learning (TL). Zhiyong et al. 2023 [30] applied Traboost TL models for Wenchaun-Yingxiu China as source area while the target area was Zingui basin. The results revealed that when the landslide inventory of the source areas was used for LSM in target area, the AUC value of decision tree (DT), support vector machine (SVM) and random forest (RF) increased by 0.07, 0.03 and 0.02 respectively. Wang et al. 2023[12] employed TL approach coupled with generalized additive models for 10 study areas using various resolutions (1, 10, and 25,) located in Austria, Ecuador and Italy. The results concluded that case-based reasoning performed well compared to domain adaption. Moreover, the authors used a similarity-based criterion for TL approach. Zhang et al. 2023[18], proposed a geo-environmental aware susceptibility map based on unsupervised adversial transfer learning. The results depicted an increase of AUC value by 7–12% compared to the traditional machine learning results. The results also revealed that TL approach effectively transfer knowledge from other regions to sample free regions. Moreover, TL approach was applied in Hong Kong region where deep learning models were used[13]. The study reported 30% increase of accuracies on the test set while 62% reduction in the logirthmic losses. Similarly,

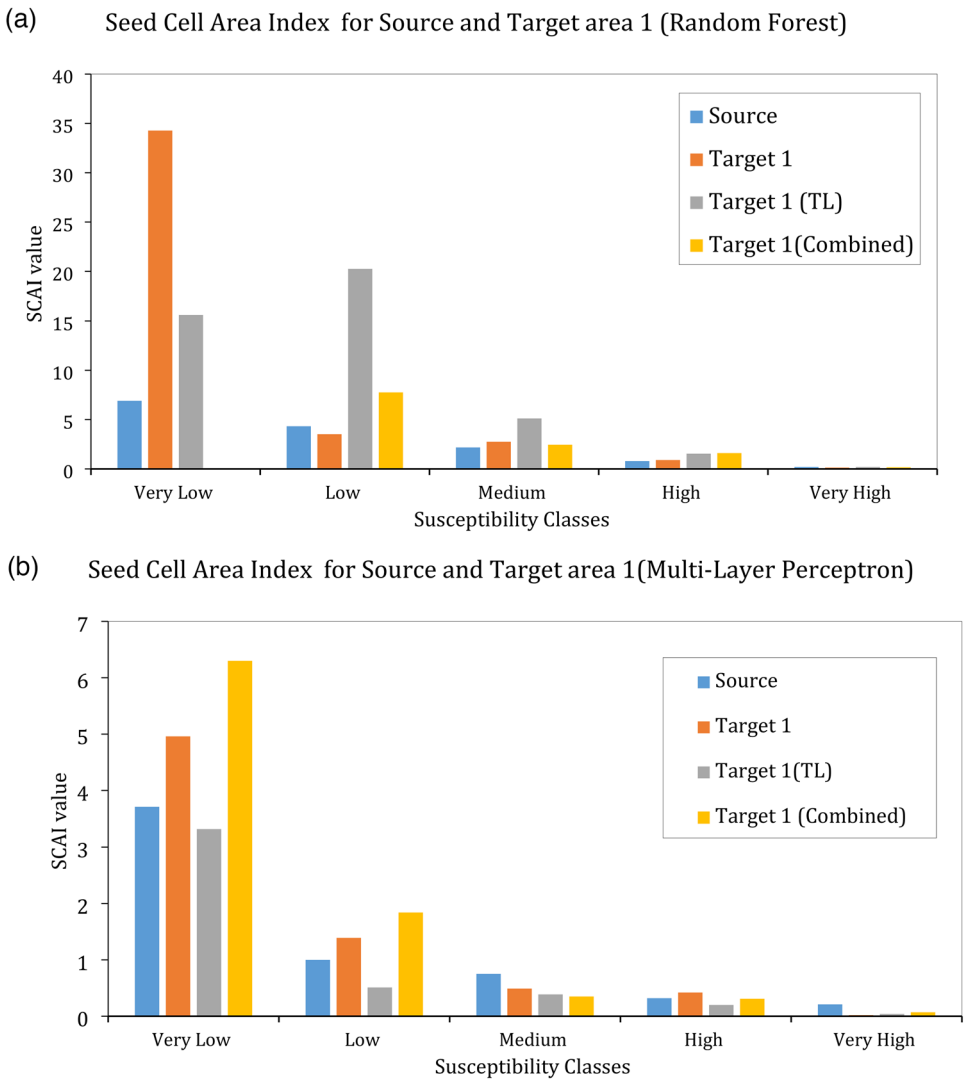


Fig. 16. SCAI values for each susceptibility zones of landslide susceptibility mapping **a** source and target area 1 using RF, **b** using MLP, **c** for target area 2 using RF and MLP.

the landslide inventory of Wenchuan and Jiuzhaigou earthquake were considered as source and target areas respectively to apply the TL approach (Ai et al. 2023). The Jiuzhaigou susceptibility map was prepared using the pre-trained model. The results of AUC-ROC value revealed that the TL approach model effectively portrayed the landslide area. Finally in Lin et al.[58] delved into the complexity of landslide data preparation by applying transfer learning attributes (TLA) based on auto encoder of generative adversarial networks. The study employed 14 landslide influencing factors and 986 landslide data to model the landslide susceptibility. The results revealed that the proposed TLA strategy performed well by increasing the AUC-ROC value by 2–7%.

In view of the previous studies, the TL approach specifically instance transfer performs better in transferring knowledge from source to target area. However, the applicability of other TL approaches such as parameter transfer, feature-representation transfer and knowledge transfer have not been used in relation to LSM preparation and as per our knowledge no attempts have been made in evaluating the applicability of the TL approach in the Himalayan region.

Himalayan mountains especially the northwestern part is experiencing an increase in the frequency of landslides due to changing climate patterns and urbanization. Hence with such an increase in the landslides, precise and timely delineation of susceptible areas is important, which is hampered significantly due to limited data availability.

Thus, this study uses TL approach in LSM preparation where the source area or the data-rich region is Mandi while two the data data-scarce region (target area) areas are Kullu and Rudraprayag districts. 12 landslide conducive factors (LCFs) and machine learning models such as random forest (RF) and multi-layer perceptron (MLP) were used in the preparation of LSM for source and target area.

Variance inflation factor was used to check the independence of each LCF, the VIF value indicated the independence of LCFs, while the information gain (IG) revealed the contribution of each LCF. Distance to road

and distance to lineament were among the most contributors in LSM in source area, while distance to road and river contributed highest in the target area 1. Further, the similarity between the source and target area was established by KL divergence, the KL value < 2 showed the extent of similarity between the source and two target areas. RF-based LSM of the source area depicted that 12% of the area covered very high susceptibility. LSM of the target area 1 trained by the data itself revealed that very high susceptibility covered 10% of the area, which was increased to 15% when the TL approach was applied, while it reduced to 13% when combined TL approach was applied. Similarly, MLP prepared LSM, classified 17% of the source area. Very high susceptibility while 14% in target area. The TL-based LSM classified 24% of area in very high susceptibility which was 10% higher than the previous one. However, a reduction to 10% in very high susceptibility areas was observed in the combined TL approach.

Furthermore, in target area 2, RF-based LSM classified 6% of the very high susceptibility, TL-based LSM covered 7% of the area while 6% area covered in very high susceptibility when the knowledge from the source as well as target area were used. Contrastingly, LSM using MLP was classified as 13% which increased by 15% when the TL from the source area was used, however, a decrease to 10% was observed when LSM was prepared using source as well as target area.

Thus, it is evident that TL approach overestimated the region of very high susceptibility while the combined TL approach classified less but near a similar area to the target-trained model. Further, the accuracy assessment using AUC-ROC, precision, recall, F-score and accuracy revealed that for RF an increased AUC value was observed in TL approach (0.034) and combined TL approach (0.051), while a decrease in a few decimals was observed in recall, F-score, and accuracy. Additionally, in the case of MLP, all the measures increased for TL and combined TL approach.

Each LS model has its highlights and challenges. The landslide inventory preparation involves detailed field investigation and time time-consuming, the parameter transfer approach used in this study highlights the applicability of knowledge obtained from the data-rich region to data scarce region. While on the other hand, the proposed TL approach has certain challenges especially when the target areas are dis-similar geomorphologically. Thus, the results highlight that the TL is effective in the Himalayan region. This highlights that although the TL approach can predict the landslide in the target area, such a trained model may overestimate the high and very high susceptibility classes. Consequently, when applying the TL approach, where landslide data is imperative the use of a combined TL approach is recommended.

Conclusion

In this study, TL approach was used to overcome the issue of data limitation. The study used Mandi district as the source and Kullu and Rudraprayag district as target areas . The primary objective was to analyse the applicability of TL where the knowledge obtained from the source area is applied to the target area. The result obtained showed that random forest (RF) and multi-layer perceptron (MLP) performed better when data was transferred from the source area. Furthermore, when the data from source as well as target area was used for LSM preparation for target area, the model performed exceptionally well portraying an increase in the AUC value of 0.051 for RF and 0.039 for MLP. However, a slight decrease in the value of recall, F-score and accuracy was observed in RF, while all the statistical measures were highest for the model trained using data from source as well as target area in MLP. Hence, the results provide a comprehensive analysis of the applicability of the TL approach in the LSM preparation particularly where the data collection is difficult and costly. Overall, the findings highlight the applicability of TL to significantly improve the predictive capability of LSM in data-scarce mountainous terrains, thereby contributing to more informed land use planning and effective risk mitigation strategies.

Data availability

Data and the codes used will be made available by the corresponding author upon reasonable request.

Received: 6 August 2024; Accepted: 15 October 2024
Published online: 23 November 2024

References

1. Varnes, D. J. Landslide hazard zonation: a review of principles and practice. *Nat. Hazards*. (3) (1984).
2. Nath, S. K., Sengupta, A. & Srivastava, A. Remote sensing GIS-based landslide susceptibility & risk modeling in Darjeeling-Sikkim Himalaya together with FEM-based slope stability analysis of the terrain. *Nat. Hazards* **108**(3), 3271–3304. https://doi.org/10.1007/s11069-021-04823-5 (2021).
3. Bhagya, S. B. et al. Landslide susceptibility assessment of a part of the Western Ghats (India) employing the AHP and F-AHP models and comparison with existing susceptibility maps. *Land* **12**(2), 468. https://doi.org/10.3390/land12020468 (2023).
4. Nwazelibe, V. E., Unigwe, C. O. & Egbueri, J. C. Testing the performances of different fuzzy overlay methods in GIS-based landslide susceptibility mapping of Udi Province, SE Nigeria. *CATENA* **220**, 106654. https://doi.org/10.1016/j.catena.2022.106654 (2023).
5. Gupta, S., Shukla, D. P. & Thakur, M. Selection of weightages for causative factors used in preparation of landslide susceptibility zonation (LSZ). *Geomatics Nat. Hazards Risk* **9**(1), 471–487. https://doi.org/10.1080/19475705.2018.1447027 (2018).
6. Michael, E. A., & Samanta, S. Landslide vulnerability mapping (LVM) using weighted linear combination (WLC) model through remote sensing and GIS techniques. *Modeling Earth Systems and Environment*, **2**, 1–15. (2016).
7. Singh, A., Chhetri, N., Nitesh, Gupta, S., & Shukla, D. P. Strategies for sampling pseudo-absences of landslide locations for landslide susceptibility mapping in complex mountainous terrain of Northwest Himalaya. *Bull. Eng. Geol. Environ.* **82**(8). https://doi.org/10.1007/s10064-023-03333-x (2023).
8. Pourghasemi, H. R. & Rahmati O. Prediction of the landslide susceptibility: Which algorithm, which precision?. *Catena*, **162**, 177–192 (2018).
9. Broeckx, J., Vanmaercke, M., Duchateau, R. & Poesen, J. A data-based landslide susceptibility map of Africa. Earth-Science Reviews, 185, 102-121 (2018).

10. Ngo, P. T. T. et al. Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. *Geoscience Frontiers*, **12**(2), 505–519 (2021).
11. Azarafza, M., Akgün, H., Atkinson, P. M., & Derakhshani, R. Deep learning-based landslide susceptibility mapping. *Sci. Reports.* **11**(1). https://doi.org/10.1038/s41598-021-03585-1 (2021).
12. Wang, Z., Goetz, J. & Brenning, A. Transfer learning for landslide susceptibility modeling using domain adaptation and case-based reasoning. *Geosci. Model Develop.* **15**(23), 8765–8784. https://doi.org/10.5194/gmd-15-8765-2022 (2022).
13. Wang, H., Wang, L. & Zhang, L. Transfer learning improves landslide susceptibility assessment. *Gondwana Res.* **123**, 238–254. https://doi.org/10.1016/j.gr.2022.07.008 (2023).
14. Defang, L., Li, J. & Fan, F. Classification of landslides on the southeastern Tibet Plateau based on transfer learning and limited labelled datasets. *Remote Sensing Letters*, **12**(3), 286–295 (2021).
15. Ai, X., Sun, B., & Chen, X. Construction of small sample seismic landslide susceptibility evaluation model based on Transfer Learning: a case study of Jiuzhaigou earthquake. *Bull. Eng. Geol. Environ.* **81**(3). https://doi.org/10.1007/s10064-022-02601-6 (2022).
16. Fu, Z., Li, C. & Yao, W. Landslide susceptibility assessment through TrAdaBoost transfer learning models using two landslide inventories. *CATENA* **222**, 106799. https://doi.org/10.1016/j.catena.2022.106799 (2023).
17. Liu, L., Zhang, Y., Xiao, T., & Yang, C. A frequency ratio–based sampling strategy for landslide susceptibility assessment. *Bull. Eng. Geol. Environ.* **81**(9). https://doi.org/10.1007/s10064-022-02836-3 (2022).
18. Bhuyan, K.et al. Generating multi-temporal landslide inventories through a general deep transfer learning strategy using HR EO data. *Sci. Reports* **13**(1). https://doi.org/10.1038/s41598-022-27352-y (2023).
19. Zhang, L. et al. Geo-environment-aware adversarial transfer learning method for landslide susceptibility evaluation of complex mountainous areas. *Trans. GIS* **27**(5), 1418–1440. https://doi.org/10.1111/tgis.13080 (2023).
20. Singh, N. N., Gupta, S., & Shukla, D. P. Analysis Of Landslide Reactivation Using Satellite Data: A Case Study Of Kotrupi Landslide, Mandi, Himachal Pradesh, India. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-3/W11*, 137–142. https://doi.org/10.5194/isprs-archives-xlii-3-w11-137-2020 (2020).
21. Chakrabarti, B. K. Geology of the Himalayan belt: deformation, metamorphism, stratigraphy (2016).
22. Bhatia, S. B., Bhargava, O. N., Singh, B. P. & Bagi, H. Sequence stratigraphic framework of the Paleogene succession of the Himalayan Foreland Basin: a case study from the Shimla Hills. *J. Palaeontol. Soc. India* **58**(1), 21–38 (2013).
23. Saha, S., Arabameri, A., Blaschke, T. & Bui, D. T. A Novel ensemble approach for landslide susceptibility mapping (LSM) in Darjeeling and Kalimpong Districts, West Bengal, India. *Remote Sens.* **11**(23), 2866. https://doi.org/10.3390/rs11232866 (2019).
24. Powers, P. M., Lillie, R. J. & Yeats, R. S. Structure and shortening of the Kangra and Dehra Dun reentrants, Sub-Himalaya, India. *GSA Bulletin*, **110** (8): 1010–1027. https://doi.org/10.1130/0016-7606(1998)110 (1998).
25. Valdiya, K. S. Later Proterozoic and early Cambrian in the Himalaya. In *Society of Earth Scientists series* (pp. 335–371). https://doi.org/10.1007/978-3-319-25029-8_11 (2015).
26. Najman, Y. et al. Reconstructing the exhumation history of the Lesser Himalaya, NW India, from a multitechnique provenance study of the foreland basin Siwalik Group. *Tectonics.* **28**(5). https://doi.org/10.1029/2009tc002506 (2009).
27. Gupta, S. K., & Shukla, D. P. Effect of scale and mapping unit on landslide susceptibility mapping of Mandakini River Basin, Uttarakhand, India. *Environ. Earth Sci.* **81**(14). https://doi.org/10.1007/s12665-022-10487-6 (2022).
28. 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. https://doi.org/10.1016/j.geomorph.2017.06.013 (2017).
29. Bozinovski, S., & Fulgosi, A. (1976, April). The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In Proceedings of symposium informatica (Vol. 3, pp. 121-126).
30. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evolut.* **12**(9), 1620–1633. https://doi.org/10.1111/2041-210x.13650 (2021).
31. Zhiyong, F., Changdong, L. & Wenmin, Y. Landslide susceptibility assessment through TrAdaBoost transfer learning models using two landslide inventories. *CATENA* **222**, 106799. https://doi.org/10.1016/j.catena.2022.106799 (2023).
32. Lombardo, L., Opitz, T., Ardizzone, F., Guzzetti, F. & Huser, R. Space-time landslide predictive modelling. *Earth-Sci. Rev.* **209**, 103318. https://doi.org/10.1016/j.earscirev.2020.103318 (2020).
33. Kalantar, B., Pradhan, B., Naghibi, S. A., Motevalli, A., & Mansor, S. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). *Geomatics, Natural Hazards and Risk*, **9**(1), 49–69 (2018).
34. Chang, K., Merghadi, A., Yunus, A. P., Pham, B. T., & Dou, J. Evaluating scale effects of topographic variables in landslide susceptibility models using GIS-based machine learning techniques. *Sci. Reports.* **9**(1). https://doi.org/10.1038/s41598-019-48773-2 (2019).
35. Ortiz, J. V. & Martínez-Graña, A. M. Morphometric evaluation and its incidence in the mass movements present in the chicamocha canyon, Colombia. *Sustainability* **15**(2), 1140. https://doi.org/10.3390/su15021140 (2023).
36. Gupta, S. K., & Shukla, D. P. Handling data imbalance in machine learning based landslide susceptibility mapping: a case study of Mandakini River Basin, North-Western Himalayas. *Landslides*, **20**(5), 933–949 (2023).
37. Dou, H., He, J., Huang, S., Wen-Bin, J., & Guo, C. Influences of non-landslide sample selection strategies on landslide susceptibility mapping by machine learning. *Geomatics Nat. Hazards Risk* **14**(1). https://doi.org/10.1080/19475705.2023.2285719 (2023).
38. Hong, H., Wang, D., Zhu, A. & Wang, Y. Landslide susceptibility mapping based on the reliability of landslide and non-landslide data. *Expert Syst. Appl.* **243**, 122933. https://doi.org/10.1016/j.eswa.2023.122933 (2024).
39. Singh, A., Dhiman, N., C, N. K. & Shukla, D. P. Improving ML-based landslide susceptibility using ensemble method for sample selection: a case study of Kangra district in Himachal Pradesh, India. *Environmental Science and Pollution Research*.https://doi.org/10.1007/s11356-024-34726-4 (2024).
40. Rabby, Y. W., Li, Y., & Hilafu, H. An objective absence data sampling method for landslide susceptibility mapping. *Scientific reports*, **13**(1), 1740 (2023)
41. Abraham, M. T., Satyam, N., Lokesh, R., Pradhan, B. & Alamri, A. Factors affecting landslide susceptibility mapping: assessing the influence of different machine learning approaches, sampling strategies and data splitting. *Land* **10**(9), 989. https://doi.org/10.3390/land10090989 (2021).
42. Saha, S. et al. Prediction of landslide susceptibility in Rudraprayag, India using novel ensemble of conditional probability and boosted regression tree-based on cross-validation method. *Sci. Total Environ.* **764**, 142928. https://doi.org/10.1016/j.scitotenv.2020.142928 (2021).
43. Zeng, T. et al. Assessing the imperative of conditioning factor grading in machine learning-based landslide susceptibilitymodeling: a critical inquiry. *CATENA* **236**, 107732. https://doi.org/10.1016/j.catena.2023.107732 (2024).
44. Qin, S. et al. landslide detection from open satellite imagery using distant domain transfer learning. *Remote Sens.* **13**(17), 3383. https://doi.org/10.3390/rs13173383 (2021).
45. Youssef, A. M., & Pourghasemi, H. R. Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geoscience Frontiers*, **12**(2), 639–655 (2021).
46. Zare, M., Pourghasemi, H. R., Vafakhah, M., & Pradhan, B. Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. *Arabian Journal of Geosciences*, **6**, 2873-2888 (2013).

47. Wu, X., Ren, F. & Niu, R. Landslide susceptibility assessment using object mapping units, decision tree, and support vector machine models in the Three Gorges of China. *Environ. Earth Sci.* **71**(11), 4725–4738. https://doi.org/10.1007/s12665-013-2863-4 (2013).
48. Niraj, K. C., Singh, A. & Shukla, D. P. Effect of the Normalized Difference Vegetation Index (NDVI) on GIS-enabled bivariate and multivariate statistical models for landslide susceptibility mapping. *J. Indian Soc. Remote Sens.* **51**(8), 1739–1756. https://doi.org/10.1007/s12524-023-01738-5 (2023).
49. Pham, B. T., Bui, D. T., Prakash, I. & Dholakia, M. B. Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA* **149**, 52–63. https://doi.org/10.1016/j.catena.2016.09.007 (2017).
50. López, O. A. M., López, A. M., & Crossa, J. Overfitting, model tuning, and evaluation of prediction performance. In *Springer eBooks* (pp. 109–139). https://doi.org/10.1007/978-3-030-89010-0_4 (2022).
51. Shirzadi, A., Bui, D. T., Pham, B. T., Solaimani, K., Chapi, K., Kavian, A., Shahabi, H., & Revhaug, I. Shallow landslide susceptibility assessment using a novel hybrid intelligence approach. *Environ. Earth Sci.* **76**(2). https://doi.org/10.1007/s12665-016-6374-y (2017).
52. Yang, C., Liu, L., Huang, F., Huang, L. & Wang, X. Machine learning-based landslide susceptibility assessment with optimized ratio of landslide to non-landslide samples. *Gondwana Res.* **123**, 198–216. https://doi.org/10.1016/j.gr.2022.05.012 (2023).
53. Yang, S., Li, D., Sun, Y., & She, X. Effect of landslide spatial representation and raster resolution on the landslide susceptibility assessment. *Environ. Earth Sci.* **83**(4). https://doi.org/10.1007/s12665-024-11442-3 (2024).
54. Lee, J., Sameen, M. I., Pradhan, B. & Park, H. Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. *Geomorphology* **303**, 284–298. https://doi.org/10.1016/j.geomorph.2017.12.007 (2018).
55. Adnan, M. S. G. et al. Improving spatial agreement in machine learning-based landslide susceptibility mapping. *Remote Sens.* **12**(20), 3347. https://doi.org/10.3390/rs12203347 (2020).
56. Manaouch, M. et al. Assessment of landslide susceptibility using machine learning classifiers in Ziz upper watershed. *SE Morocco. Phys. Geography* **45**(2), 203–230. https://doi.org/10.1080/02723646.2023.2250174 (2023).
57. Murlidhar, B. R. et al. Prediction of flyrock distance induced by mine blasting using a novel Harris Hawks optimization-based multi-layer perceptron neural network. *J. Rock Mech. Geotech. Eng.* **13**(6), 1413–1427. https://doi.org/10.1016/j.jrmge.2021.08.005 (2021).
58. Lin, M., Teng, S., Chen, G. & Bassir, D. Transfer learning with attributes for improving the landslide spatial prediction performance in sample-scarce area based on variational autoencoder generative adversarial network. *Land* **12**(3), 525. https://doi.org/10.3390/land12030525 (2023).

Acknowledgements

The authors are grateful to IIT Mandi in Kamand, Himachal Pradesh, India, for providing this opportunity and technical tools to conduct this research. Additionally, we would like to thank the GSI, Ministry of Road and Highways, Planet Imagery and Archive, and Google Earth for providing us to use of their datasets in this study. Also, the authors would like to acknowledge the USGS, the Land Processes Distributed Active Archive Centre (LP DAAC) and ASTER for providing the 30-m ASTER DEM to carry out this work. Finally, the first author is greatly thankful to Ms Sneha Shekhar, Research Scholar at IIT MANDI, for her input in the preparation of manuscript.

Author contributions

Ankit Singh: data curation, formal analysis, investigation, validation, visualization, roles/writing—original draft; Nitesh Dhiman: data curation, formal analysis, investigation, roles/writing—original draft. Niraj KC: conceptualization, formal analysis, investigation, methodology Dericks Praise Shukla: conceptualization, formal analysis, investigation, methodology, resources, software, supervision, validation, writing—review and editing.

Declarations

Competing interests

The authors declare that we have no known conflict of interest to disclose.

Ethics approval

This study does not need any ethical approval as this research does not involve humans and/or animals**.**

Consent for publication

All authors give their consent to publish this work in whole.

Additional information

Correspondence and requests for materials should be addressed to D.P.S.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024