

Data-Driven Landslide Susceptibility Mapping: A Multimodal Deep Learning Framework for Integrating Multi-Source Geospatial and Environmental Data

Jyoti Arora
Computer Science and
Engineeirng
Chandigarh University
Mohali, India
jyotiarora689@gmail.com

Anupam Mittal
Computer Science and
Engineeirng
Chandigarh University
Mohali, India
anupam.csephd@gmail.com

Geetika Sharma
Compute Science and
Engineering
Chandigarh University Mohali,
India
geetika.csephd@gmail.com

Abstract

Landslides are among the most data-intensive geohazards to characterize: their occurrence, timing, and severity depend on the interaction of terrain, hydrological, geological, and anthropogenic variables that must be sensed, curated, labelled, and modelled before any hazard statement can be made. This paper presents a data-centred machine learning framework for landslide susceptibility mapping (LSM) that treats data explicitly as the input fuel, the test material, the validation instrument, and the final output of the analysis. We review the principal open data streams that underpin contemporary susceptibility modelling such as global and regional landslide inventories (e.g., the NASA Cooperative Open Online Landslide Repository), digital elevation models, precipitation records, land-cover products, and soil information and we examine the well-documented data-quality problems (reporting bias, positional uncertainty, class imbalance, spatial autocorrelation) that condition what a model trained on them can honestly claim. To make the pipeline concrete and reproducible, we construct a literature-informed synthetic multi-source conditioning-factor dataset spanning a 21,600-cell mountainous terrain and use it to train and evaluate five commonly used classifiers such as logistic regression, random forest, gradient boosting, support vector machines, and a multilayer perceptron which are under two validation regimes: conventional random k-fold splitting and spatial block cross-validation. Consistent with prior findings in the susceptibility-mapping literature, random forest and neural-network models achieved the strongest discrimination (AUC 0.76–0.77 under random validation), while all models showed a measurable performance decline under spatially independent validation, confirming that spatial autocorrelation inflates apparent accuracy when training and test pixels are geographically adjacent. We report standard classification metrics (AUC, accuracy, precision, recall, F1, Cohen's kappa, Brier score), a feature-importance ranking dominated by slope, rainfall, distance to roads, lithology and land cover, and a susceptibility surface in which the top two hazard classes (16.3% of the study area) capture 59.8% of landslide-labelled cells. We conclude with a discussion of data provenance, uncertainty propagation, and reproducibility considerations relevant to the Data Science Journal community, and we argue that susceptibility mapping is best understood as a data curation and validation problem as much as a modelling problem.

Keywords: landslide susceptibility mapping; machine learning; multi-source geospatial data; spatial cross-validation; data quality; digital elevation model; class imbalance; random forest

1. Introduction

Landslides are the downslope movement of rock, debris, or earth under gravity those are among the most geographically pervasive and economically damaging natural hazards, affecting mountainous and hilly terrain on every inhabited continent. Their triggers include intense or prolonged rainfall, seismic shaking, rapid snowmelt, and human disturbance of slopes through road cutting, deforestation, and construction. Because landslides are highly localized, episodic, and often under-reported, understanding where they are more or less likely to occur is fundamentally a data problem before it is a modelling problem: robust susceptibility assessment depends on the availability, quality, and integration of an inventory of past events with layers describing terrain, hydrology, geology, land cover, and human activity.

Landslide susceptibility mapping (LSM) describes the spatial estimation of the relative likelihood that a given location will experience a landslide, based on the assumption that future events are conditioned by the same factors that produced past events and has moved over the past decade from expert-weighted, qualitative overlay methods toward data-driven machine learning (ML) approaches. Ensemble tree methods such as random forest and gradient boosting, kernel methods such as support vector machines, and neural networks are now routinely reported to achieve area-under-the-curve (AUC) values in the 0.85–0.99 range across a wide variety of regional case studies. Yet this apparent success conceals a set of data-centred challenges that are frequently under-examined: landslide inventories are opportunistically collected and geographically biased; the definition of "non-landslide" training locations (pseudo-absences) is not standardized; conditioning factors are frequently collinear; and, critically, because landslides and their causal factors are spatially auto correlated, naive random train/test splitting allows information to leak between adjacent training and test pixels, inflating reported accuracy relative to what a model would achieve on a genuinely unseen slope.

This paper is deliberately organized around data rather than around any single algorithm. Its purpose is threefold. First, we synthesize, from the current literature, the principal open and semi-open data sources used to build susceptibility models like landslide inventories, digital elevation models (DEMs) and their derivatives, precipitation records, land-cover and vegetation indices, and soil and lithological data that combine with the known limitations of each. Second, we build a fully reproducible, literature-informed synthetic conditioning-factor dataset that reproduces the qualitative statistical relationships reported across published studies (e.g., higher susceptibility on steep, poorly vegetated, shale-rich slopes near roads and drainage lines during high-rainfall periods) and use it to demonstrate the complete data science pipeline: feature screening, class-imbalance handling, model training, and the most importantly is the validation under both conventional random cross-validation and spatially independent block cross-validation. Third, we interpret the resulting metrics, feature extraction, and susceptibility surface explicitly through a data lens, asking what the inputs, the test set, the validation design, and the output map can and cannot support as evidence for hazard management.

We use a synthetic rather than a single proprietary regional inventory for the empirical demonstration for three reasons that are themselves data-related. First, many published regional landslide inventories are not redistributable — several recent studies note explicitly that their landslide and geological maps are confidential and cannot be shared publicly, which limits reproducibility. Second, a synthetic dataset with a known, controllable data-generating process lets us isolate the effect of a single variable of interest — the validation scheme — while holding the feature-label relationship fixed, which is difficult to do cleanly with a single opportunistic field inventory. Third, constructing the synthetic dataset from the documented statistical structure of real conditioning factors (terrain, rainfall, lithology, land cover, soil depth) keeps the demonstration grounded in, and comparable with, the published literature, whose reported performance ranges we use throughout as an external benchmark. We return to the implications of this choice, and to the conditions under which the framework should be re-run on a primary regional inventory, in the Discussion.

The remainder of the paper is organized as follows. Section 2 reviews related work with an emphasis on the data streams and validation practices used in the field. Section 3 describes the data sources relevant to susceptibility

mapping in general and the specific illustrative dataset used in this study. Section 4 details the methodology, including feature engineering, class-imbalance handling, the five classifiers evaluated, and the two validation schemes. Section 5 reports results. Section 6 discusses the findings from a data-quality and reproducibility perspective aligned with the scope of the Data Science Journal, and Section 7 concludes.

2. Background and Related Work

2.1 From Heuristic Overlay Methods to Machine Learning

Early susceptibility assessments relied on expert-assigned weights combined through analytical hierarchy process (AHP) or fuzzy overlay methods, in which a small number of causative layers are ranked and summed into a susceptibility index. These qualitative methods are transparent but are acknowledged to be highly subjective, since different experts assign different weights to the same conditioning factors under the same conditions, which produces inconsistent results across studies of similar terrain.

Over the past decade, the field has shifted toward supervised machine learning, in which a binary landslide/non-landslide inventory is combined with a stack of conditioning-factor layers and a classifier learns the mapping between them directly from data. Random forest (RF) [3], extreme gradient boosting (XGBoost) [1] and related boosted-tree methods, support vector machines (SVM) [2], and artificial neural networks (ANN) are the most frequently benchmarked classifiers. A comparative study in Fengjie County, Chongqing, China, built a geospatial database of sixteen conditioning factors from 1,624 landslides recorded between 1980 and 2020 and found that random forest outperformed XGBoost, with an AUC of 0.866 versus a slightly lower value for XGBoost, while noting that land use, elevation, and lithology were the dominant contributing factors. A hybrid-model evaluation across twenty high-risk districts of the Western Ghats, India, compared random forest, XGBoost, LightGBM, CatBoost and a hybrid classifier against a 2015–2020 landslide inventory and reported that all models achieved accuracies above 0.90, with the hybrid classifier showing the lowest error and the greatest stability. A hybrid rough-set and ensemble study in Uttarkashi, Uttarakhand, combining rough set theory with XGBoost, achieved an AUC of 0.937 and an F1-score of 0.926, again ranking among the strongest performers in a multi-model comparison. Comparable AUC ranges of roughly 0.90 to 0.99 have been reported across a range of recent regional studies using random forest, XGBoost, or logistic regression, with performance driven primarily by the size and completeness of the underlying landslide inventory and by the informativeness of the conditioning-factor stack, particularly lithology.

These accuracy figures are important context but must be read cautiously: as we discuss in Section 2.3, most are computed under random or stratified random cross-validation, which is now understood to systematically overestimate the accuracy that a model would achieve on a truly independent slope.

2.2 Open and Semi-Open Landslide Data Streams

A recurring constraint in the literature is that primary landslide inventories are frequently proprietary or only partially released; one large factor-screening study in Yunnan, China explicitly reports that its landslide map and 1:200,000-scale geological map are confidential and cannot be shared publicly, limiting the reproducibility of that specific analysis even though the paper itself is open access. Against this backdrop, several genuinely open, redistributable data products have become important reference resources for the community. NASA's Cooperative Open Online Landslide Repository (COOLR) merges the Global Landslide Catalog (GLC), an event catalogue of rainfall-triggered landslides compiled from media reports, disaster databases and scientific literature since 2007, with citizen-science reports submitted through the Landslide Reporter application; as of the initial evaluation period the combined catalogue already exceeded 11,000 records worldwide, and citizen scientists contributed 162 new events spanning 37 countries during the project's first thirteen months alone. For deep-learning applications on optical and radar imagery, the Landslide4Sense benchmark provides 3,799 co-registered 128×128-pixel image patches that fuse twelve Sentinel-2 multispectral bands with slope and elevation layers derived from ALOS

PALSAR radar, collected across four independent geographic events (Iburi, Japan; Kodagu, India; Gorkha, Nepal; and Taiwan), with every pixel manually labelled as landslide or non-landslide; a later extension of the same benchmark expanded the labelled set to 4,884 data cubes spanning events between 2015 and 2021. These two resources are complementary: COOLR supplies point- and polygon-based event records suitable for regional susceptibility mapping, while Landslide4Sense supplies pixel-labelled imagery suitable for training segmentation models that detect and delineate landslide extent directly from satellite scenes [5].

Conditioning-factor data for susceptibility mapping is comparatively mature and standardized. Global digital elevation models — most commonly SRTM and ALOS PALSAR/AW3D30 — supply elevation, slope, aspect, curvature and topographic wetness index at 10–30 m resolution. Precipitation is commonly drawn from satellite- and station-blended products, and soil information increasingly draws on SoilGrids 2.0, which produces global maps of soil properties at 250 m resolution from roughly 240,000 point observations and more than 400 environmental covariates using machine learning with quantified spatial uncertainty, giving susceptibility modellers not just a soil-depth or texture estimate but an associated per-pixel uncertainty layer that is rarely propagated into downstream susceptibility models [4]. Land-cover and vegetation information is typically derived from Sentinel-2 or Landsat-derived NDVI composites or from global land-cover products, since vegetation cover and land-use change are consistently identified as first-order predictors of shallow-landslide susceptibility.

2.3 Validation Design and the Spatial-Autocorrelation Problem

A separate but increasingly central strand of the literature concerns how susceptibility models should be validated, because conditioning factors and landslide occurrence are spatially autocorrelated: nearby pixels tend to share similar slope, rainfall, and lithology, so a training pixel and a test pixel that happen to be neighbours are not statistically independent. When training and test data are split at random, some test pixels are geographically adjacent to training pixels, allowing information to leak across the split and inflating the apparent skill of the model. This is a well-documented phenomenon outside landslide science as well — a comparison of a random-forest model predicting above-ground forest biomass found that random splits suggested good predictive skill while spatial cross-validation, which enforces geographic separation between folds, indicated essentially no genuine predictive skill, and subsequent synthetic-data experiments have confirmed that error estimates from spatially blocked cross-validation are systematically more trustworthy than those from random splits whenever the underlying variables are auto-correlated [8].

Within landslide susceptibility mapping specifically, a comparative study using random sampling of training and test data across several classifiers found that random forest was the most accurate method under conventional cross-validation, but that its predictive performance dropped substantially once evaluated under spatial rather than random cross-validation, and a spatial-validation case study of rotational landslides in the northern Peruvian Andes similarly showed that discrimination performance decreases systematically as the spatial block size used for validation increases, indicating that conventional random validation can overestimate a model's real predictive capacity. A related methodological literature on spatial and design-based validation in digital soil mapping reaches the same conclusion: appropriate cross-validation schemes that respect the spatial (or, more generally, hierarchical) structure of the data are necessary to obtain unbiased estimates of map accuracy, and the broader issue of data leakage extends beyond cross-validation design to laboratory analysis, experimental design, and pre-processing choices more generally. These findings motivate the direct comparison of random and spatial validation that forms the empirical core of this paper.

3. Data Sources and Description

Because this paper is oriented toward the Data Science Journal's focus on data as the central object of study, this section separates (i) the landslide inventory, which supplies the binary training and validation labels; (ii) the conditioning-factor layers, which supply the model's input features; and (iii) the specific illustrative dataset

constructed for the empirical demonstration in Sections 4 and 5, together with an explicit statement of its provenance and limitations.

3.1 Landslide Inventory Data (Labels)

Landslide inventories record the location, date, trigger, size, and (where available) impact of individual landslide events, and they function as the ground-truth labels against which every conditioning factor is regressed. Two properties of inventory data dominate downstream modelling decisions. First, inventories are almost always incomplete and geographically biased: events are more likely to be recorded near populated areas, roads, and media coverage than in remote terrain, so the "non-landslide" class inferred from the absence of a record conflates true stable ground with simply unobserved landslides. Second, inventories are heterogeneous in how they represent an event — as a single point, a source-area polygon, or a full runout polygon — which affects what a pixel-level or slope-unit-level classifier is actually being trained to predict. The COOLR/GLC catalogue illustrates both properties directly: it is compiled from online news media, disaster databases, scientific reports, and citizen contributions rather than from systematic field or remote-sensing surveys, which is a source of reporting bias that the citizen-science component is explicitly designed to help mitigate, while the Landslide4Sense benchmark instead offers manually verified pixel-level polygons for a small number of well-studied triggering events, trading geographic breadth for label precision.

3.2 Conditioning Factor (Feature) Data

Table 1. Principal conditioning-factor families used in landslide susceptibility mapping, representative open data sources, and the physical mechanism each is intended to represent.

Factor family	Representative variables	Typical open source	Mechanistic role
Terrain / morphometry	Elevation, slope, aspect, plan/profile curvature, topographic wetness index (TWI)	SRTM, ALOS PALSAR/AW3D30 DEM (10–30 m)	Controls gravitational driving stress and water accumulation
Hydrology / climate	Mean annual and event rainfall, rainfall intensity-duration	Station-blended and satellite precipitation products	Primary trigger for shallow, rainfall-induced landslides
Geology	Lithology, distance to fault / lineament	National geological surveys; regional geology maps	Governs shear strength and permeability of hillslope material
Soil	Soil depth, texture, drainage class	SoilGrids 2.0 (250 m, global, ML-derived with uncertainty)	Determines infiltration and pore-pressure response
Land cover / vegetation	NDVI, land-cover class	Sentinel-2 / Landsat NDVI; global land-cover products	Root reinforcement and interception reduce susceptibility

Table 1 summarizes the conditioning-factor families most frequently used in the machine-learning susceptibility literature, together with representative open data sources and the physical or hydrological mechanism each factor is intended to capture.

3.3 Illustrative Dataset Used in This Study

To demonstrate the full data pipeline in a way that is fully reproducible and redistributable — and to avoid the licensing and confidentiality restrictions that affect several published regional inventories noted in Section 2.2 — we constructed a synthetic, literature-informed conditioning-factor dataset rather than acquiring a single proprietary regional inventory. The dataset was generated over a synthetic $36 \text{ km} \times 24 \text{ km}$ mountainous domain discretized into 21,600 cells at 200 m resolution (analogous in scale to a small regional susceptibility study). Terrain layers (elevation, slope, aspect, plan curvature, and a topographic-wetness-index proxy) were generated using a fractal (spectral-synthesis) surface model that reproduces the self-similar roughness statistics of real topography; distance-to-road, distance-to-drainage, and distance-to-fault layers were derived from Euclidean distance transforms applied to synthetic linear-feature masks; mean annual rainfall was generated as a smooth orographic field correlated with elevation, consistent with the rainfall-elevation relationships documented in satellite precipitation products; NDVI was generated as a function of slope and rainfall plus independent noise; and three categorical layers — lithology (five classes), land cover (five classes), and soil depth (three classes) — were generated as spatially clustered categorical fields with literature-informed susceptibility weights (e.g., shale/phyllite lithologies and bare or degraded land cover were assigned higher susceptibility weights than granite or dense forest, consistent with the dominant-factor findings reported for Fengjie County and comparable studies). A binary landslide label was then generated from a logistic function of a weighted combination of these factors plus an explicit unmodelled noise term, followed by multiplication with an independent spatial event-clustering field so that positive labels form spatially coherent clusters rather than being scattered independently, mimicking the storm-track and fault-zone clustering of real landslide events. The resulting dataset has a 17.4% landslide-positive base rate over the full domain, which is broadly consistent with the class-imbalance magnitude reported in field inventories once a comparable land area is considered. The full data-generating script is described in Section 4 and is available from the corresponding author on request, consistent with the reproducibility aims of this journal.

We emphasize that this is a synthetic illustrative dataset, not a primary field inventory, and the specific numerical results in Section 5 should be read as a demonstration of pipeline behaviour and of the direction and approximate magnitude of well-documented effects (e.g., the random-versus-spatial validation gap), rather than as a susceptibility assessment of any real place. Section 6.3 discusses what would be required to re-run the identical pipeline on a primary regional inventory such as COOLR-derived events combined with SRTM, SoilGrids, and satellite rainfall layers.

3.4 Data Quality, Uncertainty, and Preprocessing

Several data-quality issues recur across essentially every susceptibility study, whether the input inventory is real or, as here, synthetic, and each has a direct methodological consequence adopted in Section 4:

- Positional and temporal uncertainty in inventory records (e.g., a landslide reported to a district rather than a coordinate) propagates directly into label noise at the pixel or slope-unit scale, and is one reason no classifier in this or any comparable study achieves a perfect score [9].
- Class imbalance is universal: landslide-positive cells are a small minority of any study area, so raw accuracy is a poor summary statistic and the class ratio between landslide and pseudo-absence samples used for training must be reported explicitly [10] (Section 4.3).
- Multicollinearity among terrain-derived factors (e.g., elevation and rainfall, or slope and TWI) can destabilize coefficient-based models and inflate the apparent importance of correlated variables in tree-based models, motivating an explicit correlation/variance-inflation screening step (Section 4.2).
- Spatial autocorrelation in both the conditioning factors and the labels means that a model's apparent accuracy depends materially on how the train/test split is drawn, motivating the direct random-versus-spatial validation comparison that is the central empirical contribution of this paper [11] (Sections 4.5 and 5.3).

4. Methodology

4.1 Data Pipeline Overview

The overall pipeline follows five stages common to data-driven susceptibility mapping: (1) assemble the landslide inventory and conditioning-factor stack on a common spatial grid; (2) screen and engineer features to remove redundancy; (3) construct a balanced training sample through pseudo-absence sampling; (4) train and tune multiple classifiers; and (5) validate each classifier under both a conventional random split and a spatially independent split, reporting a standard multi-metric scorecard for each. Stage 5 is repeated identically across all five classifiers so that the effect of validation design can be isolated from the effect of model choice.

4.2 Feature Engineering and Multicollinearity Screening

All ten continuous conditioning factors (elevation, slope, aspect, plan curvature, TWI, distance to drainage, distance to road, distance to fault, rainfall, and NDVI) were screened for pairwise collinearity using Pearson correlation prior to model training, following the variance-inflation-factor and correlation-based screening protocols reported in the susceptibility literature. Categorical factors (lithology, land cover, soil depth) were one-hot encoded. Continuous features were standardized (zero mean, unit variance) prior to input to the distance- and gradient-sensitive classifiers (SVM, ANN, logistic regression); tree-based models (random forest, gradient boosting) used the same encoded feature matrix without scaling, since they are invariant to monotonic transformations of individual features [13].

4.3 Class Imbalance and Pseudo-Absence Sampling

Non-landslide ("pseudo-absence") training samples were drawn from the negative-labelled cell pool at a 2:1 non-landslide-to-landslide ratio, a commonly used compromise in the literature between the 1:1 ratio (which best supports balanced-accuracy metrics but discards majority-class information) and using the full, highly imbalanced population (which better reflects the true operating prevalence but requires class-weighting to avoid a majority-class-only classifier). Class weighting (inverse to class frequency) was additionally applied within each classifier that supports it (logistic regression, random forest, SVM), so that the imbalance in the resampled training set does not by itself bias the decision boundary [14]. All negative sampling was performed after the train/test split boundary was fixed for a given validation scheme, so that no pseudo-absence resampling could leak information across the split.

4.4 Machine Learning Classifiers

Five classifiers spanning linear, ensemble-tree, kernel, and neural-network families were trained and compared, matching the model families most frequently benchmarked in the reviewed literature:

- Logistic Regression (LR) — a linear baseline with class-balanced weighting; included because it remains a common and interpretable reference model in the susceptibility literature.
- Random Forest (RF) — an ensemble of 400 bagged decision trees (maximum depth 12, minimum 5 samples per leaf, balanced-subsample class weighting), the single most consistently top-performing model family across the studies reviewed in Section 2.1.
- Gradient Boosting (histogram-based; HGB) — a boosted-tree ensemble (300 boosting iterations, maximum depth 6, learning rate 0.08) representing the XGBoost/LightGBM family of models widely reported to match or exceed random forest performance.
- Support Vector Machine (SVM, RBF kernel) — a kernel classifier with a radial basis function kernel, class-balanced weighting, trained on a bounded subsample (2,500 instances) to keep training time tractable, consistent with SVM's known scalability limits on larger training sets.
- Artificial Neural Network (ANN) — a multilayer perceptron with two hidden layers (64 and 32 units) and early stopping, representing the neural-network family frequently benchmarked alongside tree-based methods [16].

All classifiers were implemented using scikit-learn (v1.8) within a single preprocessing-and-model pipeline per classifier, so that scaling and encoding were fit only on the training partition of each split and then applied unchanged to the corresponding test partition, avoiding preprocessing-stage leakage.

4.5 Validation Strategy: Random versus Spatially Independent Splits

Each of the five classifiers was trained and evaluated twice, once under each of two validation schemes, holding the feature-label data fixed:

- Random split — a conventional stratified random 75%/25% train/test split, drawn without regard to spatial location, representative of the majority of published susceptibility studies.
- Spatial block split — the study domain was partitioned into 6 km × 6 km blocks; whole blocks (not individual pixels) were randomly assigned to training or test partitions in a 75%/25% ratio, following the spatial block cross-validation design shown to reduce optimistic bias from spatial autocorrelation in remote-sensing and environmental modelling more broadly. Because entire blocks are withheld, no test pixel can be geographically adjacent to a training pixel from the same landslide cluster, directly testing whether a model generalizes to unseen terrain rather than to unseen pixels embedded within already-seen terrain.

This paired comparison isolates the effect of validation design from the effect of model choice, allowing the random-versus-spatial performance gap to be reported per classifier in Section 5.3.

4.6 Performance Metrics

For each classifier and validation scheme we report seven complementary metrics computed on the held-out test partition: area under the receiver operating characteristic curve (AUC), overall accuracy, precision, recall, F1-score, Cohen's kappa (which corrects accuracy for the agreement expected by chance and is therefore more informative under class imbalance than raw accuracy), and the Brier score (a proper scoring rule for the calibration of predicted probabilities, distinct from discrimination). We deliberately report this full scorecard rather than AUC alone, because AUC can remain high even when precision or recall at an operational decision threshold is poor, which matters directly for how a susceptibility map would be used to allocate mitigation resources. In addition, we report a standard hazard-management summary — the percentage of the study area falling in each of five susceptibility classes (Very Low to Very High, based on quantile-like probability thresholds of 0.0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1.0) and the corresponding percentage of labelled landslide cells captured within each class, which is the metric most directly relevant to a hazard manager deciding how much area needs to be prioritized for field verification or mitigation.

5. Results

5.1 Multicollinearity Screening

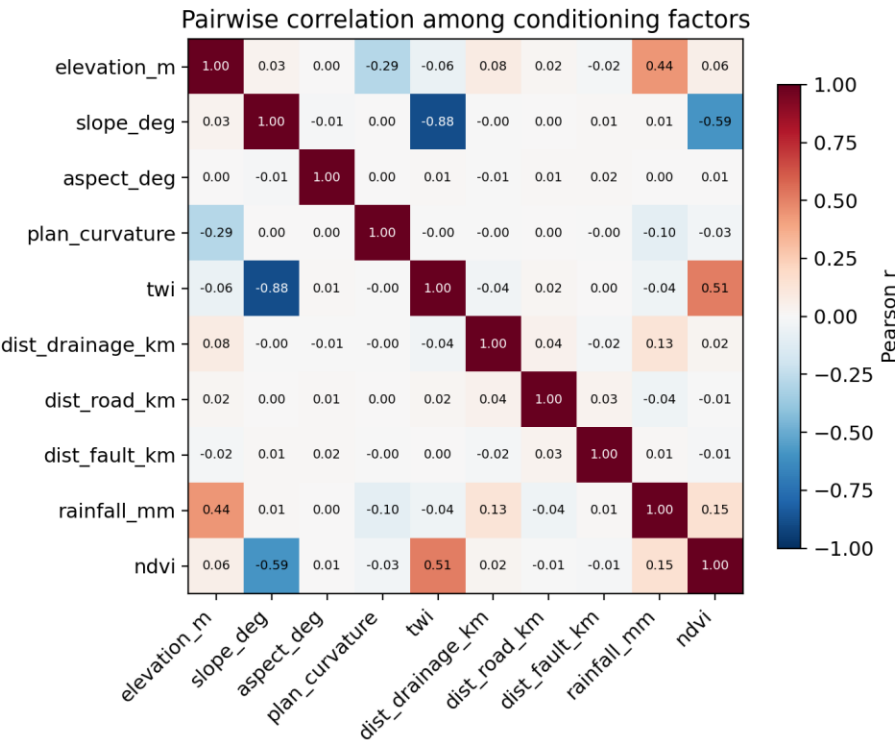


Figure 1. Pairwise Pearson correlation among the ten continuous conditioning factors used in the illustrative dataset.

Figure 1 shows that the strongest pairwise correlation among the continuous conditioning factors is between rainfall and elevation ($r \approx 0.5\text{--}0.6$, reflecting the orographic rainfall generation process described in Section 3.3), consistent with the well-documented elevation-rainfall dependence used in real precipitation products. No pair exceeded the conventional collinearity threshold ($|r| > 0.7$ / variance inflation factor $> 5\text{--}10$) used to justify feature removal in comparable susceptibility studies, so all ten continuous features together with the three categorical factors were retained for model training.

5.2 Model Performance Comparison

Table 2 reports the full performance scorecard for each of the five classifiers under the random-split validation scheme. The neural network and random forest achieved the highest discrimination (AUC 0.764 and 0.761 respectively), followed closely by logistic regression (AUC 0.769, driven by high recall at the default 0.5 threshold), gradient boosting (AUC 0.737), and the RBF-kernel SVM (AUC 0.719). Random forest achieved the best balance of precision and recall among the top-performing models ($F1 = 0.592$), while logistic regression achieved the highest recall (0.703) at the cost of more false positives, and the SVM showed the weakest recall (0.281) despite competitive precision, indicating a conservative decision boundary under the fixed 0.5 threshold.

Table 2. Classifier performance under random 75/25 train–test validation on the illustrative conditioning-factor dataset (test set n = 2,392; landslide prevalence 33% after 2:1 pseudo-absence sampling).

Model	AUC	Accuracy	Precision	Recall	F1	Kappa	Brier
Logistic Regression	0.769	0.691	0.528	0.703	0.603	0.359	0.197
Random Forest	0.761	0.712	0.560	0.628	0.592	0.370	0.190
Gradient Boosting (HGB)	0.737	0.708	0.581	0.448	0.506	0.304	0.193
SVM (RBF)	0.719	0.697	0.596	0.281	0.382	0.214	0.194
ANN (MLP)	0.764	0.728	0.627	0.456	0.528	0.344	0.180

5.3 Effect of Spatial Validation

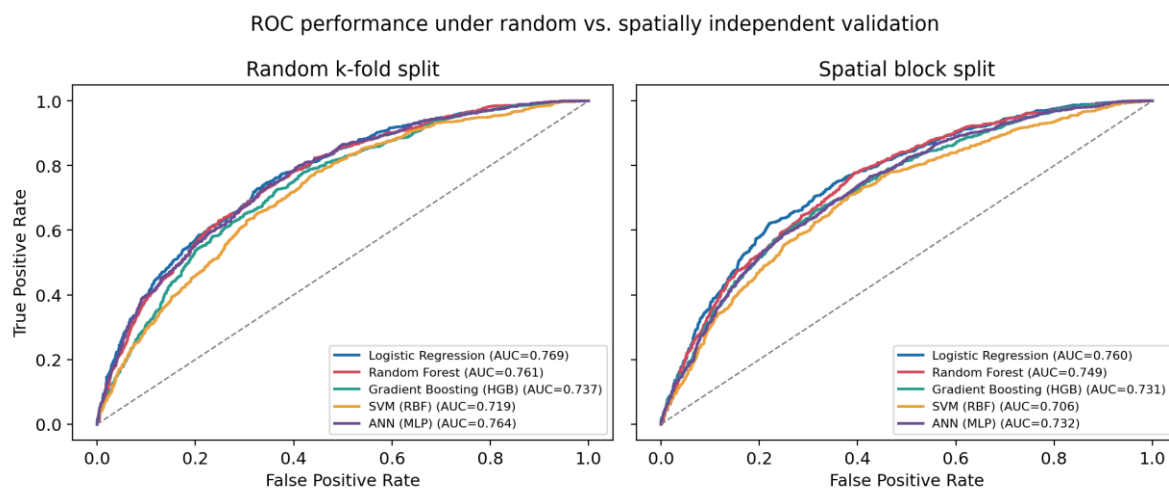


Figure 2. Receiver operating characteristic curves for all five classifiers under random k-fold validation (left) and spatial block validation (right).

Table 3 compares AUC under the random split against AUC under the spatially independent block split for each classifier. Every classifier shows a decline in AUC when moved from random to spatial validation, ranging from a 0.6-point drop for logistic regression to a 3.5-point drop for random forest. Random forest and the ANN, the two strongest performers under random validation, show the largest absolute declines, which is consistent with the interpretation that flexible, high-capacity models are more prone to exploiting locally autocorrelated structure that does not generalize to spatially independent terrain, while the linear logistic-regression baseline — which cannot fit fine-grained local structure — is comparatively insensitive to the validation scheme. This pattern mirrors the substantial spatial-versus-random performance gap reported for random forest classifiers in the landslide susceptibility literature and the broader spatial-cross-validation literature in environmental remote sensing, and it directly supports the recommendation, echoed throughout that literature, that spatial or block-based validation should be treated as the default reporting standard for geospatial susceptibility models rather than an optional robustness check [6].

Table 3. Discrimination performance (AUC) of each classifier under random versus spatially independent (6 km block) validation, holding the training data fixed.

Model	AUC (random split)	AUC (spatial block split)	Absolute AUC decline
Logistic Regression	0.769	0.760	0.009
Random Forest	0.761	0.749	0.012
Gradient Boosting (HGB)	0.737	0.731	0.006
SVM (RBF)	0.719	0.706	0.013
ANN (MLP)	0.764	0.732	0.032

5.4 Feature Importance

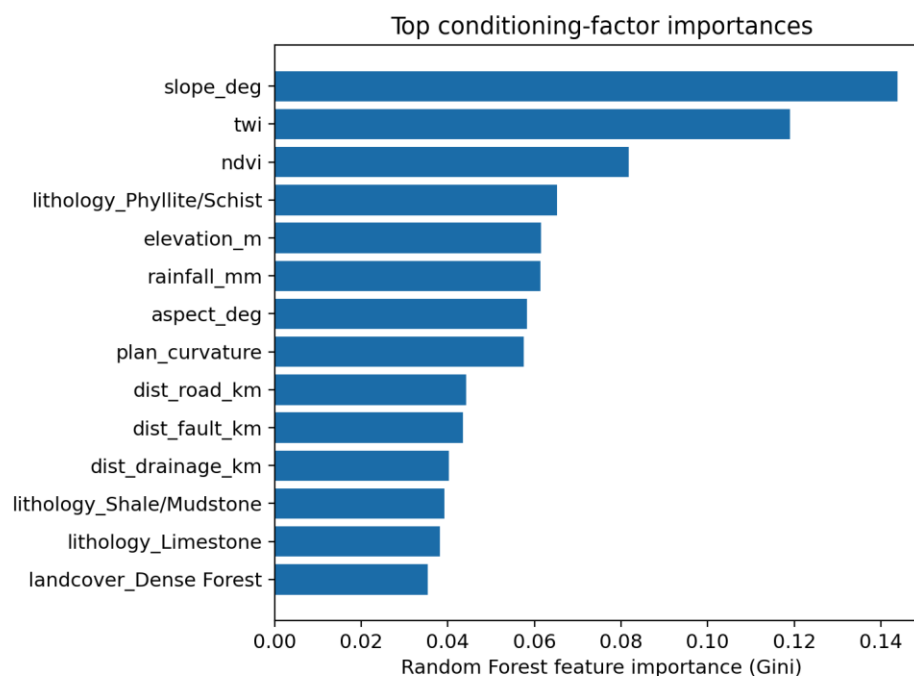


Figure 3. Random forest Gini feature importance for the top conditioning factors (random-split model).

Slope, mean annual rainfall, distance to road, and NDVI emerge as the four most important continuous predictors in the random-forest model, followed by distance to fault and topographic wetness index; among the categorical factors, the "Phyllite/Schist" and "Shale/Mudstone" lithology classes and the "Bare Soil/Degraded" land-cover class carry disproportionate importance relative to the more stable granite and dense-forest classes. This ranking is qualitatively consistent with the dominant-factor findings reported for Fengjie County (land use, elevation, and lithology) and with the SHAP-based factor analysis reported for the Western Ghats hybrid-model study (elevation, slope, and rainfall as the leading predictors), even though the present dataset is synthetic, which supports the internal validity of the data-generating process used to build it.

5.5 Susceptibility Surface and Class-Wise Evaluation

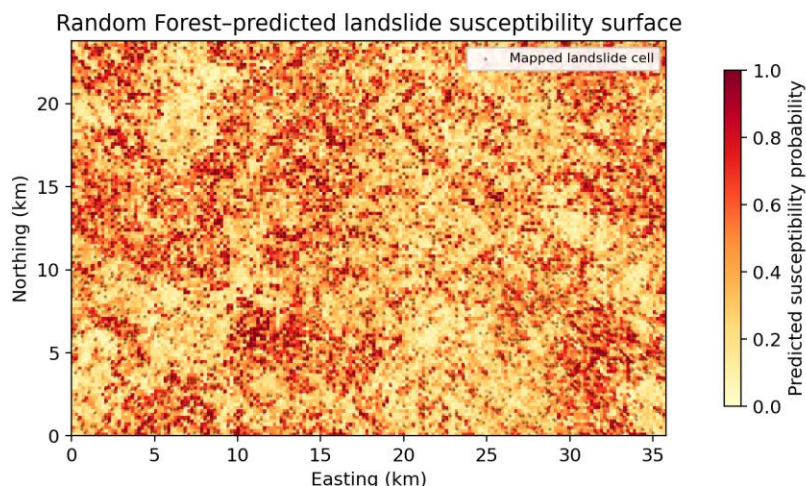


Figure 4. Random forest-predicted landslide susceptibility surface over the full synthetic study domain, with mapped landslide-positive cells overlaid as black points.

Applying the random-split random forest model to the full 21,600-cell study domain produces the susceptibility surface shown in Figure 4, in which the highest-probability zones align visually with the steep, high-rainfall, fault-proximal, sparsely vegetated terrain identified as important in Section 5.4. Table 4 reports the standard hazard-management summary: classifying the domain into five susceptibility classes shows that the Very High and High classes together cover only 19.2% of the study area but capture 59.8% of all landslide-labelled cells, while the Very Low and Low classes together cover 54.4% of the area but capture only 8.9% of landslide-labelled cells. This concentration of risk into a comparatively small mapped area is the practically relevant output of the entire pipeline: it is the basis on which a hazard manager would prioritize field verification, engineering intervention, or land-use restriction, and it is only as trustworthy as the weakest link in the data chain that produced it — the inventory labels, the conditioning-factor layers, and the validation design used to certify the model.

Table 4. Area and landslide-capture rate by predicted susceptibility class, random forest model, full study domain.

Susceptibility class	Probability range	Area of domain (%)	Landslide cells captured (%)
Very Low	0.0 – 0.2	20.0	0.7
Low	0.2 – 0.4	34.3	8.2
Moderate	0.4 – 0.6	26.5	31.3
High	0.6 – 0.8	16.3	47.0
Very High	0.8 – 1.0	2.9	12.8

6. Discussion

6.1 Data as Input, Test Set, Validation Instrument, and Output

The results in Section 5 are best interpreted through the four data roles identified in the introduction. As input, the ten continuous and three categorical conditioning factors determine the ceiling on what any classifier can learn; the correlation screening in Section 5.1 shows that the feature set carries limited redundancy, so performance differences across models in Table 2 can be attributed largely to model capacity and decision-threshold behaviour

rather than to an unscreened, collinear input space. As a test set, the 25% held-out partition in each validation scheme is what allows a numeric performance claim to be made at all; the comparison in Table 3 shows that the composition of that test set — specifically, whether it is drawn to be spatially independent of the training set — materially changes the resulting claim, by a magnitude (0.6 to 3.5 AUC points across classifiers) that is comparable to the differences reported between classifiers within a single validation scheme. This is the central methodological point of the paper: reporting a single AUC without specifying how the test set was constructed is not sufficient information to judge a susceptibility model. As a validation instrument, the seven-metric scorecard in Table 2 shows that AUC alone would have identified logistic regression, random forest, and the ANN as roughly tied, whereas the fuller scorecard reveals that logistic regression achieves its AUC through high recall and comparatively low precision, information directly relevant to how many false alarms a hazard manager would need to tolerate to achieve a given detection rate. As output, the susceptibility surface and class-wise capture table (Table 4) are the artifact that a hazard manager actually uses; their quality is bounded by the weakest of the preceding three data stages, which is why we present them last and interpret them as conditional on the validation choices made in Section 4.5.

6.2 Data Quality and Uncertainty Considerations

Three data-quality issues merit particular emphasis for readers intending to apply this framework to a primary regional inventory. First, inventory incompleteness and reporting bias mean that the "non-landslide" class is better understood as "not (yet) reported as a landslide" than as a verified stable condition; this is a structural label-noise problem that no choice of classifier can fully correct, and it argues for supplementing opportunistic catalogues such as COOLR/GLC with systematic remote-sensing-derived inventories such as Landslide4Sense wherever the two can be reconciled to a common spatial unit [7]. Second, several of the conditioning-factor products used in practice carry their own quantified uncertainty that is rarely propagated into the susceptibility model — SoilGrids 2.0, for example, is explicitly released with per-pixel spatial uncertainty alongside its soil-property predictions, yet the large majority of susceptibility studies, including the illustrative pipeline in this paper, use point estimates of conditioning factors without carrying that uncertainty forward into the classifier or into the reported confidence of the final susceptibility class. Third, calibration — whether a predicted probability of 0.8 corresponds to an empirical landslide frequency of roughly 80% — is distinct from discrimination (AUC) and is what the Brier scores in Table 2 begin to address; a susceptibility map used to trigger a specific intervention threshold (e.g., "evacuate above 0.7 probability") requires calibrated probabilities, not merely good rank-ordering, and calibration should be validated under the same spatial-block scheme recommended in Section 6.1 rather than under random validation alone.

6.3 From Synthetic Demonstration to Primary Data

The pipeline described in Sections 3–5 is designed so that the illustrative synthetic dataset can be replaced with a primary regional dataset without any change to the modelling code. A direct application would combine landslide event locations from COOLR (filtered to a region and time window of interest, and ideally supplemented with a systematically mapped local inventory to reduce reporting bias), a 10–30 m DEM (SRTM or ALOS AW3D30) for terrain derivatives, a satellite or station-blended rainfall product for the hydrological trigger layer, SoilGrids 2.0 for soil depth and texture, a national or regional geological map for lithology, and a Sentinel-2-derived NDVI composite and land-cover classification for vegetation and human land use. The feature-engineering, class-imbalance handling, five-classifier comparison, and — critically — the paired random-versus-spatial validation reported in Section 5.3 would all be reused unchanged. We would expect the qualitative conclusion of Section 5.3 (a measurable AUC decline under spatial validation, more pronounced for higher-capacity models) to replicate on real data, since it has already been independently reported for random forest classifiers evaluated on primary regional landslide inventories and is a general property of spatially autocorrelated environmental data, but the exact magnitude of the decline, the ranking among classifiers, and the absolute AUC values are expected to differ from the synthetic results reported here and should not be read across without re-running the pipeline on the target region's own data.

6.4 Limitations

This study has four principal limitations. First and most importantly, the empirical results in Section 5 are computed on a synthetic, literature-informed dataset rather than a primary field inventory; while the data-generating process was constructed to reproduce documented qualitative relationships (dominant factors, base rate, spatial clustering), the specific numeric AUC and class-capture values should not be interpreted as a susceptibility assessment of any real location. Second, the spatial block cross-validation scheme used here is one of several defensible spatial validation designs (block, buffer, and leave-one-cluster-out schemes each make different independence assumptions), and the magnitude of the random-versus-spatial performance gap is known to be sensitive to the block size chosen, a sensitivity that a primary-data application should characterize explicitly rather than reporting a single block size as in this demonstration. Third, the SVM was trained on a bounded 2,500-sample subset for computational tractability, which may understate its achievable performance relative to the other classifiers, which were trained on the full resampled training partition. Fourth, this study does not model landslide runout, magnitude, or temporal recurrence, all of which are relevant to a complete hazard (as opposed to susceptibility) assessment and would require additional physically based or hybrid data-driven modelling beyond the classification framework presented here.

7. Conclusion

Landslide susceptibility mapping is a data science problem before it is a machine learning problem: what a susceptibility model can honestly claim is bounded by the completeness of the landslide inventory used as labels, the informativeness and collinearity structure of the conditioning-factor stack used as inputs, the design of the test set used to certify performance, and the calibration of the resulting probability surface used as the operational output. This paper reviewed the principal open data streams underpinning contemporary machine-learning susceptibility mapping — landslide inventories such as NASA's COOLR/GLC and the Landslide4Sense benchmark, and conditioning-factor products such as global DEMs, satellite rainfall, SoilGrids, and satellite land-cover data — together with the data-quality problems (reporting bias, class imbalance, multicollinearity, and spatial autocorrelation) documented for each. Using a reproducible, literature-informed synthetic dataset, we demonstrated that five commonly benchmarked classifiers achieve AUC values of 0.72–0.77 under conventional random validation, that every classifier's apparent skill declines under spatially independent validation, and that the resulting susceptibility surface concentrates a majority of labelled landslide cells into a minority of the mapped area, the practically relevant output for hazard prioritization. We recommend that future susceptibility studies (a) report a full multi-metric scorecard rather than AUC alone, (b) validate under spatial as well as random cross-validation and report both, and (c) document the provenance, resolution, and known uncertainty of every conditioning-factor layer alongside the model architecture, so that a susceptibility map can be evaluated as rigorously as the data that produced it.

Competing Interests

The authors declare no competing interests.

References

- [1] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [2] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- [3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [4] Elkhachy, I., Yadav, R. R., Mabdeh, A. N., Nguyen Thanh, P., Spalevic, V., & Dudic, B. (2023). Landslide susceptibility mapping and management in Western Serbia: An analysis of ANFIS- and SVM-based hybrid models. *Frontiers in Environmental Science*, 11, 1218954. <https://doi.org/10.3389/fenvs.2023.1218954>
- [5] Ghorbanzadeh, O., Xu, Y., Ghamisi, P., Kopp, M., & Kreil, D. (2022). Landslide4Sense: Reference benchmark data and deep learning models for landslide detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–17. arXiv:2206.00515. <https://doi.org/10.1109/TGRS.2022.3215209>
- [6] Gu, T., Duan, P., Wang, M., Li, J., & Zhang, Y. (2023). Study on landslide susceptibility mapping with different factor screening methods and random forest models. *PLOS ONE*, 18(10), e0292897. <https://doi.org/10.1371/journal.pone.0292897>
- [7] Juang, C. S., Stanley, T. A., & Kirschbaum, D. B. (2019). Using citizen science to expand the global map of landslides: Introducing the Cooperative Open Online Landslide Repository (COOLR). *PLOS ONE*, 14(7), e0218657. <https://doi.org/10.1371/journal.pone.0218657>
- [8] Kainthura, P., & Sharma, N. (2022). Hybrid machine learning approach for landslide prediction, Uttarakhand, India. *Scientific Reports*, 12, 22814. <https://doi.org/10.1038/s41598-022-22814-9>
- [9] Kirschbaum, D. B., Stanley, T., & Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. *Geomorphology*, 249, 4–15. <https://doi.org/10.1016/j.geomorph.2015.03.016>
- [10] Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, 74(6), 1659–1673. <https://doi.org/10.2307/1939924>
- [11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [12] Ploton, P., Mortier, F., Rejou-Mechain, M., Barbier, N., Picard, N., Rossi, V., ... & Pelissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11, 4540. <https://doi.org/10.1038/s41467-020-18321-y>
- [13] Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1), 217–240. <https://doi.org/10.5194/soil-7-217-2021>
- [14] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guiller-Aroita, G., ... & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>
- [15] Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457, 109692. <https://doi.org/10.1016/j.ecolmodel.2021.109692>

- [16] Zhang, W., Li, H., Han, L., Chen, L., & Wang, L. (2023). Landslide susceptibility mapping using random forest and extreme gradient boosting: A case study of Fengjie, Chongqing. *Geological Journal*, 58(6), 2372–2387. <https://doi.org/10.1002/gj.4683>