communications
earth & environment

ARTICLE

https://doi.org/10.1038/s43247-023-00806-5 OPEN

Landslide susceptibility modeling by interpretable neural network

K. Youssef[1,5,6], K. Shao[2,6], S. Moon[2]✉ & L.-S. Bouchard[1,3,4]✉

Landslides are notoriously difficult to predict because numerous spatially and temporally varying factors contribute to slope stability. Artificial neural networks (ANN) have been shown to improve prediction accuracy but are largely uninterpretable. Here we introduce an additive ANN optimization framework to assess landslide susceptibility, as well as dataset division and outcome interpretation techniques. We refer to our approach, which features full interpretability, high accuracy, high generalizability and low model complexity, as superposable neural network (SNN) optimization. We validate our approach by training models on landslide inventories from three different easternmost Himalaya regions. Our SNN outperformed physically-based and statistical models and achieved similar performance to state-of-the-art deep neural networks. The SNN models found the product of slope and precipitation and hillslope aspect to be important primary contributors to high landslide susceptibility, which highlights the importance of strong slope-climate couplings, along with microclimates, on landslide occurrences.

[1] Department of Chemistry and Biochemistry, University of California Los Angeles, 607 Charles E. Young Dr. East, Los Angeles, CA 90095-1569, USA. [2] Department of Earth, Planetary, and Space Sciences, University of California, Los Angeles, 595 Charles E. Young Dr. East, Los Angeles, CA 90095-1567, USA. [3] Department of Bioengineering, University of California Los Angeles, 607 Charles E. Young Dr. East, Los Angeles, CA 90095-1569, USA. [4] California NanoSystems Institute, University of California Los Angeles, 607 Charles E. Young Dr. East, Los Angeles, CA 90095-1569, USA. [5]Present address: Krannert Cardiovascular Center, Indiana University School of Medicine, Indianapolis, USA. [6]These authors contributed equally: K. Youssef, K. Shao. ✉email: sgmoon@ucla.edu; lsbouchard@ucla.edu

Landslides are a major natural hazard that cause billions of dollars in direct damages and thousands of deaths globally each year[1,2]. Landslides can also cause various secondary hazards, such as damming and flooding, which often leave a region prone to subsequent damage following the initial event[3]. Additionally, landslide debris may cause instability by perturbing river sedimentation and disrupting ecosystems[3,4]. As landslide hazards are expected to increase due to climate change, scientists have sought to more accurately assess landslide susceptibility[5–10], an estimate of the probability that a landslide may occur in a specific area, with the goal of mitigating the impact of landslides on the economy, public safety, and local ecosystems.

Landslide occurrences are influenced by various factors including physical attributes of the terrain, such as slope, relief, and drainage areas, and material properties such as the density and strength of soil and bedrock[11–14]. Also, environmental conditions such as climate, hydrology, ecology, and ground motion due to earthquakes may contribute to slope instability[15–17]. Landslide susceptibility is calculated from these various controlling factors either through physically-based models[12,13,16,18], data-driven approaches utilizing statistical analysis[19,20], or machine learning techniques (ML), including random forest, support vector machines, and deep neural networks (DNN)[6,21–26].

While substantial work has been devoted to assessing susceptibility, each model has shortcomings. Physically- or mechanistically-based approaches, based on the equilibrium between driving and resisting forces, have been widely applied to assess slope stability[11–13,27]. However, mechanistic models have limitations, including a limited number of variables, simplified assumptions of landslide geometry and certain environmental conditions (e.g., antecedent moisture, bedrock structure), and the high cost of geotechnical exploration necessary to estimate and calibrate for accurate subsurface properties (e.g., cohesive strength, pore pressure, weathering profile)[15]. Alternatively, data-driven approaches, including statistical and ML methods, can handle a large number of controls to assess susceptibility. Statistical methods such as logistic regression and likelihood ratios[19,20,28] can utilize a multitude of landslide controls as inputs. Scientists using these data-driven approaches have obtained a measurable degree of success in determining areas susceptible to landslides[6,19,20]. However, these data-driven models also rely on the expert's choices, preconditions, and classifications of input variables. The outcome of these models' results, the landslide susceptibility map, does not decouple individual feature contributions to landslide susceptibility nor account for their interdependencies due to the limited computational capabilities in conventional approaches[28].

Machine learning approaches, such as fuzzy logic algorithms, support vector machines, and DNNs, have been applied to landslide studies for mapping landslide susceptibility[22,24,29]. DNNs have achieved improved performance compared to both statistical methods and other ML approaches due to their use of nonlinearities, complex interdependencies of interlayer connections, as well as internal representations of data[21–24,30–32]. However, the black box nature of DNNs has been a major hurdle for their adoption in practice and research, making it difficult for experts to understand and trust their outcomes. With DNNs, it is nearly impossible to determine the exact relation between individual inputs and outputs[30–32]. Lack of interpretability is a weakness of DNNs and a fundamental drawback for high-stakes applications such as landslide mitigation where decisions impact lives and result in untold costs of insurance and reconstruction[2,3,33]. Interpretability would ideally provide decision-makers with a list of contributing factors ranked in order of importance, as well as any possible interplay between these factors.

The DNN's lack of interpretability has prompted the Defense Advanced Research Projects Agency's (DARPA) third wave of AI call in 2017 and the European Union's 2018 General Data Protection Regulation, which grants a right to an explanation, for algorithmic decisions that are made[34]. Next-generation AI systems refer to the so-called explainable or interpretable AI (XAI) models. The latter must be able to construct explanatory models for classes of real-world phenomena that can be communicated to humans[32]. Various XAI categories have since been defined in the literature based on factors such as application and methodology, where each category is further divided into subclasses[35]. Although the use of XAI in research is expanding, existing approaches aimed at explaining black box models exhibit a trade-off between accuracy and interpretability, resulting in a large gap in performance (e.g. ref. [36]). Recently, Rudin[30] showed that with proper feature engineering, and a shift from explaining existing black box models to creating methods with inherently interpretable models, the trade-off between accuracy and interpretability can be circumvented.

To this end, we propose a framework that bridges the gap between explainability and accuracy for landslide susceptibility models. This framework utilizes a hybrid of model extraction methods and feature-based methods to generate a fully interpretable additive ANN model while simultaneously pruning features and feature interdependencies that are redundant or suboptimal to model performance and generalizability. Additive ANN are a type of generalized additive models (GAM) that have been recently gaining popularity[37–40]. They combine separate ANNs, each specializing in a single feature, to optimize a common outcome. Unlike other additive XAI methods such as Shapley additive explanations (SHAP) that aim to explain the local behavior of a black box model[41], additive neural networks are inherently interpretable models with both local and global interpretability. Model extraction methods aim to train an explainable "student" model to mimic the behavior of a "teacher" model, and feature-based methods aim to analyze and quantify the influence or the importance of each input feature[35]. Our optimization framework possesses full interpretability, high accuracy, high generalizability, and low model complexity. Most notably, toy problems included in the Supplementary Note 1 demonstrate the capability of our framework to generate fully interpretable additive ANNs with controlled complexity and accuracy that can match state-of-the-art DNNs, as well as find globally optimal unique solutions. Furthermore, we utilize dataset division and outcome interpretation techniques uniquely suitable for landslide susceptibility modeling applications with spatially dependent data structures. We refer to the approach as superposable neural network (SNN) optimization in reference to the automated way of incrementally generating the additive ANN model and determining the contributing features. Our approach is different from the more commonly followed approach of designing a fixed network architecture with a fixed set of manually selected input features where the entire network is jointly trained in an end-to-end fashion[40].

In this study, we model three different regions of the easternmost Himalaya using SNNs. For comparison, we include results from a physically-based slope stability model (SHALSTAB), two statistical methods (logistic regression and likelihood ratios), in addition to state-of-the-art DNN teacher models. Finally, we examine the SNN-determined relationship and relative importance of each feature's contribution to landslide susceptibility and discuss how information extracted from the SNN can provide insights into the physical controls of landslides in our studied regions. Our results highlight underappreciated, important controls such as the product of slope and precipitation and hillslope aspects in the studied region. Controls that consist of

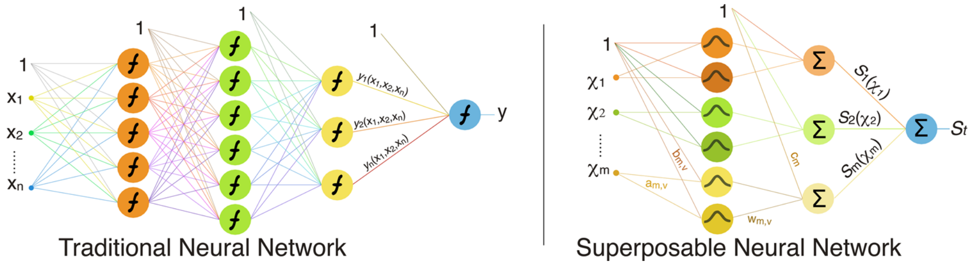

Fig. 1 Conventional DNN architecture vs SNN architecture. In a conventional DNN, features are interconnected and interdependencies are embedded in the network, making them virtually impossible to separate. In a SNN, features and feature interdependencies that contribute to the output are found in advance and explicitly added as independent inputs. Radial basis (Gaussian) activation functions are used in the SNN, where each neuron is connected to one input only. The $x_1, x_2, \ldots x_n$ refer to a set of n original features, and $\chi_1, \chi_2, \ldots \chi_M$ refer to a set of M composite features. y and S_t refers to DNN and SNN outcomes of total susceptibility, respectively. The symbols in this figure are defined and explained in the main text, Eq. (1).

products of input features can help unveil the influences from feature interactions.

Superposable neural networks. SNNs are an additive ANN architecture that enforces no interconnections between inputs (Fig. 1). The lack of interconnections between features is the key to explainability. Unlike DNNs where interdependencies between features are embedded in layers of network connections, interdependencies in SNNs are explicitly created as a product function of more than one original input feature. We refer to these products as "composite features" (see "Methods" for details). Important interdependencies between features are automatically determined by isolating composite features contributing to the desired outcome. Contributing composite features are explicitly added as independent inputs to the model, while non-contributing composite features are discarded (see SNN training flow diagram in Fig. 2 as well as "Methods"). Furthermore, we label SNNs according to the highest level of composite features used in training the model, which refers to the maximum number of features allowed in multivariate interactions. For example, a Level-3 SNN can include Level-1, Level-2 and Level-3 composite features. Using composite features, SNNs can approximate any continuous function for inputs within a specific range as a polynomial expansion to any desired precision. This ability allows SNNs to retain a level of accuracy on par with state-of-the-art DNNs.

The SNN is represented mathematically by the function (Eq. (1)):

$$S_t(\{\chi_j\}) = \sum_j \left(\sum_k w_{j,k} e^{-(a_{j,k}\chi_j + b_{j,k})^2} + c_j \right). \tag{1}$$

It contains only two hidden layers of neurons with radial basis activation functions in the first layer and linear activation functions in the second layer. The choice of radial basis activation functions allows the user to minimize the number of neurons in the model, maximizing the efficiency of our method. Each input χ_j is exclusively connected to a group of neurons to form an independent function $S_j = \sum_k w_{j,k} e^{-(a_{j,k}\chi_j + b_{j,k})^2} + c_j$ and the SNN output $S_t = \sum_j S_j$ is the sum of all independent functions, where $j =$ 1 : number of features (M), $k = 1$: number of neurons per feature (v), and χ_j is the j^{th} composite feature. In addition to determining the features and interdependencies between features that contribute to the outcome, the SNN architecture enables the quantification of their exact contributions to the output.

The model simplicity and lack of connections between neurons associated with different features makes our model fully interpretable and mathematically analyzable. However, this aspect also makes the model highly constrained, which poses challenges on its training. Jointly training the model with commonly used gradient descent-based optimizers proved to be extremely difficult to converge, especially as the number of features increases. Our optimization approach enables the separate training of individual neural networks by utilizing several state-of-the-art ML techniques (multi stage training, knowledge distillation, second order optimization[42–47]) to deliver a model that is optimal in terms of performance and remarkably simple in terms of architecture. The reduction in model complexity, while maintaining an accuracy that rivals that of DNNs, which are orders of magnitude more complex in terms of number of parameters and redundancies in interconnectivities, presents a substantial advance.

A validation of our approach using toy models is included in Supplementary Note 1.1 and 1.2. In the first application, we create a synthetic dataset by adding known functions of composite features and test the ability of the SNN to find the contributing features and extract their functions from the data. The second application incorporates up to Level-4 feature interactions and demonstrates the impressive ability to extract boolean relationships from synthetic data. Boolean inference tasks are notoriously difficult because of the high degree of stiffness and nonlinearity between input and output. The SNN optimization algorithm is described in "Methods".

Landslides in the easternmost Himalaya. Asia holds the majority of human losses due to landslides globally, with a high concentration in the Himalayan Arc[1,2]. In particular, the easternmost Himalaya has a high susceptibility to numerous landslides from steep slopes, extreme precipitation events, flooding, and frequent earthquakes[48–52] (Fig. 3 and Supplementary Fig. 1). We generated a landslide inventory of the easternmost Himalaya by combining the manual delineation of landslide areas with a semi-automatic detection algorithm[53,54] (Fig. 4a–c; a flowchart diagram in Supplementary Fig. 2, exemplary landslides in Supplementary Fig. 3). Within the entire study area of 4.19 × 10^9 m^2, the total number of mapped landslides is 2289, and their areas range from 900 to 1.96 × 10^6 m^2 (Supplementary Table 1, Supplementary Fig. 4)[55]. Landslide densities calculated over a 2.25 km^2 window are generally high in the range front (max 0.121) and low in the hinterland (~0.039).

Within the easternmost Himalaya, we selected three regions (the Dibang, Lohit, and range front regions) with varying ranges of landslide controls to test the performance and application of the SNN model (Fig. 3). Hereafter, we refer to Dibang, Lohit, and range front regions as the N-S, E-W, and NW-SE regions, respectively. Testing the SNN over these three regions with varying environmental conditions will allow us to examine the following: (1) whether the SNN can identify universal or distinctly different controls of landslides, and (2) whether SNN-determined functions of feature contributions to susceptibility, S_j, are similar or different across these three regions. We used 15 single features in the SNN model (Supplementary Fig. 5, Supplementary Table 2). The

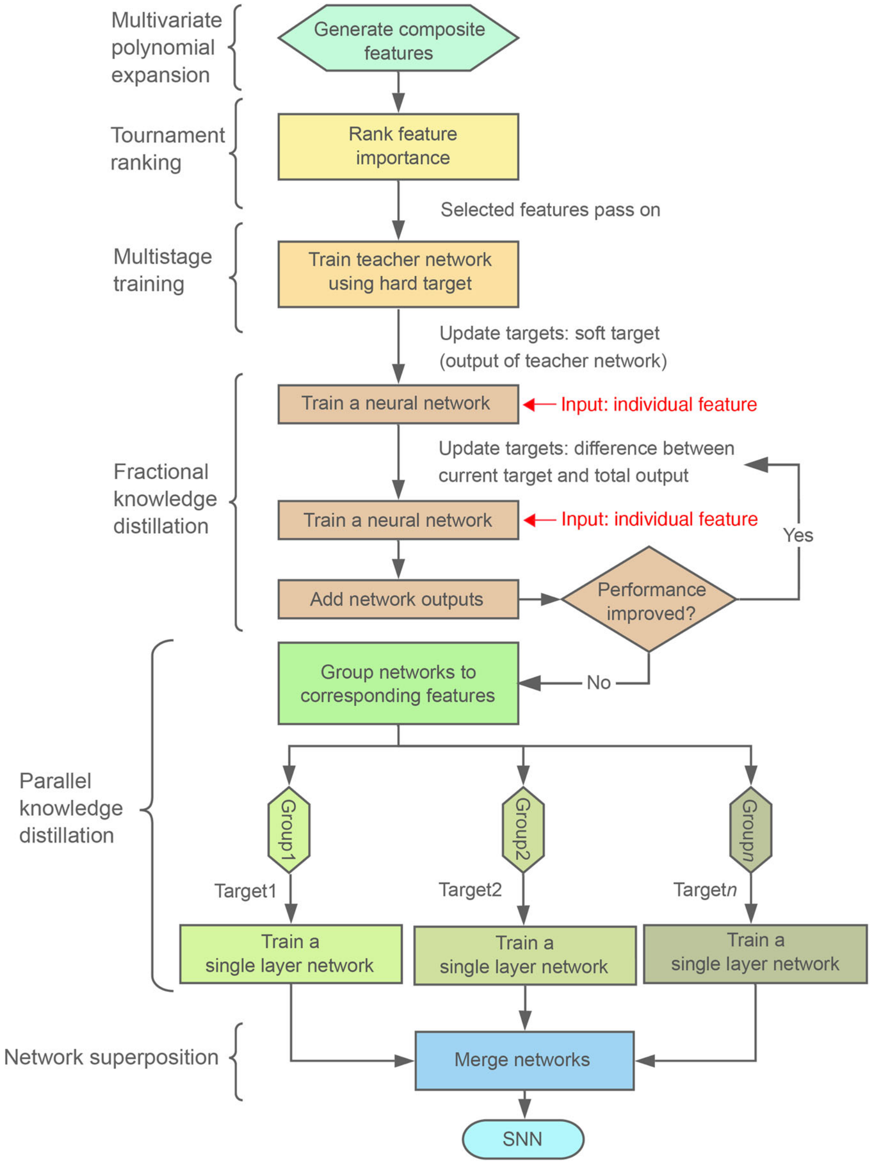
Multivariate polynomial expansion
Generate composite features
Tournament ranking
Rank feature importance
Selected features pass on
Multistage training
Train teacher network using hard target
Update targets: soft target (output of teacher network)
Train a neural network
Input: individual feature
Fractional knowledge distillation
Update targets: difference between current target and total output
Train a neural network
Input: individual feature
Yes
Add network outputs
Performance improved?
No
Group networks to corresponding features
Parallel knowledge distillation
Group1
Group2
Groupn
Target1
Target2
Targetn
Train a single layer network
Train a single layer network
Train a single layer network
Network superposition
Merge networks
SNN

Fig. 2 Superposable neural network training flow diagram. The flow diagram shows the methods used in our study, which include the feature-selection model and multistage training. Our feature-selection model based on multivariate polynomial expansion and tournament ranking allows for the exploration of multiple combinations of parameters without relying on an expert's choices, precondition, or classification of input features and identify a set of optimal composite features that are relevant to the landslide susceptibility. Then, multiple steps of knowledge distillation are used to quantify each control's contribution to susceptibility (S_j, where j corresponds to single layer network). By superposing S_j, we create an additive, superposable neural network (SNN) model for total landslide susceptibility. The details of each methodology are explained in "Methods".

15 single features include aspect (*Asp*), mean curvature ($Curv_M$), planform curvature, profile curvature, total curvature, discharge, distance to channel ($Dist_C$), distance to faults ($Dist_F$), distance to the Main Frontal Thrust and suture zone ($Dist_{MFT}$), drainage area, elevation (*Elev*), local relief (*Relief*), mean annual precipitation (*MAP*), number of extreme rainfall events (*NEE*) and slope. The inclusion of these variables is based on previous studies that examined landslide controls in the Himalayan region[20,56–58]. The details of study area, landslide inventory, input data sources and calculation are presented in "Methods".

Results and discussion

SNN Implementation. We modeled landslide susceptibility of the easternmost Himalaya using Level-1, 2 and 3 SNN models. We find that the Level-3 SNN is able to achieve over 99% of the accuracy of the state-of-the-art teacher DNN, and the Level-2 SNN is able to achieve over 98%. Given the small difference, we assume the explainability of the Level-2 SNN to be sufficient for our analysis. Due to the nature of this application, a special data partitioning method was devised to partition each region into roughly 70% for training and 30% for validation, which utilizes Pythagorean tiling to partition the regions in a spatially representative manner (Fig. 5) (see "Methods" for details).

A threshold value of S_t is used as a binary classifier to predict landslides and compare them with observed landslides from our inventory. We selected a threshold susceptibility corresponding to the closest point to a perfect classifying model with 100% true positive rate and 0% false positive rate on a receiver operating characteristic (ROC) curve. Areas with S_t greater and lower than this threshold are classified as landslide (*ld*) and non-landslide (*nld*) areas, respectively, in the model (Fig. 4d–f).

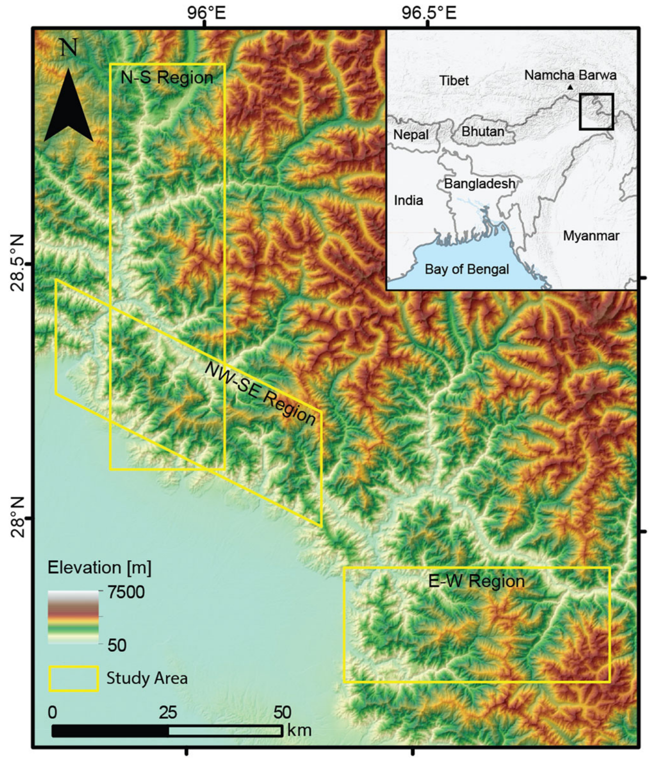

Fig. 3 Study area in the easternmost Himalaya. Colors represents the elevation[71], and yellow boxes indicate our N-S (Dibang), NW-SE (range front), and E-W (Lohit) oriented study regions. The inset map shows the eastern Himalayan region with our study area shown in a black box and national borders shown in dark gray lines.

Comparison with traditional landslide susceptibility modeling. In addition to the comparison against the state-of-the-art DNN teacher model, we provide comparisons of Level-1 and Level-2 SNN performance to a number of traditional methods, all applied to the same regions and using the same inventory data. Comparison of different models on the same area is needed since model performance cannot be directly compared to model performance published in other papers, since those papers focused on different regions.

First, we investigated each of the 15 single features as individual classifiers for landslide occurrences. Second, we applied a physically-based slope stability model (SHALSTAB) for soil landslides[12,27,59] that couples infinite slope stability and steady-state hydrology for cohesionless material. Considering that most landslides in our inventory are soil landslides (Methods), SHALSTAB was assumed to be suitable for our analysis. We modified SHALSTAB and calculated a metric called the failure index (*FI*), as the ratio of driving to resisting forces on a hillslope. *FI* is equivalent to the inverse of the factor-of-safety, which represents the propensity for landslide occurrence. Third, we used two commonly used statistical models, logistic regression and likelihood ratios, to model landslide susceptibility[28,60,61]. Logistic regression (hereafter, *LogR*) is based on a multivariate regression between a binary response of landslide occurrence and a set of predicting features that are continuous, discrete, or a combination of both types[60]. Likelihood ratios (*LR*) are calculated as the ratio of the percentage of landslide pixels relative to total landslide pixels divided by the percentage of pixels relative to the total area within a specific range of feature values[60,61]. Previous studies have quantified the ratio of the probability of landslide occurrences to the probability of non-occurrences or all-occurrences within a range of feature values and referred to it as the likelihood ratio, frequency ratio, or probability ratio[28,60,61]. A ratio of 1, >1, or <1 indicates an average, above-average, or below-average likelihood of landslide occurrence, respectively, within the feature range compared to that of the study area. Landslide susceptibility for each pixel is calculated as the sum of the corresponding *LR* from each feature's value. A threshold value of modeled landslide susceptibility from *LogR* and *LR* can be used as a binary classifier to predict landslides following a similar procedure that we used for the SNN.

We assessed model performance based on various metrics including area under the receiver operating characteristic curve (AUROC). In addition, we calculated the statistical measures of accuracy, sensitivity (probability of detection, POD), specificity (probability of false detection, POFD), and POD-POFD. We also calculated the 95% confidence interval of mean AUROC from the statistical and neural network model outputs based on a 10-fold cross validation. The 95% confidence intervals of mean AUROC can be used to determine whether model performances are statistically different (model and method details in Supplementary Note 2).

We show that the SNN model's performance is comparable to that of the teacher, second-order-optimized DNN, while providing a statistically significant improvement over commonly used physically-based and statistical models. AUROCs of Level-1 and Level-2 SNNs are 0.856 and 0.890, respectively, calculated as the averages from the three study regions. The value for each region is presented in Supplementary Table 3. The Level-2 SNNs captured over 98% of the teacher model (MST) performance across all three study regions. The Level-2 SNN is optimal in the sense that it provides high accuracy (comparable to deep nets) and relatively simple model complexity (hereafter, SNN refers to Level-2 SNN).

The SNN achieved ~ 21% average improvement in AUROC over the top performing single original features (i.e., *MAP* or slope, AUROC = 0.737), ~22% over a physically-based model (SHALSTAB) (AUROC = 0.727), and ~5–8% over logistic regression (AUROC = 0.848) and likelihood ratios (AUROC = 0.823) in our three study regions. The 95% confidence intervals of the mean AUROC of the SNN lie above and do not overlap with those of the statistical models (Supplementary Table 4). In addition, the vast majority of other performance metrics such as accuracy, POD, POFD, and POD-POFD from the SNN are improved over these other methods as well (Supplementary Table 5).

SNN model explainability. The SNN-determined independent functions S_j show varying relationships between both features and feature interdependencies, and their absolute susceptibility contribution (Fig. 6). $S_{MAP^{*}Slope}$ and $S_{NEE^{*}Slope}$ generally exhibit steep increases with feature value, followed by asymptotic behavior (Fig. 6a, d, g). These nonlinear relationships between landslide susceptibility and the product of slope and climatic features of *MAP* and *NEE* are similar in all three regions. In addition, S_{Asp} shows a peak around 145° to 180°, which indicates a preference for south-facing slopes, likely due to moisture from the Bay of Bengal[49] (Supplementary Fig. 6, Supplementary Note 3). These functional relationships are similar to those deduced by the *LR* statistical method that represent the likelihood of landslide occurrence. However, unlike *LR*, which assume the same, average likelihood ($LR = 1$) for each feature, S_j corresponding to $LR = 1$ varies depending on a feature's absolute, decoupled contribution to landslide susceptibility.

The SNN provides the exact contribution of each individual feature to the total susceptibility outcome, which allows us to

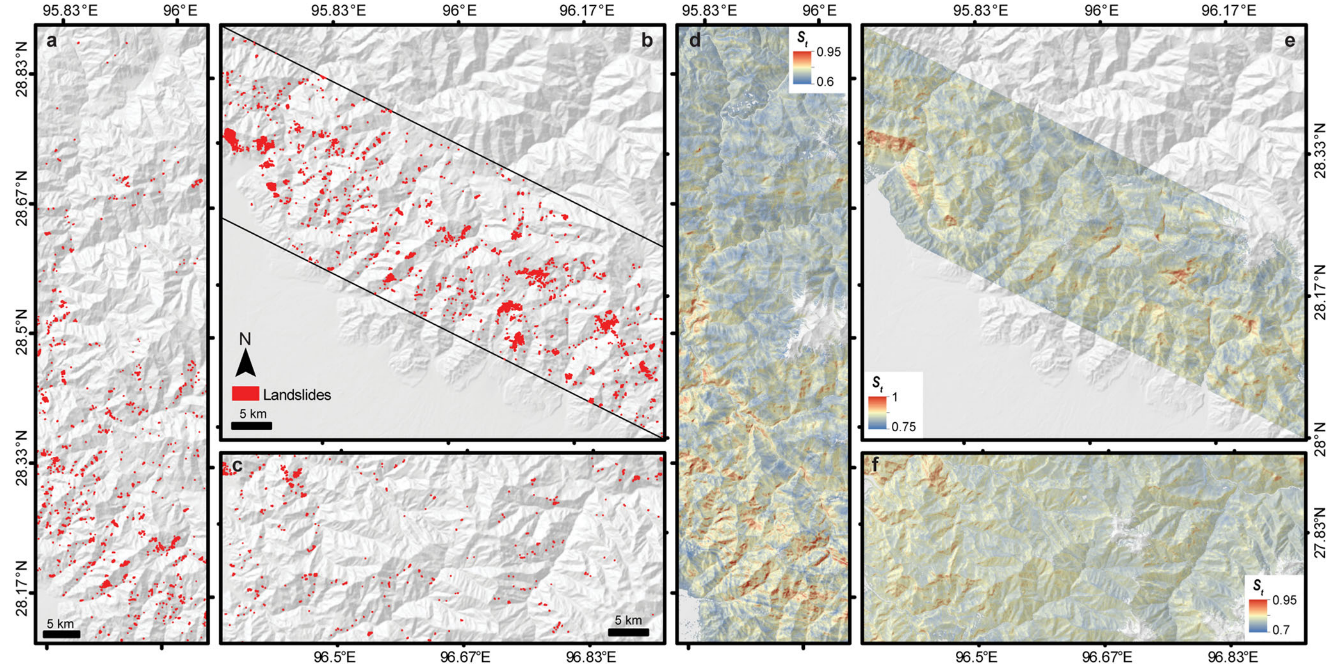

Fig. 4 Mapped landslides and modeled susceptibility. Spatial distribution of **a**–**c** mapped landslides and **d**–**f** modeled landslide susceptibility for the **a**, **d** N-S, **b**, **e** NW-SE, and **c**, **f** E-W study regions. **a** 959, **b** 1536, and (**c**) 386 landslides are shown in red polygons in (**a**–**c**). Total susceptibility at the pixel scale (S_t) from the Level-2 superposable neural network are shown in (**d**–**f**). The threshold S_t values that are used to classify landslide and non-landslide pixels in the model are **d** 0.767, **e** 0.861, and **f** 0.816, respectively.

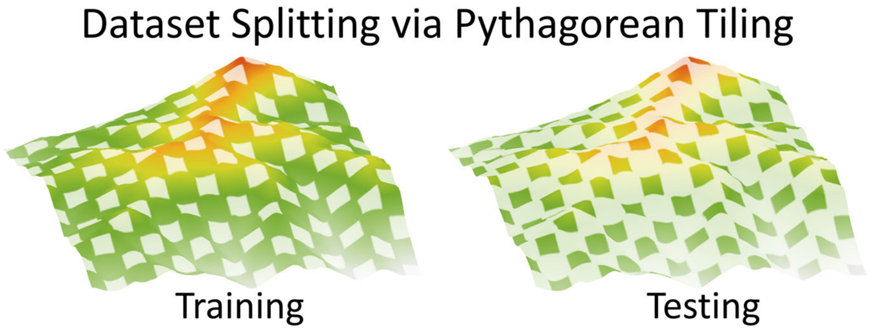

Fig. 5 Illustration of spatial data partitioning using Pythagorean tiling. Pythagorean tiling is used to divide data from the modeled region in a spatially representative manner that maintains variability between training and testing partitions. Using Pythagorean tiling, we generate a checkerboard-like pattern with a 70/30% square ratio, where bigger squares correspond to training and smaller squares correspond to testing.

quantify the relative importance of landslide controls in different localities and across varying spatial scales (Fig. 7d–f). Causal rankings of individual features that drive landslides can be obtained by calculating the susceptibility difference between *ld* v.s. *nld* pixels, $\Delta\bar{S}_j$, within a region of interest for each individual feature. This is demonstrated both globally (Fig. 7a–c), where the region of interest is the entire region of study, and locally (Fig. 8a–c), where the region of study is divided into hundreds of smaller regions of interest, each consisting of a 2.25 km^2 window. For comparison, we also identified the primary controls of landslides and their relative contributions from the Level-1 SNN and weights determined by the logistic regression model (Supplementary Note 2, Supplementary Fig. 7).

Composite features involving topographic and climate features are identified as important landslide controls for our study area. Namely, the product of slope and *NEE* or *MAP*, *Asp*, and the product of *Asp* and *Relief* tend to have large $\Delta\bar{S}_j$ across all three regions (Fig. 7a–c). In addition, those features are identified as locally important, primary features when analyzing using a 2.25 km^2 window throughout the area (Fig. 8a–c). The primary features of *MAP***Slope* and *NEE***Slope* are consistent among our three study regions in the easternmost Himalaya, despite differences in the spatial distribution and magnitude of precipitation and proximity to a major fault with a history of earthquakes (Supplementary Fig. 1). Although these composite features may not be the largest contributor for total susceptibility (Fig. 7d–f), they tend to have different contributions for *ld* and *nld* areas and lead to a large $\Delta\bar{S}_j$ (Fig. 7a–c).

SNN-derived individual feature contributions are used to assess the relative importance between climate and slope features. The feature independence in the SNN additive architecture and the use of composite features allows us to isolate the effect of slope or climate in the model. (1) The exact marginal contribution is calculated for Level-2 features involving slope or climate (i.e., *Asp*, *NEE*, and *MAP*). (2) Level-1 slope and Level-2 slope marginal contributions are added together to produce the total

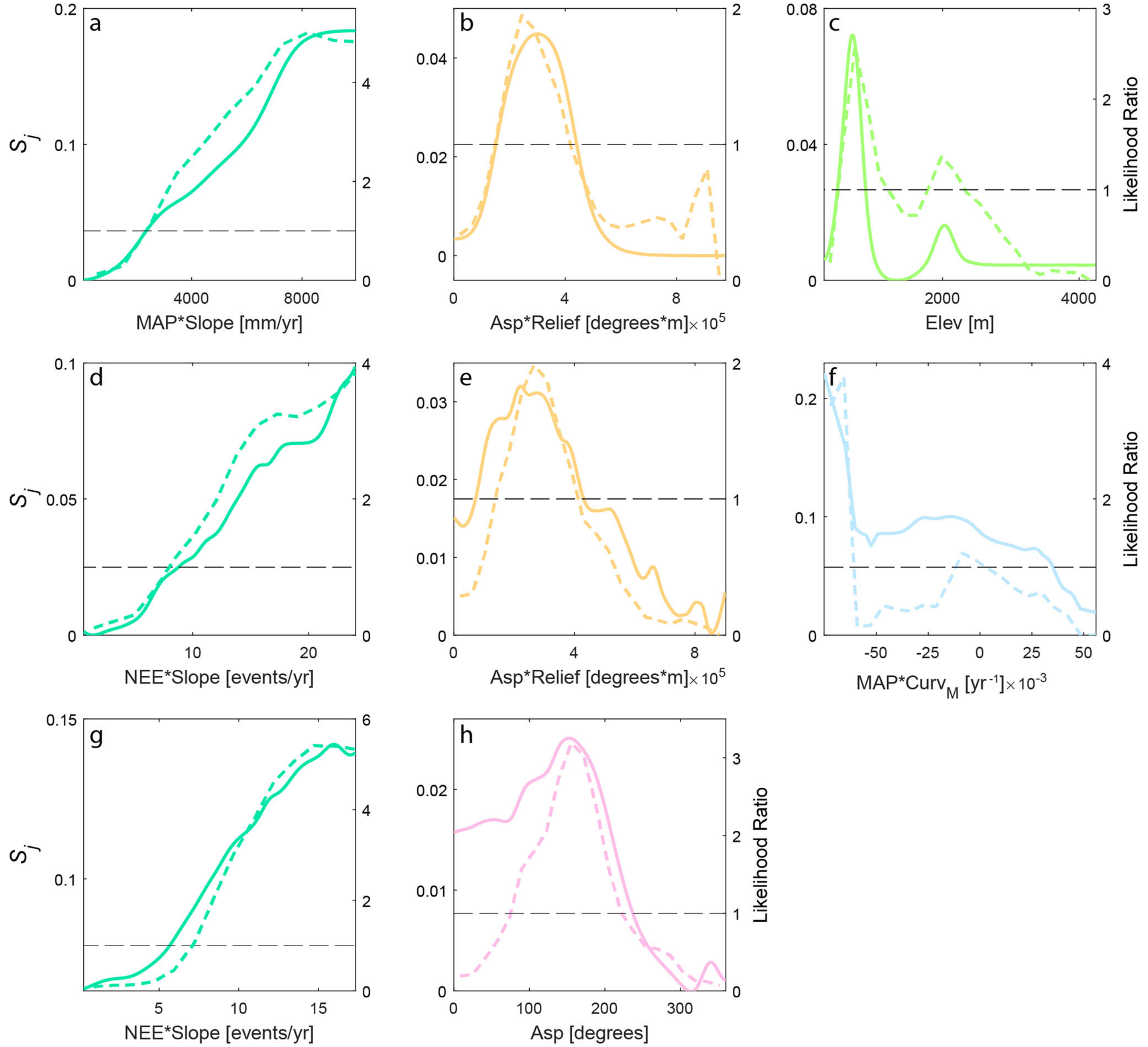

Fig. 6 Individual feature contributions to total susceptibility. Independent functions of S_j identified as primary landslide controls are shown for the **a**–**c** N-S, **d**–**f** NW-SE, and **g**, **h** E-W study regions. Likelihood ratios (*LR*), representing the likelihood of landslide occurrence for a specific range of feature values, are shown as short, dashed, colored lines with corresponding right-side *y*-axes for reference. $LR = 1$ and $LR > 1$ represent the average and above-average likelihood of landslide occurrence, respectively. Note that S_j corresponding to $LR = 1$, shown as long-dashed black lines, differ between features because the SNN quantifies the absolute contributions of S_j decoupled from other features. Features related to topography, aspect, climate, and geology are shown in green, pink, blue, and brown or combinations thereof, respectively. Mean annual precipitation (*MAP*), number of extreme rainfall events (*NEE*), aspect (*Asp*), elevation (*Elev*), mean curvature (*Curv*$_M$), and local relief (*Relief*). The asterisk * indicates algebraic multiplication of two features.

susceptibility contribution from the slope, $S_{t,Slope}$. (3) Level-1 climate and Level-2 climate marginal contributions are added together to produce total susceptibility contribution from climate features, $S_{t,Climate}$. In Fig. 8d–f, we compare the relative importance of slope and climate features using our approach that separates their contributions between *ld* and *nld* pixels throughout the region. Then, we calculate the difference between $\Delta\bar{S}_{t,Slope}$ and $\Delta\bar{S}_{t,Climate}$, divided by the threshold susceptibility value, $S_{t,threshold}$, for each respective region. We find that ~74%, 54%, and 54% of localities have a larger contribution from climate features than that of slope for the N-S, NW-SE, and E-W regions, respectively, emphasizing an overall importance of climatic features that drive landslides.

Accurate and interpretable landslide susceptibility from the SNN. Whereas many XAI efforts involve a trade-off between accuracy and interpretability, our SNN does not compromise accuracy. Given the SNN's inherent and unique ability to decouple individual feature contributions and select feature interdependencies, we can easily isolate local contributions from primary controls discovered by the SNN (Fig. 8). Our local analyses for assessing landslide controls indicate that the contribution of climate features, such as *NEE*, *MAP*, and *Asp*, to landslide susceptibility tends to surpass that of slope for a majority of landslide occurrences in this area. These results highlight a prevalent climatic control on landslide occurrences in the easternmost Himalayan region. Due to the eastward

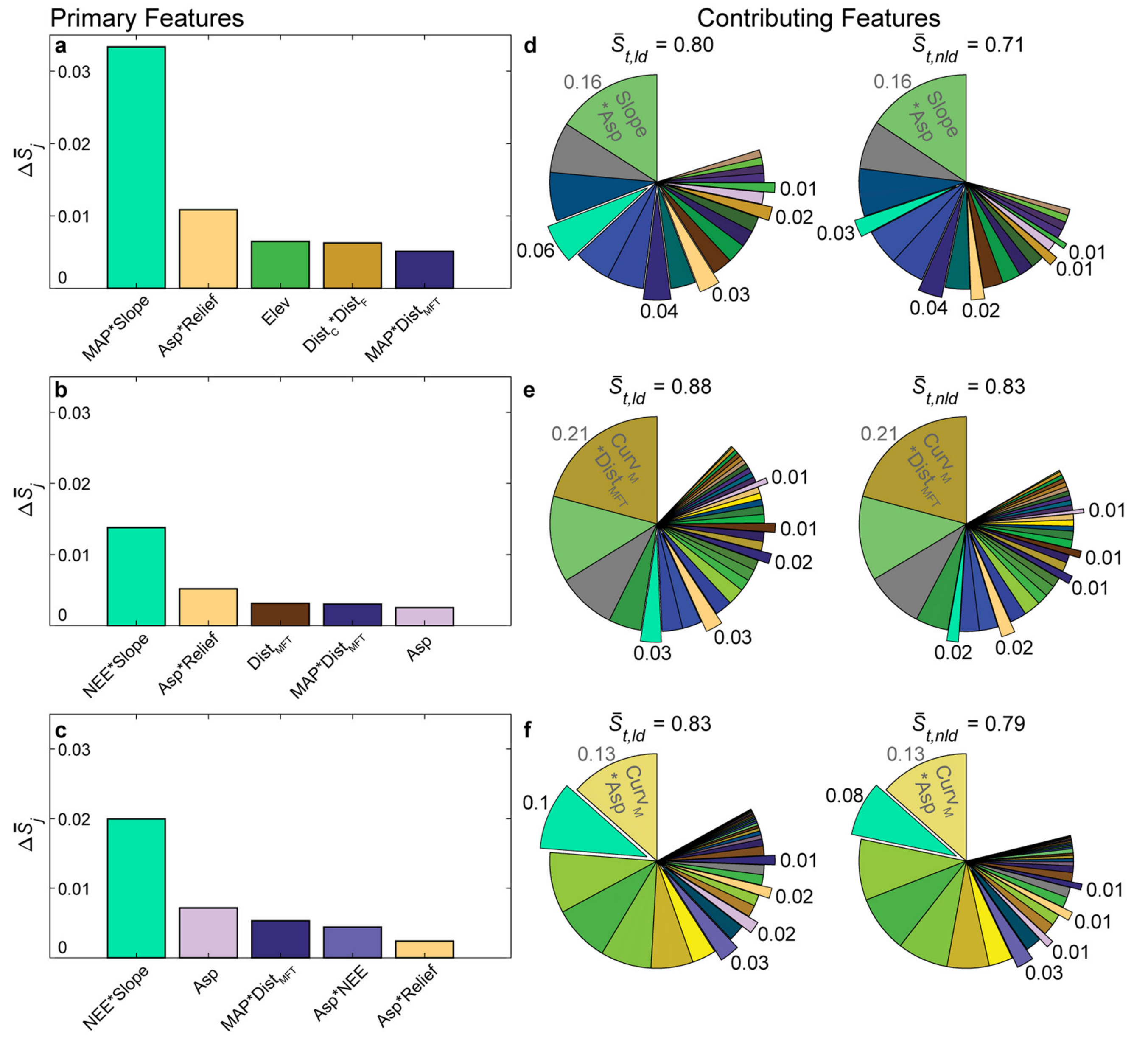

Fig. 7 Feature contributions to total susceptibility. (**a**, **d**) N-S, (**b**, **e**) NW-SE, and (**c**, **f**) E-W study regions. Bar charts in (**a**–**c**) represent $\Delta\bar{S}_j$ in descending order, and pie charts in (**d**–**f**) represent average S_j ($\bar{S}_j$) contributions to landslide (*ld*) and non-landslide (*nld*) areas. $\Delta\bar{S}_j$ represents the difference in average contribution between areas of *ld* and *nld* in each region. Extruding pie chart features are features with large $\Delta\bar{S}_j$ found in the corresponding bar chart on the left. Features related to topography, aspect, climate, and geology are shown in green, pink, blue, and brown or combinations thereof, respectively. Mean annual precipitation (*MAP*), number of extreme rainfall events (*NEE*), aspect (*Asp*), elevation (*Elev*), mean curvature (*Curv$_M$*), distances to channel (*Dist$_C$*), all faults (*Dist$_F$*), and the Main Frontal Thrust and suture zone (*Dist$_{MFT}$*), and local relief (*Relief*). The asterisk * indicates algebraic multiplication of two features. Information regarding features is provided in "Methods".

increasing trends of precipitation rate and variability along the Himalaya, the easternmost Himalaya contains one of the largest strike-perpendicular climatic variations across the steep mountain range[49]. This considerable climate gradient from the range front to the hinterland likely impacts landslide susceptibility in the easternmost Himalaya.

The transparency of our SNN model offers insight into potential mechanisms of landslides and the relative importance of controlling factors. First, the SNN highlights the important, yet under-appreciated controls of *NEE*Slope*, *MAP*Slope*, *Asp*, and *Asp*Relief* (Fig. 8), which implies a dominant occurrence of precipitation-induced landslides in our study site. However, these topography-climate composite features reveal the importance of both incorporated features. These features comprising the product between slope and precipitation rates and intensity as well as that of aspect and relief suggest that landslides are affected by strong slope-climate couplings and aspect-related microclimates.

The nonlinear asymptotic function of $S_{MAP*Slope}$ and $S_{NEE*Slope}$ (Fig. 6a, d, g) can be explained by a physical mechanism of rainfall-induced landslides that induces slope failure due to an increase in pore-water pressure and subsurface saturation[62]. The modeled total landslide susceptibility (S_t) is analogous to the physically-derived failure index (*FI*), which is equivalent to the inverse of the factor-of-safety. *FI* is formulated from equilibrium on an infinite, cohesionless slope considering a pore

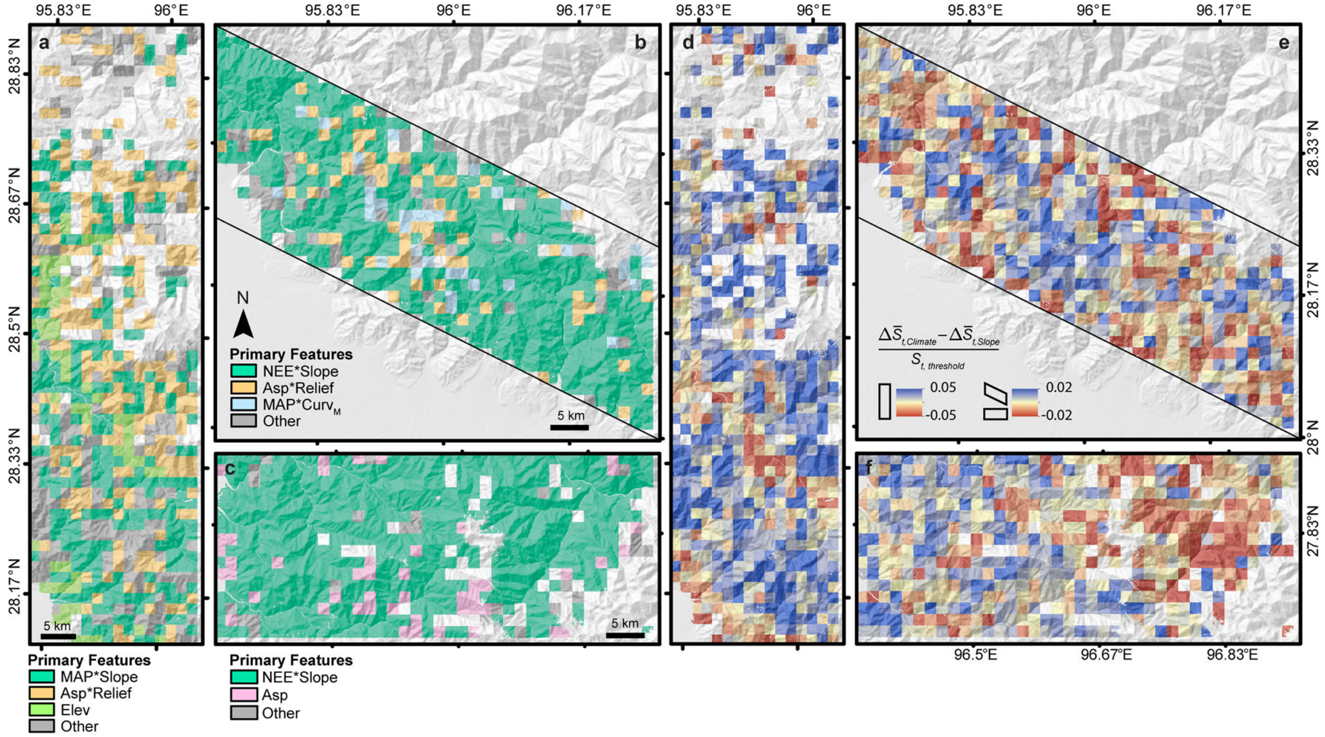

Fig. 8 Important controls for landslides. Spatial distribution of **a**–**c** primary features identified as locally important controls of landslides and **d**–**f** relative climate vs slope susceptibility contributions for the **a**, **d** N-S, **b**, **e** NW-SE, and **c**, **f** E-W study regions. The locally important control in (**a**–**c**) is identified as the feature with the largest difference in average contribution ($\Delta\bar{S}_j$) between areas of landslides (*ld*) and non-landslides (*nld*) within a 2.25 km^2 window. The contribution from climate features ($\Delta\bar{S}_{t,Climate}, j = Asp, NEE, MAP$) relative to that of slope ($\Delta\bar{S}_{t,Slope}$) normalized by the corresponding threshold S_t is shown in (**d**–**f**). Windows with a higher climate contribution are colored blue while those with a greater slope contribution are colored red. Windows of no data contain a majority of unmapped areas or indicate lack of modeled landslides. Features related to topography, aspect, climate, and geology are shown in green, pink, blue, and brown or combinations thereof, respectively. Mean annual precipitation (*MAP*), number of extreme rainfall events (*NEE*), aspect (*Asp*), elevation (*Elev*), mean curvature (*Curv$_M$*), and local relief (*Relief*). The asterisk * indicates algebraic multiplication of two features.

pressure effect based on SHALSTAB[12,59] as:

$$FI = \frac{S}{S_0}\left(1 - W\frac{\rho_w}{\rho_s}\right)^{-1} \qquad (2)$$

where S_0 is the threshold slope, S is the local slope, ρ_s is the wet bulk density of soil (2.0 g/cm^3), ρ_w is the bulk density of water (1.0 g/cm^3), and W is wetness. W is calculated as a ratio between local hydraulic flux from a given steady-state precipitation rate relative to that of soil profile saturation[12]:

$$W = \frac{h}{z} = \frac{qA}{bT\sin\theta} \qquad (3)$$

where h is the saturated height of the soil column (L), z is the total height of the soil column (L), q is the steady-state precipitation during a storm event (L/T), A is the drainage area (L^2) draining across the contour length b (L), T is the soil transmissivity when saturated (L^2/T), and θ is the local slope in degrees. W varies from 0 (unsaturated) to 1 (fully saturated). See Supplementary Note 2 for model details.

Expansion of the denominator in a geometric series gives:

$$FI = \frac{S}{S_0}\left(1 + W\frac{\rho_w}{\rho_s} + W^2\left(\frac{\rho_w}{\rho_s}\right)^2 + O(W^3)\right) \equiv \frac{S}{S_0}k(W). \qquad (4)$$

The approximated FI has three components: local slope S, threshold slope S_0, and $k(W)$, which represents the degrees that landslides are promoted by subsurface saturation. $k(W)$ varies from 1 (unsaturated) to 2 (fully saturated). The multiplication of local slope and $k(W)$, which has an upper bound, mimics the nonlinear asymptotic function of S_{MAP^*Slope} and S_{NEE^*Slope}. This asymptotic increase in susceptibility is similar to observations of other precipitation-induced landslides, but different from earthquake-induced landslides whose occurrences increase nonlinearly with increasing slope[63,64].

Second, the identified controls of *MAP*, *NEE*, and *Asp* imply that local precipitation infiltration on steep slopes may be the dominant contributor to subsurface saturation in the easternmost Himalaya. A change in climatic conditions can raise volumetric water content and porewater pressure. This rise leads to an increased degree of subsurface saturation (i.e., *W*) and subsequently induces slope failure. Previous physically-based slope stability models consider various climatic factors (e.g., rainfall amount and intensity, subsurface convergence flow) to deduce the degree of subsurface saturation to model rainfall-induced landslide occurrences[12,16,18]. For example, SHALSTAB[12,27] uses the topographic wetness index, proposed by Beven and Kirkby[65], to calculate subsurface saturation considering the convergence of shallow subsurface flow from up-slope drainage areas for a given steady-state precipitation. On the other hand, the Transient Rainfall Infiltration and Grid based Regional Slope stability model (TRIGRS)[16,18] calculates transient pore pressure development due to vertical rainfall infiltration from rainfall intensity. In reality, both subsurface convergence and rainfall infiltration are essential contributors to subsurface saturation and need to be implemented in physically-based slope stability models. However, measuring precipitation intensity, moisture availability, or subsurface convergence and saturation in the field is difficult, especially in rural mountainous areas with limited accessibility.

According to our SNN model results, the most important, controlling features for landslides in this area are the product of slope and *MAP* (N-S region) or that of slope and *NEE* (NW-SE and E-W regions). This result implies that local precipitation infiltration influenced by precipitation rate and intensity, represented by *MAP* and *NEE*, may serve as a first-order control on *W* or *k(W)* in Eq. (4). The absence of drainage area or discharge as a dominant contributing feature to susceptibility may suggest that subsurface flow convergence may be a second-order contributor to landslides in the easternmost Himalaya. However, we cannot rule out the possibility that the importance of topographic convergence was masked due to the low-resolution of our input topographic and rainfall data[66]. These factors can be further examined in future studies using high-resolution topographic and climate data in SNN models.

Nonetheless, identifying the exact trigger for a landslide requires dense field measurements and historic records of soil, hydrologic, and climatic conditions (e.g., soil moisture, antecedent rainfall, rainfall intensity)[9,67], which are often difficult to obtain, especially in rural mountainous areas with limited accessibility. We have shown that our SNN model can identify key controls and quantify their potential contributions to susceptibility, highlighting the essence of strong slope-climate coupled controls on landslide occurrences. The composite features identified by the SNN such as *NEE***Slope* or *MAP***Slope* are consistent with previous understandings of landslide mechanisms. However, they were not explicitly implemented in previous data-driven statistical models. In DNNs, such couplings would likely be identified, but if that were the case, the information would be implicitly contained in the network weights and not readily available to the user. By incorporating climatic composite features including *MAP***Slope*, *NEE***Slope*, and *Asp***Relief*, the performance of the SNN improved, increasing average AUROC by 5–22% compared to those of statistical or physically-based models[12,27,60,61] (Supplementary Note 2, Supplementary Table 3). This performance enhancement is statistically significant according to our confidence interval estimates from a 10-fold cross validation.

Implications, limitations, and future directions. Our work presents a substantial advance in XAI applications to natural hazards and circumvents the "black box" nature of common AI models. SNNs provide quantitative analyses of controlling factors and further highlight the important, mechanistic interpretations of landslides. Our AI-based decision-making approach provides a comprehensive framework that allows for the examination of numerous composite features and identification of key controls while retaining high accuracy. As natural perturbations increase due to urban development and climate change, the SNN may provide a promising, data-driven predictive tool that will enable communities to confidently tailor plans for hazard mitigation.

While a variety of explainable AI methods are available today, our proposed SNN method offers unique advantages that are not simultaneously present in any other method. SNN is a fully explainable model that achieves a level of explainability comparable to linear regression, while delivering state-of-the-art performance that matches that of black box models like deep neural networks. Furthermore, unlike other additive models, SNN can incorporate multivariate functions without compromising full explainability. Additionally, the model features adaptive optimization of both feature selection and network architecture during training. A comprehensive comparison of SNN with other explainable AI methods must take all of these factors into account. This requires an in-depth study beyond the scope of this paper. For instance, other additive model methods generally rely on fixed architectures and preselected feature sets that lack feature interactions beyond bivariate interactions. On the other hand, decision trees utilize highly nonlinear interactions between multiple features through a different approach that theoretically offers full explainability, but is often difficult to interpret for large number of features or complex problems requiring numerous branches. It is also worth noting that SNN is not restricted to MST as the teacher model, and its accuracy can be further improved when more accurate teacher models are found. A viable alternative to MST for applications with small datasets is random forest, which is an ensemble of decision trees trained on randomly selected feature and dataset subsets using bootstrapping. While decision trees are explainable, random forest is considered a black box since its outcome is an aggregate of multiple trees. In such cases, SNN can leverage random forest as a teacher model to achieve similar accuracy while maintaining full explainability.

We acknowledge that the overall importance of slope and climatic features and their functional relationships with susceptibility revealed by the SNN are qualitatively similar to those inferred from statistical models. However, the SNN is more useful for landslide susceptibility assessment because it decouples individual feature contributions and quantifies absolute contributions from features and feature interdependencies. For example, the relative and absolute importance of SNN decoupled features are different from those determined by the weights set by logistic regression. In addition, our analysis shows that S_j corresponding to $LR = 1$ differs depending on a feature's absolute, decoupled contribution to landslide susceptibility. The SNN approach reveals the important coupling between slope and climatic factors (e.g., *MAP***Slope*, *NEE***Slope*) as a primary driver for landslide occurrence. Accounting for these under-appreciated features and feature interdependencies that are not generally implemented in statistical methods or physically-based models can lead to a substantial increase in performance. We note that these results are specific to the region analyzed herein (easternmost Himalaya), and other regions may feature a different set of dominant factors.

We acknowledge there are limitations of our method in the easternmost Himalaya. Our input features are averaged over time and space, making it impossible to relate them directly to specific events (e.g., intense rainstorms or earthquakes) inducing landslides in our inventories. In addition, our inventory is based on optical satellite images acquired at a specific time (e.g., 2017 Landsat) and post-failure spectral signatures. Thus, our model lacks information about the precise timing or types of landslides (e.g., fast- or slow-moving landslides, soil or bedrock landslides). This makes it difficult to assess the timescales and spatial dependencies of landslide-triggering events (e.g., rainfall intensity or duration) for specific landslides or landslide types. Previous studies from the Nepal Himalaya suggest that the spatial distribution of landslides can vary with triggering events such as cloud outbursts, flooding and large-magnitude earthquakes[68,69].

However, for this study region, our method properly captures the first-order climatic controls of landslide occurrences. Our primary feature datasets may capture a representative, spatial distribution of landslide-triggering events such as intense precipitation and rock damage over the decadal timescale of concern. In the easternmost Himalaya, both MAP and NEE from TRMM and APHRODITE datasets covering 12 and 50 years show similar southward increasing trends[49,70]. This spatial pattern likely emerges from the aggregation of intense precipitation events influenced by orographic precipitation[49]. In the 30 years prior to the mapped inventory, there were no earthquakes with a magnitude larger than M_W 5.0 (Incorporated Research Institutions for Seismology, www.iris.edu), which can induce abundant landslides. In future studies, a time-series landslide inventory from multiple years and information on nonrepresentative or infrequent extreme events can be used to assess the spatial and temporal correspondence between triggering events and landslides[69].

Additionally, landslide and input feature data have relatively coarse spatial resolutions and are based on limited temporal information (e.g., 30 m resolution Landsat satellite images from 2017[71], 90 m resolution SRTM DEM[71], and ~5 km^2 resolution TRMM data over 12 years[49]). We do not have access to high-quality, high-resolution data of topography, surface materials (e.g., soil depth, bedrock structures, lithology), and climatic and ecohydrologic conditions (e.g., landslide-triggering storm intensity, time-series precipitation intensity, vegetation types). Due to the extremely rugged mountains in the Himalaya, the highest available DEM resolution without extensive data gaps, suitable for regional-scale landslide susceptibility analysis, is 90 m[9,10]. Also, there are no readily available time-series precipitation data with a resolution < 5 km^2 in this area. We used relatively coarse 30 m resolution Landsat images to map landslides even though limited high-resolution satellite imagery is available (e.g., Planetscope Scene). This is because: (1) Landsat images are globally available, open-source satellite images with a ~40-year historic archive, (2) reliable topographic, climatic, and geologic feature data have coarser resolutions than 30 m, and (3) we cover a large region of the easternmost Himalaya (a total area of 4.19×10^9 m^2, 4.66×10^6 pixels at 30 m). When applying a regional-scale model covering a large area with limited input data resolution and high computational costs, the use of 30 m resolution imagery for our model was inevitable. Although our inventory is based on coarse 30 m resolution Landsat images, our landslide inventory adequately captures the regional-scale spatial distributions of landslide occurrences and provides essential information for regional-scale landslide susceptibility models (see "Methods"). However, it is possible that our results from both physically-based or data-driven models may be biased due to the inherited uncertainties and limitations of our input data that are resolution-sensitive (e.g., topographic metrics, mapped landslides).

Despite data limitations and uncertainties, our method is general and adaptable to other regions as well as sets and formats of contributing factors and available datasets. Our SNN analysis of the easternmost Himalaya alone presents an important contribution to landslide hazard studies. High mountains in Asia hold the majority of human losses due to landslides globally, according to a global analysis conducted using 2004–2016 data[1,2]. Due to the associated high risks, there have been efforts to model landslide susceptibility in the Himalayan regions based on currently available data with limited resolutions[9,20,56–58]. Our work aims to capture the regional-scale spatial distributions of landslide susceptibility, differentiate controls of landslide occurrences, and provide interpretable, empirical functional relationships between landslide controls and susceptibility. The decoupled SNN-identified functions combined with future changes in environmental conditions (e.g., extreme precipitation)[9,72] may provide a promising tool for assessing potential landslide hazards in this area. Additionally, a modified version of the semi-automatic detection algorithm can be extended further to incorporate InSAR data from sources such as Copernicus Sentinel-1 satellites alongside time-scale optical satellite imagery[73,74] to specifically detect slow-moving landslides in future studies. With these datasets, we can apply SNN methods to slow-moving landslides and assess the controls of surface deformation while accounting for temporal changes in environmental conditions[75]. Our method is easily applicable to other locations, different datasets, and other physical hazards, such as earthquakes and wildfires. The SNN is remarkably simple consisting of only two hidden layers, yet its performance rivals that of DNNs. Our SNN can also be easily updated and improved when global, open-source, high-resolution datasets and high-performance computational resources become more available in the future.

Methods

Study area. Numerous landslides in the Himalayan region come from steep topography, intense rainfall and flood events, and seismic activities[48,49,58,76,77]. In particular, the easternmost Himalaya (Fig. 3) has a high susceptibility to landslides due to the following reasons. First, this area exhibits a dramatic precipitation gradient due to moisture originating from the Bay of Bengal in the south[49–51] (Fig. 3). Previous studies have calculated daily and mean annual precipitation rates based on 90-min measurements from the Tropical Rainfall Measuring Mission (TRMM) 2B31 over 12 years (January 1998 to December 2009), with a spatial resolution of ~5 km^2[49]. According to these datasets, our region has mean annual precipitation rates (*MAP*) varying from ~7000 mm/yr in the range front to ~200 mm/yr in the hinterland[49] with the number of extreme rainfall events (*NEE*), calculated as the number of days that exceed the 90th percentile of daily rainfall rates, reaching ~13 and ~2 events/yr in the range front and hinterland, respectively[49]. The dramatic orographic patterns of precipitation magnitude and variability are also observed in the 57-yr Asian Precipitation-Highly Resolved Observational Data Integration Towards Evaluation of Water Resources project (APHRODITE)[70]. Second, this area has consistently steep slopes from the range front, where Holocene Himalayan shortening is concentrated near and along the Main Frontal Thrust, into the hinterland, which is affected by deglaciations from the last glacial maximum[78–81]. Third, this area is prone to active seismicity. The 1950 M_W 8.6 Assam earthquake, one of the largest earthquakes in the Himalayan range, struck the nearby Namche Barwa region[52]. Since 1973, this region has experienced >450 earthquakes with $M_W > 4$ according to the Incorporated Research Institutions for Seismology data archive (www.iris.edu, accessed on 10/01/2020). Many of these factors contribute to landslide occurrences in our study site.

Within the easternmost Himalaya, we selected three regions (the Dibang, Lohit, and range front regions) with varying ranges of landslide controls to test the performance and application of the SNN model (Fig. 3 and Supplementary Fig. 1). Both Dibang and Lohit regions extend from the active range front to the hinterland, from north to south and east to west, respectively. The Dibang region consists of metasedimentary rocks in the range front and crystalline rocks in the hinterland. The Lohit region is mainly composed of crystalline rocks. The active range front region is oriented in a northwest-southeast direction and mainly composed of metasedimentary rocks.

Landslide Inventory. We generated a landslide inventory of the easternmost Himalaya using a semi-automatic detection algorithm that combines manual delineation of landslide areas with an automatic detection algorithm based on convolutional neural networks (CNN)[53–55] (Fig. 4a–c; the method illustrated using a flowchart diagram in Supplementary Fig. 2). The basic procedure is as follows. We initially mapped landslides using 30 m resolution Landsat 8 imagery from November 2017 with bands 2, 3, 4, 5, and 7[71]. These satellite images were used to generate natural and false color imagery to show information of landcover types. High degrees of vegetation in the area allow for the easy detection of vegetation removal due to landslides and clear delineation of a landslide polygon. Most landslides are mapped as a combination of source and deposit, which are difficult to distinguish in coarse resolution Landsat bands. Whenever possible, we excluded debris transport or deposits and only mapped landslide scars associated with source areas. Because our landslide mapping is based on spectral signatures of post failures, our inventory likely includes both shallow, soil landslides and deep, bedrock landslides.

We only assessed regions where landslides generally have the potential to occur or be detectable. Thus, areas of topographic slope less than 0.06 and alpine areas without vegetation cover were excluded from our landslide mapping and analysis. A slope threshold of 0.06 was determined to be the minimum slope along which landslides occur based on a cumulative distribution function of slope from observed landslides in the easternmost Himalaya. Similar criteria based on terrain characteristics such as slope or local relief have been used in previous studies to constrain the area of landslide analysis[82]. Alpine areas were classified using spectral signatures representing snow cover in Landsat 8 imagery from February 2018.

Then, we used a CNN to detect landslides automatically, following previous works[53,54] (Supplementary Fig. 2). The CNN is used as a segmentation model for identifying landslides from 5 Landsat 8 bands and 7 input features (i.e., mean curvature, elevation, local relief, mean annual precipitation, slope, failure index, and wetness). The model takes a $32 \times 32 \times 12$ patch as an input, where 12 represents the sum of 5 satellite bands and 7 input features. The model produces a 32×32 binary patch as an output, where landslide pixels are given a value of 1, and non-landslide pixels are given a value of 0. The model segments a full region by dividing the region into 32×32 patches, segmenting each patch individually, then stitching the model outputs back together to obtain a fully segmented region. The training dataset was prepared by manually annotating a small percentage of each studied region to be used as the ground truth targets for training the CNN. The manually annotated areas were selected as a number of randomly distributed 50×50 pixel square sections throughout the studied regions. The manually annotated sections were selected such that half of them include landslides and half of them do not. Hundreds of 32 × 32 patches were extracted from each 50 × 50 square section to augment the size of the training dataset. Once the CNN model is trained and used to segment the full region, the result is reviewed manually by an expert and modifications are made.

We manually corrected landslides from the automatic detection method using Landsat 8 images, high-resolution satellite images from Google Earth, and a 4-band Planetscope Scene with a 3 m resolution. Manual correction is necessary because of potentially inaccurate representations of landslide areas in automatically mapped inventories. Common issues include large detected features aggregated from multiple, adjacent landslides and small detected features that are not related to landslides[82,83]. We divided aggregated features into multiple landslides following suggestions from a previous study[83]. Most landslide polygons in all study regions were checked for aggregated features, which were divided based on the spectral signatures of recent scars and debris flows shown in high-resolution imagery. We used the manually corrected, automatically mapped landslides for our final landslide inventory (referred to as semi-automated landslides)[55]. The spatial distributions and extents of landslides from our inventory are shown in Fig. 4a–c.

The manually and semi-automatically detected landslides show a good correspondence [>90% match for landslides > 4 pixels (3600 m^2)] based on object identification that examines the existence of overlapping areas. Generally, most landslides missing from the manually detected inventory are objects with a small number of pixels that are not easily and objectively detected by humans. Semi-automated landslides with ≤4 pixels comprise ~7.5% of total landslide areas. When comparing these pixels with 3 m resolution Planetscope Scene satellite images during the post-processing procedure, we found that many of these pixels are indeed small landslides showing different spectral signatures (e.g., Supplementary Fig. 3). Thus, we included these semi-automatic landslides with ≤4 pixels in our final inventory. Areas commissioned by semi-automatic detection, but not manual mapping, were ~0.1, ~0.4, and ~0.1%, while areas omitted by semi-automated detection were ~0.2, ~0.6, and ~0.1% of the N-S, NW-SE, and E-W study areas, respectively.

The area frequency distribution of our landslides from manual and semi-automatic mappings before 2017 shows a similar distribution to that of pre-2007 landslides from a nearby eastern Himalayan region that were manually mapped using 15–30 m resolution ASTER and Landsat images[48,84] (Supplementary Fig. 4). According to a global compilation of geometrical measurements and types of 4231 landslides[84], soil landslides from all examined regions including the Himalayan region do not appear to exceed an area of 100,000 m^2. Below this threshold, soil landslides tend to be dominant[48,84]. In our landslide inventory, <1% of individual landslides and <20% of total landslide area are greater than 100,000 m^2 (Supplementary Table 1). Thus, we assume that most mapped landslides are likely soil landslides. In addition, we find that more abundant small landslides detected using the semi-automated method are similar to those observed in the landslide area-frequency distribution based on high resolution imagery (~4–15 m) from an eastern Himalayan region nearby (Supplementary Fig. 4)[48]. This supports that our semi-automatically mapped landslide inventory likely includes many small landslides missed by humans that were detected by a CNN-based automatic detection algorithm.

The total number of semi-automatically mapped landslides in our inventory is 2289, whose areas range from 900 to 1.96 × 10^6 m^2 (Fig. 4a–c). The total mapped landslide area is 2.83 × 10^7 m^2, which produces a landslide density of 0.007 within the entire study area of 4.19 × 10^9 m^2 (Supplementary Table 1). Landslide density is also calculated within a 2.25 km^2 window, which is greater than the largest landslide size (1.96 km^2). Landslide densities calculated over a 2.25 km^2 window are high in the range front (maximum of 0.121) and low in the hinterland (maximum of 0.039).

Model input feature descriptions. We quantified the spatial distribution of 15 topographic, climatic, and geologic controls and used them as input features for the SNN (Supplementary Fig. 5, Supplementary Table 2). Topographic controls include aspect (the direction of topographic slope face; *Asp*), mean curvature ($Curv_M$), planform curvature, profile curvature, total curvature, distance to channel ($Dist_C$), drainage area, elevation (*Elev*), local relief calculated as an elevation range within a 2.5 km radius circular window (*Relief*), and slope. Climatic or hydrologic controls include discharge, mean annual precipitation (*MAP*), and number of extreme rainfall events (*NEE*). Last, geologic controls include the distance to lithologic boundaries (i.e., mostly faults) ($Dist_F$) and distance to the Main Frontal Thrust and suture zone ($Dist_{MFT}$). These features were selected from literatures that examined landslide occurrences in the Himalayan region[20,56–58]. We mostly used features directly measured through satellite data including a 90 m digital elevation model from the Shuttle Radar Topography Mission (SRTM)[71] and rainfall magnitude and variability from TRMM[49], as well as published regional geologic maps[79,85]. Utilizing open-source satellite data with a long-term historic archive allows anyone to easily implement our approach in other regions (e.g., Himalayan Arc) with limited accessibility, high landslide potential, and a long landslide history[1,2,9,86].

Below are the details of our data sources and methods of calculation. First, topographic variables such as slope, aspect, local relief, curvature, distance to channel, and drainage area were calculated from a 90 m SRTM digital elevation model (DEM)[71]. Although a higher-resolution 30 m DEM is available, it contains missing values within our study area. Thus, we used a 90 m DEM for calculating topographic variables. Slope was calculated as the steepest descent gradient using an 8-direction (D8) flow routing method[87]. We calculated aspect, the direction of slope face, as the angle in degrees clockwise from north given by the components of the 3-D surface normal. The surface normal was calculated using the x, y, and z components of each pixel. Local relief was calculated as the range in elevation within a 2.5 km radius circular window. We used a 2.5 km radius window because it is similar to the length scale of across-valley widths in the range front where most landslides are. Local relief at this scale allowed us to quantify the spatial variation of topographic relief relevant to landslides on these fluvial valleys. Curvature was calculated as the second derivative of the 90 m SRTM DEM. We calculated mean, planform, profile, and total curvatures using TopoToolbox 2[87,88].

To calculate distance from channel, we first determined flow direction using D8 flow routing. The flow direction was carved through topographic depressions and flat areas to avoid sinks and generate a continuous drainage system. We then imposed a minimum drainage area of 1 km^2 needed to initiate a stream before extracting a stream network based on the flow direction. Using the stream network, we calculated the distance of each pixel in the DEM to the nearest location in the stream network.

We acquired *MAP* and *NEE* from a previous study[49] that analyzed the Tropical Rainfall Measuring Mission (TRMM) 2B31 datasets from January 1998 to December 2009. Daily rainfall and *MAP* values were integrated from 90-min measurements over 12 years. To calculate *NEE*, the 90th percentile of daily rainfall total for each pixel was determined for the 12-year measurement period[49]. Only days with measured rainfall were included in calculating the probability density function. The number of days per year with a daily rainfall total above the 90th percentile was counted as *NEE*[49,89]. The resolution of the original *MAP* and *NEE* datasets in our study area is ~5 km^2, which we resampled to 30 m resolution to be consistent with the resolution of our landslide inventory. To calculate the drainage area, we first calculated D8 flow directions of stream networks and calculated the number of upstream cells that contribute to each pixel. The number of cells can then be converted into a drainage area. Discharge was calculated by summing upstream contributing cells weighted by their *MAP* to account for spatially varying precipitation patterns. Using these weights, cells with higher *MAP* values will contribute more to total discharge than cells with lower precipitation values.

Previous studies[82,90] have shown that distance to fault ruptures is a good predictor for the occurrence of earthquake-induced landslides. We do not have information on active fault planes at depth and ground peak acceleration patterns for past earthquakes in these regions. Thus, we calculated $Dist_{MFT}$ for our study regions as each pixel's Euclidean distance from the closest point on traces of the Main Frontal Thrust (MFT) and suture zones mapped by Taylor and Yin[85]. These faults represent potentially active faults in our study area[79,80]. Because the suture zone is located far to the north, $Dist_{MFT}$ largely reflects the distance to the MFT. In addition, we calculated $Dist_F$ as the Euclidean distance of each pixel from boundaries separating all lithologic units reported in[79]. We included $Dist_F$ because bedrock tends to be more damaged near major lithologic boundaries due to faulting, which may influence landslide occurrences. The Euclidean distance was calculated using ArcGIS 10.6.

SNN training method: composite features. We categorize composite features by the number of product operations involved. For example, given a problem with n original input features $x_1, x_2, \ldots x_n$, we can generate a set of $M \geq n$ composite features $\chi_1, \chi_2, \ldots \chi_M$, where Level-1 features are the single original features (first-degree monomials such as x_i) and Level-2 features are composite features equal to the product of two Level-1 features. As an example, we may form the product $x_1{*}x_2$ (second-degree monomial), where the monomials x_1 and x_2 are Level-1 features. Level-3 features are composite features consisting of a product of three Level-1 features, such as $x_1{*}x_2{*}x_3$ or $x_1 * x_2^2$, and so on, resulting in third-degree monomials. Composite features are restricted to functions that cannot be derived from another function by elementary algebraic transformations. For example, $x_1^2 * x_2^2$ and $2{*}x_1{*}x_2$ are not permitted since they can be derived from $x_1{*}x_2$ by elementary operations (namely, by squaring and scaling, respectively). In mathematics, composite features differing from each other by a finite number of elementary operations could define an equivalence class.

SNN training method: optimization. The flow diagram of the superposable neural networks (SNN) training method is presented in Fig. 2. The SNN is an additive model[91,92] with a unique architecture described by Eq. (1) and Fig. 1, and a unique training method explained here.

The method can be summarized by the following steps:

(1) Multivariate polynomial expansion: composite features are generated.
(2) Tournament ranking: an automated feature selection method we have designed for finding the features that are most relevant to the model.
(3) Multistage training (MST): a second-order deep learning technique for generating a high-performance teacher network.
(4) Fractional knowledge distillation: a technique we designed for separating the contribution of each feature to the final output.
(5) Parallel knowledge distillation: standard knowledge distillation individually applied to networks corresponding to each feature.

(6) Network superposition: merging single layer networks corresponding to each feature into one SNN.

The two stages of knowledge distillation are key in facilitating the optimization of the highly constrained SNN architecture in a way that maximizes accuracy while minimizing the number of neurons for optimal model simplicity. The multi stage training (MST) DNN used as the teacher model due to its high performance and regularization properties, was tuned to minimize the difference between training and testing accuracy to guide the SNN model into a regularized solution that avoids over-fitting. The steps are further explained in detail below.

SNN training method: multivariate polynomial expansion. Given n features $x_1, x_2, \ldots, x_n$, we generate M composite features $\chi_1, \chi_2, \ldots, \chi_M$ according to a pre-determined maximum composite feature level.

Example 1. If the original number of features is 3 and the maximum composite feature level is Level-3, then we generate 13 composite features $[\chi_1, \chi_2, \cdots, \chi_{13}] = [x_1, x_2, x_3, x_1 * x_2, x_1 * x_3, x_2 * x_3, x_1 * x_2 * x_3, x_1^2 * x_2, x_1^2 * x_3, x_2^2 * x_1, x_2^2 * x_3, x_3^2 * x_1, x_3^2 * x_2]$.

In this work, we have used 15 original features with a maximum composite feature Level-2. Because Level-3 performs marginally better than Level-2, we consider the Level-2 SNN as our optimal SNN. With 15 original features and the maximum composite feature Level-2, we generate a total 120 composite features. All features are standardized with zero-mean and unit-variance. The Level-1 SNN inputs are single features, and the Level-2 SNN inputs are single and composite features. The SNN output is the estimated total landslide susceptibility (S_t) at a specific location, which is the sum of the susceptibility contributions from all individual features. Our optimization approach allows for the exploration of multiple combinations of parameters (e.g., 120 composite features for Level-2) without relying on an expert's choices, preconditions, or classifications of input features. The initial set of potentially relevant features is determined by the tournament ranking step. The most relevant features are then iteratively determined during the training process, where the contribution of each control to susceptibility (S_j, where j corresponds to a single or composite feature) is quantified using multiple steps of knowledge distillation. By superposing S_j, we produced (pixel-by-pixel) the total landslide susceptibility map, S_t, with values ranging from 0 to 1 as the final product (Fig. 2).

SNN training method: tournament ranking. Our feature selection technique is based on a point system and uses a combination of backwards elimination and forward selection[93] as building blocks. The composite features generated in the previous steps are randomly arranged into groups, with each group containing a subset of the features. Each feature group is used to train a simple neural network model. After the network is trained, backwards elimination is applied using area under the receiver operating characteristic curve (AUROC) as the performance criterion (Supporting Information). The top performing feature in the group receives a point. This process is repeated many times; several thousand groups were generated in the training of each SNN in this work. Features are ranked according to the points they accumulated. Forward selection is then applied in the order of the feature ranking to select the features that will be passed on to the next step.

The second-order Levenberg-Marquardt algorithm[45] was used in training the individual neural networks models. It should be noted that using second-order training is essential for the practicality of this step. Unlike first-order training algorithms (based on gradient descent) that require manual hyper parameter tuning, second-order training algorithms are robust. In addition, second-order training can achieve better performance with fewer parameters[45,94–99]. This allows for the automation of the process, and reduces the memory requirements for training the networks, yielding a more efficient parallel implementation on multicore processors.

SNN training method: multistage training. The high-ranked features that are passed on from the previous step are used to train a high-performance DNN. We chose MST as our DNN model, since it has shown superior performance in similar applications as well as regularization properties that counteracts over-fitting[42–44].

SNN training method: fractional knowledge distillation. Knowledge distillation is a technique to reduce model complexity, by using the soft output of a more complex teacher DNN as the target of a less complex student DNN[46]. The MST in the previous step acts as our teacher network.

We have designed a variation of knowledge distillation that allows us to isolate the contribution of each feature to the estimated output. We call this variation fractional knowledge distillation (FKD), a term that is inspired by the fractional distillation technique in chemistry. We illustrate this using a step-by-step example for the case of two features. This can be easily generalized to any number of features.

Example 2. Assume that two composite features $[\chi_1, \chi_2]$ are passed on from the feature selection stage, and ordered according to importance where χ_1 is the most important. Let ts_0 be the set of soft targets obtained from the MST output:

(1) Save a copy of ts_0, named ts_{0c}
(2) Train a simple DNN $net_{1,1}$ using only χ_1 as input and ts_0 as an output
(3) Obtain $o_{1,1}$, the set of outputs of $net_{1,1}$
(4) Update ts_0 to $ts_0 - o_{1,1}$
(5) Train a simple DNN $net_{2,1}$ using only χ_2 as input and ts_0 as an output
(6) Obtain $o_{2,1}$, the set of outputs of $net_{2,1}$
(7) Update ts_0 to $ts_0 - o_{2,1}$
(8) Evaluate performance by calculating AUROC using $\sum_{i=1}^{2}\sum_{j=1}^{1} o_{i,j}$ and ts_{0c}
(9) Train a simple DNN $net_{1,2}$ using only χ_1 as input and ts_0 as an output
(10) Obtain $o_{1,2}$, the set of outputs of $net_{1,2}$
(11) Update ts_0 to $ts_0 - o_{1,2}$
(12) Train a simple DNN $net_{2,2}$ using only χ_2 as input and ts_0 as an output
(13) Obtain $o_{2,2}$, the set of outputs of $net_{2,2}$
(14) Update ts_0 to $ts_0 - o_{2,2}$
(15) Evaluate performance by calculating AUROC using $\sum_{i=1}^{2}\sum_{j=2}^{2} o_{i,j}$ and ts_{0c}
(16) Repeat n times until the performance stops improving

Each DNN above consists of only a few neurons and is trained for a small number of epochs where the contribution of each feature is gradually determined to avoid numerical instabilities. The number of neurons and epochs are hyper parameters that can be tuned based on the data. We note that the DNNs in the FKD are part of the optimization, not the final optimized SNN model. Each DNN has a single input and a single output. The function of the input relative to the output is completely known in this case regardless of the model. Each feature has multiple DNNs that gradually improve its functional relationship to the output in an aggregate manner. This enables grouping all the DNNs corresponding to a specific feature and adding their outcome (the aggregate DNN model will still have a single input and a single output). Once the functional relationship of the feature is learned, it's distilled to a single transparent layer in the next step which is what is actually used in the optimized SNN model, making it fully interpretable.

SNN training method: parallel knowledge distillation. The outputs from groups of networks, corresponding to each feature from the previous step, are added together to yield one soft target per feature. Knowledge distillation is separately used to train a single SNN layer for each feature.

Example 3. Following the previous example:

(1) Create two soft targets: $ts_1 = \sum_{j=1}^{n} o_{1,j}$, and $ts_2 = \sum_{j=1}^{n} o_{2,j}$
(2) Train a single layer network net_1 using χ_1 as input and ts_1 as an output
(3) Train a single layer network net_2 using χ_2 as input and ts_2 as an output

SNN training method: network superposition. The single layer networks from the previous step are merged together to create the SNN, by adding an output layer that sums up the outputs of all the networks from the previous step. The connection weights at the output layer are set to one. The output of the SNN is a continuous value between 0 and 1, which determines the network's estimation of landslide susceptibility at a specific location.

Example 4. Following the previous example, an SNN is created with χ_1 and χ_2 as inputs and $O = o_1 + o_2$ as the output, where o_1 is the output of net_1 and o_2 is the output of net_2.

SNN training method: implementation. In this work, we have created three SNNs for three regions. The data samples from each region were partitioned into roughly 70% for training and 30% for testing. All reported performance metric results in the paper were obtained using the testing portion of the data. Class imbalance was taken into consideration when training the networks. Given that the percentage of positive targets (locations containing a landslide) in each region is substantially smaller than negative targets (locations with no landslide), positive targets were weighted higher than negative targets in the training cost functions following the approach in ref. [44].

Pythagorean tiling. While applying the SNN to landslide susceptibility modeling, we aimed to satisfy a number of conditions: (1) Full model interpretability, both locally and globally. (2) Minimizing the number of features included in the model. (3) Maximizing prediction accuracy. (4) Optimizing generalizability, such that the model is equally representative across each region.

Due to the nature of this application, special attention should be paid to the last requirement. The standard practice in ML is to divide available data into two main partitions. One partition is used for training/validation (typically 70% of the data) and the other one for testing (typically 30% of the data). Traditionally, the goal is to maximize the reported accuracy of the testing partition where to a certain extent, over-fitting in the training portion of the data is not a primary concern. A key difference in this application is that a model generated for a certain region must be equally representative of and applicable to the entire region after training, both in accuracy and explainability. To meet this requirement, we use a special data partitioning technique that utilizes Pythagorean tiling to divide our data in a spatially representative manner that maintains variability between training and testing partitions. Using Pythagorean tiling, we generate a checkerboard pattern with a 70/30% square ratio, where bigger squares correspond to training and smaller squares correspond to testing (Fig. 5). Instead of primarily aiming to obtain the highest accuracy on the testing portion of the data, our algorithm is designed to find a more conservative solution with optimal balance between maximizing testing accuracy and minimizing the difference between training and testing accuracies.

Data availability

The manual and semi-automatically mapped landslide inventories and environmental control datasets used within this manuscript are provided through the UCLA Dataverse: https://doi.org/10.25346/S6/D5QPUA.

Code availability

The SNN code associated with this paper is available on GitHub: https://github.com/GeoSNN/GeoSNN.git. The associated DOI is: https://doi.org/10.5281/zenodo.7833891.

Received: 22 September 2022; Accepted: 17 April 2023;
Published online: 10 May 2023

References

1. Petley, D. Global patterns of loss of life from landslides. *Geology* **40**, 927–930 (2012).
2. Froude, M. & Petley, D. Global fatal landslide occurrence from 2004 to 2016. *Nat. Hazards Earth Syst. Sci.* **18**, 2161–2181 (2018).
3. Huang, R. & Fan, X. The landslide story. *Nat. Geosci.* **6**, 325–326 (2013).
4. Fan, X. et al. Earthquake-induced chains of geologic hazards: patterns, mechanisms, and impacts. *Rev. Geophys.* **57**, 421–503 (2019).
5. Tien Bui, D., Pradhan, B., Lofman, O., Revhaug, I. & Dick, O. Landslide susceptibility assessment in the Hoa Binh province of Vietnam: a comparison of the Levenberg–Marquardt and Bayesian regularized neural networks. *Geomorphology* **171**, 12–29 (2012).
6. Tien Bui, D. et al. Shallow landslide prediction using a novel hybrid functional machine learning algorithm. *Remote Sens.* **11**, 931 (2019).
7. Phong, T. et al. Landslide susceptibility modeling using different artificial intelligence methods: a case study at Muong Lay district, Vietnam. *Geocarto Int.* **36**, 1685–1708 (2021).
8. Dikshit, A., Pradhan, B. & Alamri, A. M. Pathways and challenges of the application of artificial intelligence to geohazards modelling. *Gondwana Res.* **100**, 290–301 (2021).
9. Kirschbaum, D., Kapnick, S., Stanley, T. & Pascale, S. Changes in extreme precipitation and landslides over High Mountain Asia. *Geophys. Res. Lett.* **47**, e2019GL085347 (2020).
10. Stanley, T. & Kirschbaum, D. B. A heuristic approach to global landslide susceptibility mapping. *Nat. Hazards* **87**, 145–164 (2017).
11. Dietrich, W., Reiss, R., Hsu, M. & Montgomery, D. A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. *Hydrol. Process.* **9**, 383–400 (1995).
12. Montgomery, D. & Dietrich, W. A physically based model for the topographic control on shallow landsliding. *Water Resour. Res.* **30**, 1153–1171 (1994).
13. Montgomery, D., Sullivan, K. & Greenberg, H. Regional test of a model for shallow landsliding. *Hydrol. Process.* **12**, 943–955 (1998).
14. Radbruch-Hall, D. *Landslide Overview Map of the Conterminous United States*, vol. 1183 (US Government Printing Office, 1982).
15. Guzzetti, F., Carrara, A., Cardinali, M. & Reichenbach, P. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology* **31**, 181–216 (1999).
16. Baum, R., Savage, W. & Godt, J. TRIGRS—a Fortran program for transient rainfall infiltration and grid-based regional slope-stability analysis. *US Geol. Surv. open-file Rep.* **424**, 38 (2002).
17. Meunier, P., Hovius, N. & Haines, J. Topographic site effects and the location of earthquake induced landslides. *Earth Planet. Sci. Lett.* **275**, 221–232 (2008).
18. Baum, R. L., Godt, J. W. & Savage, W. Z. Estimating the timing and location of shallow rainfall-induced landslides using a model for transient, unsaturated infiltration. *J. Geophys. Rese.: Earth Surface* **115**, F03013 (2010).
19. Lee, S. & Sambath, T. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. *Environ. Geol.* **50**, 847–855 (2006).
20. Regmi, A. et al. Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in central Nepal Himalaya. *Arab. J. Geosci.* **7**, 725–742 (2014).
21. Van Dao, D. et al. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *Catena* **188**, 104451 (2020).
22. Tien Bui, D., Tsangaratos, P., Nguyen, V.-T., Van Liem, N. & Trinh, P. Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. *Catena* **188**, 104426 (2020).
23. 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).
24. Gómez, H. & Kavzoglu, T. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. *Eng. Geol.* **78**, 11–27 (2005).
25. Lee, S., Ryu, J.-H., Won, J.-S. & Park, H.-J. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. *Eng. Geol.* **71**, 289–302 (2004).
26. Stanley, T. et al. Building a landslide hazard indicator with machine learning and land surface models. *Environ. Model. Softw.* **129**, 104692 (2020).
27. Dietrich, W., Bellugi, D. & Real De Asua, R. Validation of the shallow landslide model, SHALSTAB, for forest management. *Water Sci. Application* **2**, 195–227 (2001).
28. Reichenbach, P., Rossi, M., Malamud, B., Mihir, M. & Guzzetti, F. A review of statistically-based landslide susceptibility models. *Earth-Sci. Rev.* **180**, 60–91 (2018).
29. Pradhan, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers Geosci.* **51**, 350–365 (2013).
30. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
31. Gunning, D. et al. XAI—Explainable artificial intelligence. *Sci. Robot.* **4**, p.eaay7120 (2019).
32. Adadi, A. & Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).
33. Cui, Y. et al. The cost of rapid and haphazard urbanization: lessons learned from the Freetown landslide disaster. *Landslides* **16**, 1167–1176 (2019).
34. European Commission. White paper on artificial intelligence–a European approach to excellence and trust (2020).
35. Li, X.-H. et al. A survey of data-driven and knowledge-aware eXplainable AI. *IEEE Trans. Knowl. Data Eng.* **34**, 29–49 (2022).
36. Leiva, R. G., Anta, A. F., Mancuso, V. & Casari, P. A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design. *IEEE Access* **7**, 99978–99987 (2019).
37. Hastie, T. & Tibshirani, R. *Generalized Additive Models*, vol. 43 (CRC press, 1990).
38. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* (Routledge, 2017).
39. Friedman, J. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
40. Agarwal, R. et al. Neural additive models: Interpretable machine learning with neural nets. *Adv. Neural. Inf. Process. Syst.* **34**, 4699–4711 (2021).
41. Lundberg, S. M. & Lee, S. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
42. Youssef, K., Jarenwattananon, N. & Bouchard, L.-S. Feature-preserving noise removal. *IEEE Trans. Med. Imaging* **34**, 1822–1829 (2015).
43. Bouchard, L.-S. & Youssef, K. Feature-preserving noise removal. US Patent 9,953,246 (2018).
44. Youssef, K. et al. Machine learning approach to rf transmitter identification. *IEEE J. Radio Frequency Identif.* **2**, 197–205 (2018).
45. Yu, H. & Wilamowski, B. Levenberg-Marquardt training. *Ind. Electron. Handb.* **5**, 1 (2011).
46. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at https://arxiv.org/abs/1503.02531 (2015).
47. Tan, S., Caruana, R., Hooker, G. & Lou, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proc. 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 303–310. arXiv:1710.06169 [stat.ML] (2018).
48. Larsen, I. & Montgomery, D. Landslide erosion coupled to tectonics and river incision. *Nat. Geosci.* **5**, 468–473 (2012).
49. Bookhagen, B. & Burbank, D. W. Toward a complete Himalayan hydrological budget: spatiotemporal distribution of snowmelt and rainfall and their impact on river discharge. *J. Geophys. Res. Earth Surf.* **115**, F03019 (2010).
50. Barros, A., Kim, G., Williams, E. & Nesbitt, S. Probing orographic controls in the Himalayas during the monsoon using satellite imagery. *Nat. Hazards Earth Syst. Sci.* **4**, 29–51 (2004).
51. Yang, Y., Zhao, T., Ni, G. & Sun, T. Atmospheric rivers over the Bay of Bengal lead to extreme northern Indian rainfall. *Int. J. Climatol.* **38**, 1010–1021 (2018).
52. Ben-Menahem, A., Aboodi, E. & Schild, R. The source of the great Assam earthquake—an interplate wedge motion. *Phys. Earth Planet. Inter.* **9**, 265–289 (1974).

53. Ghorbanzadeh, O. et al. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sens.* **11**, 196 (2019).
54. Prakash, N., Manconi, A. & Loew, S. Mapping landslides on EO data: Performance of deep learning models vs. traditional machine learning models. *Remote Sens.* **12**, 346 (2020).
55. Shao, K., Youssef, K., Moon, S. & Bouchard, L. S. Landslide inventories and environmental control datasets. *UCLA Dataverse* https://doi.org/10.25346/S6/D5QPUA (2023).
56. Devkota, K. et al. Landslide susceptibility mapping using certainty factor, index of entropy and logistic regression models in GIS and their comparison at Mugling–Narayanghat road section in Nepal Himalaya. *Nat. Hazards* **65**, 135–165 (2013).
57. Mandal, S. & Mandal, K. Modeling and mapping landslide susceptibility zones using GIS based multivariate binary logistic regression (LR) model in the Rorachu river basin of eastern Sikkim Himalaya, India. *Model. Earth Syst. Environ.* **4**, 69–88 (2018).
58. Chowdhuri, I. et al. Torrential rainfall-induced landslide susceptibility assessment using machine learning and statistical methods of eastern Himalaya. *Nat. Hazards* **107**, 697–722 (2021).
59. Moon, S. et al. Climatic control of denudation in the deglaciated landscape of the Washington Cascades. *Nat. Geosci.* **4**, 469–473 (2011).
60. Lee, S. Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. *Int. J. Remote Sens.* **26**, 1477–1491 (2005).
61. Akgun, A. A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at Izmir, Turkey. *Landslides* **9**, 93–106 (2012).
62. Iverson, R. Landslide triggering by rain infiltration. *Water Resour. Res.* **36**, 1897–1910 (2000).
63. Meunier, P., Hovius, N. & Haines, J. Topographic site effects and the location of earthquake induced landslides. *Earth Planet. Sci. Lett.* **275**, 221–232 (2008).
64. Huang, A.-L. & Montgomery, D. Topographic locations and size of earthquake-and typhoon-generated landslides, Tachia River, Taiwan. *Earth Surf. Process. Landf.* **39**, 414–418 (2014).
65. Beven, K. & Kirkby, M. A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. J.* **24**, 43–69 (1979).
66. Leonarduzzi, E., Maxwell, R., Mirus, B. B. & Molnar, P. Numerical analysis of the effect of subgrid variability in a physically based hydrological model on runoff, soil moisture, and slope stability. *Water Resour. Res.* **57**, e2020WR027326 (2021).
67. Orland, E., Roering, J., Thomas, M. & Mirus, B. Deep learning as a tool to forecast hydrologic response for landslide-prone hillslopes. *Geophys. Res. Lett.* **47**, e2020GL088731 (2020).
68. Jones, J. N., Boulton, S. J., Stokes, M., Bennett, G. L. & Whitworth, M. R. 30-year record of Himalaya mass-wasting reveals landscape perturbations by extreme events. *Nat. Commun.* **12**, 1–15 (2021).
69. Jones, J. N., Boulton, S. J., Bennett, G. L., Stokes, M. & Whitworth, M. R. Temporal variations in landslide distributions following extreme events: implications for landslide susceptibility modeling. *J. Geophys. Res.: Earth Surf.* **126**, e2021JF006067 (2021).
70. Yatagai, A. et al. APHRODITE: constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *Bull. Am. Meteorol. Soc.* **93**, 1401–1415 (2012).
71. United States Geological Survey EarthExplorer. accessed September 26, 2020, https://earthexplorer.usgs.gov/ (2020).
72. Stanley, T. A. et al. Building a landslide hazard indicator with machine learning and land surface models. *Environ. Model. Softw.* **129**, 104692 (2020).
73. Bekaert, D. P., Handwerger, A. L., Agram, P. & Kirschbaum, D. B. InSAR-based detection method for mapping and monitoring slow-moving landslides in remote regions with steep and mountainous terrain: an application to Nepal. *Remote Sens. Environ.* **249**, 111983 (2020).
74. Singh, S., Raju, A. & Banerjee, S. Detecting slow-moving landslides in parts of Darjeeling–Sikkim Himalaya, NE India: quantitative constraints from PSInSAR and its relation to the structural discontinuities. *Landslides* **19**, 2347–2365 (2022).
75. Finnegan, N. J., Perkins, J. P., Nereson, A. L. & Handwerger, A. L. Unsaturated flow processes and the onset of seasonal deformation in slow-moving landslides. *J. Geophys. Res.: Earth Surf.* **126**, e2020JF005758 (2021).
76. Coudurier-Curveur, A. et al. A composite rupture model for the great 1950 Assam earthquake across the cusp of the East Himalayan Syntaxis. *Earth Planet. Sci. Lett.* **531**, 115928 (2020).
77. Kent, W. & Dasgupta, U. Structural evolution in response to fold and thrust belt tectonics in northern Assam. A key to hydrocarbon exploration in the Jaipur anticline area. *Mar. Pet. Geol.* **21**, 785–803 (2004).
78. Burgess, W., Yin, A., Dubey, C., Shen, Z.-K. & Kelty, T. Holocene shortening across the Main Frontal Thrust zone in the eastern Himalaya. *Earth Planet. Sci. Lett.* **357**, 152–167 (2012).
79. Haproff, P. et al. Geologic framework of the northern Indo-Burma ranges and lateral correlation of Himalayan-Tibetan lithologic units across the eastern Himalayan syntaxis. *Geosphere* **15**, 856–881 (2019).
80. Haproff, P., Odlum, M., Zuza, A., Yin, A. & Stockli, D. Structural and thermochronologic constraints on the Cenozoic tectonic development of the northern Indo-Burma Ranges. *Tectonics* **39**, e2020TC006231 (2020).
81. Salvi, D., Mathew, G., Kohn, B., Pande, K. & Borgohain, B. Thermochronological insights into the thermotectonic evolution of Mishmi Hills across the Dibang Valley, NE Himalayan Syntaxis. *J. Asian Earth Sci.* **190**, 104158 (2020).
82. Parker, R. et al. Mass wasting triggered by the 2008 Wenchuan earthquake is greater than orogenic growth. *Nat. Geosci.* **4**, 449–452 (2011).
83. Marc, O. & Hovius, N. Amalgamation in landslide maps: effects and automatic detection. *Nat. Hazards Earth Syst. Sci.* **15**, 723–733 (2015).
84. Larsen, I., Montgomery, D. & Korup, O. Landslide erosion controlled by hillslope material. *Nat. Geosci.* **3**, 247–251 (2010).
85. Taylor, M. & Yin, A. Active structures of the Himalayan-Tibetan orogen and their relationships to earthquake distribution, contemporary strain field, and Cenozoic volcanism active structures on the Tibetan Plateau and surrounding regions. *Geosphere* **5**, 199–214 (2009).
86. Zhu, Z. et al. Benefits of the free and open landsat data policy. *Remote Sens. Environ.* **224**, 382–385 (2019).
87. Schwanghart, W. & Scherler, D. TopoToolbox 2–MATLAB-based software for topographic analysis and modeling in Earth surface sciences. *Earth Surf. Dyn.* **2**, 1–7 (2014).
88. Schmidt, J., Evans, I. & Brinkmann, J. Comparison of polynomial models for land surface curvature calculation. *Int. J. Geographical Inf. Sci.* **17**, 797–814 (2003).
89. Bookhagen, B. Appearance of extreme monsoonal rainfall events and their impact on erosion in the Himalaya. *Geomat., Nat. Hazards Risk* **1**, 37–50 (2010).
90. Xu, C., Xu, X., Yao, X. & Dai, F. Three (nearly) complete inventories of landslides triggered by the May 12, 2008 Wenchuan Mw 7.9 earthquake of China and their spatial distribution statistical analysis. *Landslides* **11**, 441–461 (2014).
91. Hooker, G. Discovering additive structure in black box functions. In *Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, 575–580 (2004).
92. Tan, S., Caruana, R., Hooker, G., Koch, P. & Gordo, A. Learning global additive explanations for neural nets using model distillation. Preprint at http://arxiv.org/abs/1801.08640 (2018).
93. Xu, L. & Zhang, W.-J. Comparison of different methods for variable selection. *Analytica Chim. Acta* **446**, 475–481 (2001).
94. Ozyildirim, B. & Kiran, M. Do optimization methods in deep learning applications matter? Preprint at https://arxiv.org/abs/2002.12642 (2020).
95. Le, Q. et al. On optimization methods for deep learning. In *Proc. 28th International Conference on Machine Learning, Bellevue, WA, USA*, 265–272 (2011).
96. Battiti, R. First- and second-order methods for learning: between steepest descent and newton's method. *Neural Comput.* **4**, 141–166 (1992).
97. Tan, H. & Lim, K. Review of second-order optimization techniques in artificial neural networks backpropagation. *IOP Conf. Ser.: Mater. Sci. Eng.* **495**, 012003 (2019).
98. Montavon, G., Orr, G. & Müller, K.-R. *Neural Networks: Tricks of the Trade*, vol. 7700 (Springer, 2012).
99. Wilamowski, B. & Yu, H. Improved computation for Levenberg-Marquardt training. *IEEE Trans. Neural Netw.* **21**, 930–937 (2010).

Acknowledgements

L.-S.B. would like to thank NVIDIA for donating the Quadro P6000 GPU for this project. S.M. acknowledges the support of NSF EAR-1945431. We thank George Hilley, Dino Bellugi, Eitan Shelef, and Justin Higa for discussions and Ben Mirus, and three anonymous reviewers for constructive comments.

Author contributions

L.-S.B. initiated the collaboration with S.M., and all authors contributed to the design of the study. K.Y. developed and implemented the XAI method, and K.S. quantified landslide data, controls, and other landslide models. K.Y. and K.S. collected data and performed experiments, calculations and data analysis. All authors have critically examined the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-023-00806-5.

Correspondence and requests for materials should be addressed to S. Moon or L.-S. Bouchard.

Peer review information *Communications Earth & Environment* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Rahim Barzegar and Joe Aslin.