

Landslide Detection and Susceptibility Mapping Using Machine Learning and Deep Learning: A Comprehensive Review

***Corresponding Author:**
D.B.Mirajkar, Assistant
Professor, Computer
Science & Engineering
Department, Tatyasaheb
Kore Institute of
Engineering and
Technology,
Warananagar,
Kolhapur, India.

D.B.Mirajkar*, **Yasmeen Shaikh²**

¹ Assistant Professor, Computer Science & Engineering
Department, Tatyasaheb Kore Institute of Engineering and
Technology, Warananagar, Kolhapur, India.

² Professor, Computer Science & Engineering Department,
S G. Balekundri Institute of Technology, Belgavi, India.

ABSTRACT

Every year, landslides kill thousands of people and wipe out infrastructure worth billions of dollars, yet predicting where and when they will occur remains stubbornly difficult. Over the last decade or so, the combination of freely available satellite imagery and fast-maturing machine learning (ML) and deep learning (DL) techniques has begun to change that picture in meaningful ways. This paper reviews the trajectory of that change examining how researchers have moved from simple statistical classifiers to sophisticated transformer-based segmentation models, and what that progression has actually delivered in terms of practical detection capability.

We surveyed around 45 peer-reviewed studies published between 2015 and 2024, covering landslide susceptibility mapping, pixel-level inventory mapping, change detection from multi-temporal imagery, and real-time early warning integration. Our analysis draws on work using optical sensors (Sentinel-2, Landsat), SAR platforms (Sentinel-1, ALOS-2), and a growing variety of public benchmark datasets such as Bijie, COOLR, and HR-GLDD. We find that while ensemble classifiers like Random Forest still hold their own for susceptibility mapping, encoder-decoder architectures particularly U-Net variants have become the workhorse for segmentation tasks, with more recent transformer hybrids pushing IoU scores above 0.85 on standard benchmarks.

That said, the field carries some persistent and underappreciated weaknesses: almost all top-performing models have been trained and tested within narrow geographic windows; labeled data remains scarce outside China, Italy, and Central Europe; and the jump from research prototype to operational warning system has proven far harder than benchmark numbers suggest. We close the review by pointing to foundation models, physics-

informed learning, and federated training as directions that may genuinely move the needle on these limitations.

Keywords:-*Landslide detection; susceptibility mapping; deep learning; remote sensing; U-Net; change detection; early warning systems; SAR; Google Earth Engine; geohazard monitoring.*

1. INTRODUCTION

Spend enough time reading post-disaster reports and a pattern quickly emerges: landslides rarely surprise geologists, but they almost always catch civil protection authorities off guard. The disconnect is not scientific ignorance we understand the physics of slope failure reasonably well but a practical inability to monitor at the scale and speed that operational warning demands. Field surveys are slow. Manual photo-interpretation is laborious. And the mountainous terrain where landslides concentrate is precisely the terrain that roads and surveyors find hardest to reach [1, 2].

The promise of satellite remote sensing has been understood since at least the 1990s, but the tools needed to act on that promise only converged within the past decade. The launch of Sentinel-1 in 2014 and Sentinel-2 in 2015 gave researchers free, systematic, near-global coverage at spatial resolutions 10 to 20 metres that are genuinely useful for mapping moderate to large landslides [3]. Around the same time, GPU-accelerated deep learning was becoming accessible outside a handful of well-resourced labs. The convergence was predictable in retrospect: papers applying CNNs and U-Net to satellite landslide mapping started appearing around 2018–2019 and have grown steeply since [4, 7]. What does that literature actually tell us? In broad terms, quite a lot that is encouraging and some things that should give pause. Encouraging: modern segmentation models can delineate freshly triggered landslides with F1 scores above 0.85 on held-out test tiles, which is genuinely better than earlier pixel-based classifiers. Cautionary: most of those test

tiles come from the same geographic region as the training data, which makes the numbers optimistic. Generalisation across climate zones, lithologies, and image acquisition conditions remains an open problem, and it is arguably the problem that matters most for any system meant to work globally [10, 23].

This paper attempts a critical synthesis of where the field stands. Rather than cataloguing every published method, we focus on the threads of development that have most shaped current practice: susceptibility mapping with ensemble classifiers, pixel-level detection with encoder-decoder networks, bi-temporal change detection with Siamese and transformer architectures, and the still-nascent integration of ML into operational early warning pipelines. Along the way we highlight methodological inconsistencies that make cross-study comparison harder than it should be, and we flag the dataset gaps that limit what any model trained today can honestly claim to know.

The remainder of the paper is organised as follows. Section 2 provides background on landslide processes, remote sensing data, and the ML/DL building blocks that matter most for this domain. Section 3 describes how we selected and categorised the reviewed literature. Sections 4 through 6 present the core review across methods, datasets, and evaluation practice respectively. Sections 7 and 8 discuss open challenges and future directions, and Section 9 concludes.

2. BACKGROUND

What Makes Landslides Hard to Detect Automatically

Landslides are not a single phenomenon. They range from shallow debris flows that mobilise within minutes during intense rainfall to deep-seated rotational slides that creep imperceptibly for decades before catastrophic failure [11, 12]. That variability matters enormously for automated detection: the spectral and textural signature of a fresh debris flow scar bare soil, disturbed vegetation, sharp boundaries look very different from the subtle surface bulging that precedes a slow rockslide. Most published ML studies implicitly focus on the former because it is visible in optical imagery and easier to label. Slow-moving failures, which arguably represent a larger share of total landslide risk in many regions, are largely absent from benchmark datasets.

Triggering factors add another layer of complexity. Rainfall is responsible for the majority of global landslide events, but the relevant variables antecedent soil moisture, storm intensity, duration, spatial pattern interacts in ways that remain difficult to capture with coarse gridded precipitation products [13]. Seismically triggered failures have their own spatial signature. Human activity road cutting, reservoir filling, deforestation has become an increasingly important trigger in densely populated hillslope regions of South and Southeast Asia, including parts of India, Nepal, and Indonesia [2]. Any model that treats landslide occurrence as a function of terrain attributes alone will systematically miss events driven by these dynamic factors.

Remote Sensing Platforms and What They Offer

Sentinel-2 is probably the most widely used satellite source in recent landslide studies, for good reason: 10 m resolution in visible and near-infrared bands, 5-day revisit at mid-latitudes, and entirely free through the Copernicus Open Access Hub

or Google Earth Engine [3, 37]. Its limitations are also worth stating plainly: it cannot see through clouds, which are ubiquitous over tropical mountains during and after rainfall events precisely when landslides occur. Sentinel-1 C-band SAR addresses this directly, providing backscatter and coherence measurements regardless of cloud cover or illumination [15]. The trade-off is that SAR interpretation is harder coherence loss from vegetation and surface roughness creates change signals that can mimic or mask landslide signatures.

ALOS-2 PALSAR-2 L-band SAR penetrates vegetation canopy more effectively than C-band and has been especially useful for detecting landslides in heavily forested terrain in Japan, Southeast Asia, and parts of South America [15]. Commercial very-high-resolution platforms (WorldView, Pleiades, SkySat) at 30–50 cm resolution enable detection of small features invisible to Sentinel sensors, though cost and limited coverage restrict their use to targeted post-event studies. UAV surveys now fill a complementary niche: centimetre-level resolution over localised areas, deployable within hours of an event, with emerging automation of flight planning and image processing [16]. LiDAR deserves particular mention. Airborne LiDAR-derived digital elevation models at 1 m or better resolution can reveal pre-failure topographic signatures tension cracks, displaced scarps, asymmetric drainage that are entirely invisible in optical imagery. Where LiDAR coverage exists, it has consistently improved susceptibility model performance. The problem is that systematic LiDAR coverage remains restricted to a handful of high-income countries, limiting its contribution to global-scale studies.

Geospatial Feature Engineering

For susceptibility mapping in particular, the quality and selection of conditioning

factors can matter as much as the choice of ML algorithm. Standard practice extracts slope, aspect, plan and profile curvature, topographic wetness index (TWI), and stream power index from DEMs [17]. Lithology and proximity to geological faults capture rock mass susceptibility. Land cover, NDVI, and distance to roads and rivers round out the typical feature set [18]. What is less often acknowledged is that the relative importance of these factors shifts substantially across geographic and climatic settings: lithology dominates in some alpine environments; rainfall-derived soil moisture indices matter more in humid tropical settings; anthropogenic factors are increasingly important near expanding road networks. Feature importance analyses from RF models provide useful regional insights but are rarely transferred across study areas systematically.

ML and DL Foundations

Classical ML approaches logistic regression, support vector machines (SVM), Random Forest, and gradient boosting variants treat landslide prediction as a tabular classification problem: each spatial unit (pixel or mapping polygon) is represented by a vector of conditioning factor values, and the model learns a decision boundary in that feature space [5, 6, 19]. These methods are well understood, computationally cheap, and perform remarkably well on susceptibility mapping tasks given adequate training data. Their weakness is an inability to exploit spatial context: the texture, shape, and neighbour relationships visible in imagery are invisible to a model that sees only a feature vector.

Deep learning addresses this through hierarchical spatial feature learning. CNNs process image patches and learn filters that capture progressively more abstract spatial patterns through successive convolutional layers [20]. For the specific task of per-pixel labelling the output we want for

landslide mapping encoder-decoder architectures with skip connections, most famously U-Net [21], have become the standard approach. Skip connections allow fine spatial detail from early encoder layers to be combined with semantic information from deeper layers, producing precise segmentation boundaries even for small targets. Transformer-based architectures more recently introduced self-attention mechanisms that can capture long-range spatial dependencies relevant for capturing the contextual cues (surrounding topography, drainage patterns) that distinguish true landslide scars from superficially similar disturbances [9, 22].

3. REVIEW METHODOLOGY

Our search was conducted across Scopus, Web of Science, Google Scholar, and IEEE Xplore, covering the period January 2015 to December 2024. We used keyword combinations including 'landslide detection', 'landslide susceptibility', 'deep learning remote sensing', 'CNN landslide', 'SAR change detection landslide', 'landslide early warning machine learning', and closely related variants. An initial pool of over 800 documents was reduced to approximately 45 core papers through a screening process that prioritised: (a) methodological contribution rather than purely applied case studies; (b) use of satellite or airborne remote sensing data; (c) peer-reviewed journal publication; and (d) sufficient methodological detail for replication assessment. We additionally required that selected studies report quantitative performance metrics to enable cross-study comparison, even where the specific metrics varied.

We acknowledge a selection bias inherent in any review of this kind: highly cited papers from well-resourced research groups in China, Europe, and North America are overrepresented relative to work from South Asia, Africa, and Latin America, where landslide impacts are

severe but publication rates are lower. We have attempted to partially compensate by deliberately including recent work from Indian and Southeast Asian study areas where available, given that these regions represent some of the highest landslide exposure globally.

4. ML AND DL METHODS: WHAT HAS BEEN TRIED AND WHAT HAS WORKED

Susceptibility Mapping with Classical and Ensemble Classifiers

Landslide susceptibility mapping (LSM) estimating where future failures are most likely was one of the first landslide tasks to attract serious ML attention, and it remains one of the best-studied. The appeal is straightforward: it is a supervised classification problem with a reasonably well-defined feature space and an outcome variable (landslide presence/absence) that can be assembled from existing inventories. Random Forest emerged early as the dominant method, owing partly to its robustness to correlated features and irrelevant inputs, and partly to its straightforward implementation in standard Python and R libraries [5].

Merghadi et al.'s (2020) systematic comparison across 11 algorithms and multiple study areas remains the most rigorous benchmarking exercise in this literature [23]. Their headline finding that tree ensemble methods (RF, XGBoost, gradient boosting) consistently outperform single classifiers, with AUC-ROC values above 0.90 has been widely replicated. What is less often highlighted from the same study is the large performance variance across geographic settings: an algorithm that achieves AUC 0.95 in one study area may drop to 0.82 in another, even with identical hyperparameter settings. This instability suggests that reported benchmark numbers should be interpreted cautiously when considering operational deployment in new regions.

Neural network approaches to LSM have followed two broad strategies. The first adapts standard tabular classifiers by replacing RF or XGBoost with multilayer perceptrons or shallow CNNs operating on small image patches centred on each mapping unit [7]. The second builds full spatial models that simultaneously map susceptibility across entire images or tiles, implicitly incorporating neighbourhood context. Both strategies have shown competitive performance, though direct comparison with ensemble classifiers remains mixed in data-rich settings with high-quality imagery, DL models tend to win; in data-sparse settings with fewer than a few thousand labeled samples, RF usually holds its own.

Inventory Mapping and Semantic Segmentation

Generating accurate landslide inventory maps the spatial catalogues of where past failures have occurred is both a scientific output in itself and a prerequisite for training data-hungry DL models. Traditional manual mapping from aerial photographs is painstaking work; automating it with image segmentation models has therefore attracted considerable research energy.

U-Net, adapted from biomedical image segmentation [21], has become the dominant architecture for this task. Ji et al.'s (2020) application to Sentinel-2 imagery over Bijie City, Guizhou [24], was among the first systematic demonstrations of U-Net's effectiveness for landslide segmentation and established the Bijie dataset as a community benchmark. Their model achieved F1 scores of 0.85 on held-out test tiles competitive with manual interpretation at comparable resolution while processing an area that would have taken human analysts weeks in a matter of minutes. Subsequent work has extended the architecture in several directions: attention-gated U-Net variants that selectively weight

informative feature channels; multi-scale feature pyramid extensions for handling the wide size range of natural landslides; and dual-branch models that separately process spectral and elevation inputs before fusion [24, 25].

One limitation that runs through much of this segmentation literature is class imbalance. Landslide pixels typically constitute between 1% and 5% of any training image, and standard cross-entropy loss functions trained on such data tend to predict background everywhere. The community has converged on focal loss and weighted Dice loss as partial remedies, though the fundamental data constraint remains. Some groups have explored self-supervised pretraining on unlabeled satellite imagery as a way to build richer representations before fine-tuning on scarce landslide labels a direction we return to in Section 8.

Change Detection Between Pre- and Post-Event Imagery

When an event has already occurred, the goal shifts from susceptibility estimation to rapid mapping of where failures happened. Bi-temporal change detection comparing co-registered imagery from before and after an event is the natural approach, and it has been tackled with increasingly sophisticated DL methods over the past five years.

Siamese networks, which share weights between two parallel encoding branches processing pre- and post-event images, provide an elegant architecture for this task: they learn what kinds of spectral and textural changes are diagnostically meaningful for landslides rather than relying on hand-crafted difference indices [26]. Chen and Shi's (2021) Binary change detection with Image Transformer (BIT) model [27] represented a significant step forward by applying transformer self-attention to the encoded feature tokens, allowing the model to relate spatially distant but contextually linked regions of

the scene. On standard change detection benchmarks, BIT improved F1 by 3–5 percentage points over the previous CNN baselines.

SAR-based change detection deserves separate treatment. Optical methods fail entirely when post-event imagery is cloud-obscured a common situation given that landslide-triggering rainfall events are associated with persistent cloud cover. Mondini et al.'s (2021) work using Sentinel-1 coherence loss and backscatter change demonstrated that SAR-based approaches can achieve recall rates above 80% for landslides larger than 0.5 ha in Italy, with image acquisition within 24–48 hours of event occurrence [28]. The persistent limitation is that SAR change detection produces a higher false alarm rate than optical methods in vegetated terrain, where wind and seasonal phenology create coherence loss unrelated to landsliding.

Object Detection for Rapid Post-Event Screening

Pixel-level segmentation, while accurate, can be computationally expensive when applied to large post-event image archives covering hundreds of square kilometres. Object detection frameworks YOLO variants and Faster R-CNN in particular offer a faster alternative for initial screening: identifying bounding-box locations of likely landslides that can then be forwarded for detailed analysis [29]. Yi et al. (2020) applied Mask R-CNN to UAV imagery for instance-level landslide detection and segmentation [30], demonstrating that precise individual landslide boundaries could be automatically delineated at scales relevant for volume estimation and damage assessment. This kind of instance-level analysis is increasingly important for quantitative hazard characterisation beyond simple presence/absence mapping.

Transformer Architectures and Foundation Models

The current frontier in the landslide detection literature involves large transformer-based models, some pre-trained at massive scale on diverse imagery. SegFormer [31] applies a hierarchical transformer encoder with a lightweight decoder and has shown competitive segmentation performance with substantially fewer parameters than earlier architectures. More provocatively, Kirillov et al.'s (2023) Segment Anything Model (SAM) [32] trained on over one billion image masks has been applied to satellite landslide imagery in zero-shot and prompt-based modes, with early results suggesting that it can delineate plausible landslide boundaries without any domain-specific training on individual scenes. Whether this translates to reliable inventory mapping at scale is still being assessed, but the prospect of a model that requires no labeled landslide data for deployment is significant.

Geospatial foundation models pre-trained specifically on Earth observation imagery including SkyScript and variants of CLIP adapted to multispectral data provide an alternative route to the same goal: representations rich enough to support downstream landslide tasks with minimal fine-tuning data [33]. This line of work is moving quickly and its implications for data-scarce regional applications could be substantial.

Integration into Early Warning Systems

The ultimate purpose of landslide detection research is to help save lives.

Early warning systems represent the most direct pathway from model output to operational impact, and they remain the most underdeveloped part of the ML landslide literature relative to their importance. Most operational warning systems today rely on rainfall intensity-duration thresholds that define empirical triggering conditions at the regional scale [13]. ML-derived susceptibility maps have been incorporated into some systems as spatial layers that modulate warning levels areas with high susceptibility generate alerts at lower rainfall thresholds but the integration is typically post-hoc rather than architecturally central.

More integrated approaches are beginning to appear. Fang et al. (2022) combined LSTM-based rainfall time-series modelling with static geospatial susceptibility features in a hybrid architecture that achieved 6–12 hour warning lead times with false alarm rates below 20% for rainfall-triggered events in Sichuan Province [35]. IoT-based sensor networks measuring real-time pore water pressure and slope displacement are increasingly being combined with ML models to issue site-specific warnings that go beyond coarse regional thresholds [34]. These developments are promising but still largely restricted to well-instrumented pilot sites; scaling them to data-sparse regions with sparse sensor infrastructure requires approaches that can function with incomplete and intermittent data inputs.

Table 1:-Summary of representative ML/DL studies reviewed, illustrating the spread of methods, data sources, and reported performance across tasks and geographic regions.

Author(s)	Year	Method	Data	Task	Performance	Region
Ghorbanzadeh et al.	2019	CNN vs. RF/SVM	Sentinel-2 + DEM	Susceptibility	AUC: 0.92	Austria
Merghadi et al.	2020	RF, XGBoost,	GIS layers	Susceptibility	AUC: 0.91–0.95	Algeria

		SVM (11 models)				
Ji et al.	2020	U-Net (attention)	Sentinel-2	Segmentation	F1: 0.85	Guizhou, China
Yi et al.	2020	Mask R-CNN	UAV imagery	Instance detection	AP: 0.87	Loess Plateau
Chen & Shi	2021	BIT (Transformer)	Sentinel-2	Change detection	F1: 0.89	Multi-region
Mondini et al.	2021	SAR coherence + DL	Sentinel-1 SAR	Post-event mapping	Recall: 0.82	Italy
Prakash et al.	2021	RF + OBIA	Landsat-8	Inventory mapping	OA: 88%	India
Ding et al.	2022	CNN-Transformer hybrid	VHR optical	Segmentation	F1: 0.91	China
Fang et al.	2022	LSTM + geospatial fusion	IoT + CHIRPS	Early warning	FAR: <20%	Sichuan, China
Huang et al.	2023	SegFormer	Sentinel-2	Segmentation	IoU: 0.83	SW China

Note: OA = Overall Accuracy; FAR = False Alarm Rate; AP = Average Precision; AUC = Area Under ROC Curve.

5. DATASETS AND DATA INFRASTRUCTURE

Benchmark Datasets

The Bijie Landslide Dataset [24] has become the ImageNet of this sub-field not because it is perfect, but because enough groups have used it that results are at least nominally comparable. Comprising 770 aerial image patches from Guizhou Province annotated with pixel-level landslide masks, it provides a manageable entry point for segmentation model development. Its limitations are real, however: samples come from a single climatic and lithological setting, annotation quality is uneven at boundaries, and the dataset is small enough that deep models can memorise it with sufficient augmentation. Work on more globally

representative benchmarks is underway; the HR-GLDD dataset attempts to address geographic bias by drawing samples from multiple continents, though coverage of sub-Saharan Africa and large parts of South America remains thin.

The NASA Cooperative Open Online Landslide Repository (COOLR) [36] takes a different approach: rather than providing imagery, it offers a georeferenced catalogue of over 11,000 globally distributed landslide events with metadata on trigger type, material class, and spatial extent. COOLR is invaluable for susceptibility model training and for temporal analysis of event frequency, but it has well-documented spatial bias toward regions with strong reporting infrastructure meaning that its event density is partly a

map of journalistic coverage rather than purely physical hazard.

Satellite Data Access

The infrastructure for accessing satellite data has improved dramatically over the past decade. Google Earth Engine [37] deserves particular credit for democratising access to the Sentinel and Landsat archives: researchers who would previously have needed terabytes of local storage and substantial preprocessing pipelines can now write a few dozen lines of JavaScript and access imagery for any location and time period within minutes. This has been transformative for groups in institutions with limited computing infrastructure, and its effect on the geographic diversity of published research more case studies from South Asia, Africa, and Latin America is detectable in the literature from roughly 2018 onward.

CHIRPS rainfall data [38] has become the de facto standard for triggering factor analysis in studies without access to dense rain gauge networks, though its 5 km resolution misses the sub-kilometre spatial variability in convective rainfall that matters for shallow landslide initiation. GPM IMERG offers higher temporal resolution (30 minutes) at comparable spatial resolution and is increasingly used for near-real-time event monitoring. Neither product captures local orographic enhancement well, which is a persistent source of error in rainfall-threshold-based warning approaches.

Persistent Data Gaps

Several important gaps are worth stating explicitly. First, annotated landslide data for slow-moving failures which require InSAR time-series analysis rather than optical change detection is almost entirely absent from publicly available datasets. Second, the overwhelming majority of benchmark samples come from China, Italy, and Central Europe; South and Southeast Asia, despite carrying very high

landslide exposure, contributes a small fraction of publicly available labeled data. Third, temporal depth is lacking: most datasets capture a single post-event snapshot, making it impossible to train or evaluate models for monitoring progressive failure over time. Addressing these gaps would require coordinated international data collection efforts that are politically and logistically challenging but scientifically essential.

6. EVALUATION PRACTICE: WHAT WE MEASURE AND WHAT WE MISS

The landslide ML literature has not converged on a single evaluation framework, and this creates real difficulties for synthesis. Susceptibility mapping studies almost universally report AUC-ROC, which is convenient because it does not require selecting a classification threshold but AUC is insensitive to the calibration of predicted probabilities and can be inflated by spatial autocorrelation between training and test samples [10, 39]. Segmentation studies report IoU, F1, precision, and recall in varying combinations. Change detection studies sometimes use different formulations of these metrics. The result is that a claim like 'our method outperforms prior work' almost always rests on a narrow comparison within a single study's evaluation protocol rather than on genuine benchmark superiority.

The deeper problem is spatial cross-validation. Standard random train-test splits that draw samples from across an image will almost always yield optimistic performance estimates because spatially proximate pixels share spectral and topographic characteristics a model that memorises local texture will appear to generalise when it has not. Spatially blocked cross-validation, which holds out entire geographic sub-regions for testing, produces more honest estimates and more useful information about geographic

transferability, but it has not yet become standard practice [23].

Table 2:-*Evaluation metrics used in landslide ML/DL studies, with definitions and recommended application context.*

Metric	Formulation	Primary Application	Key Limitation
Intersection over Union (IoU)	$TP / (TP + FP + FN)$	Segmentation benchmarking	Sensitive to boundary precision
F1-Score (Dice)	$2TP / (2TP + FP + FN)$	Segmentation, change detection	Ignores true negatives
Precision	$TP / (TP + FP)$	Early warning (false alarm cost)	Ignores missed detections
Recall (Sensitivity)	$TP / (TP + FN)$	Life-safety applications	Ignores false alarms
AUC-ROC	Area under ROC curve	Susceptibility mapping	Inflated by spatial autocorrelation
Overall Accuracy	$(TP+TN) / N$	General classification	Misleading under class imbalance

7. CHALLENGES

The field has made genuine progress, but it is worth being candid about the problems that routine benchmark improvements have not resolved.

Geographic generalisation is, in our assessment, the most pressing unsolved problem. A model trained on landslides in the Chinese loess plateau will encounter a fundamentally different spectral and topographic signature when deployed in the wet tropical mountains of Kerala or the glacially sculpted terrain of the Swiss Alps. Feature distributions shift, class boundaries shift, even the morphological definition of what constitutes a discrete landslide shifts. Domain adaptation methods including instance re-weighting, adversarial feature alignment, and style transfer for satellite imagery have been explored as mitigations [41], but demonstrated cross-regional generalisation remains fragile in most published work. Operationally, this means that a new regional deployment effectively requires

new labeled data, which is expensive and slow to acquire.

Data scarcity and class imbalance interact in ways that compound the generalisation problem. In a typical post-event Sentinel-2 scene, landslide pixels are outnumbered by background pixels by factors of 20 to 100 or more. Standard loss functions respond by learning to predict background everywhere, which yields high overall accuracy but useless recall. Focal loss, weighted sampling, and synthetic data augmentation address symptoms rather than the underlying scarcity. Semi-supervised and self-supervised methods that leverage the much larger pool of unlabeled satellite imagery are promising but have only recently begun to appear in the landslide literature [33].

Explainability remains a practical barrier to operational adoption. Emergency managers and civil protection authorities are rightly sceptical of black-box predictions with no interpretable rationale. Gradient-based attribution methods like Grad-CAM and feature attribution through

SHAP values provide post-hoc explanations that are better than nothing, but they are not integrated into most published models, and their fidelity to the model's actual decision process is contested [43]. Physics-informed models that incorporate geotechnical relationships explicitly would provide more trustworthy explanations but require a different modelling philosophy than the pure data-driven approaches that dominate the literature.

Real-time and edge deployment constraints are underappreciated in a literature that typically reports results from offline analysis of archived imagery. An operational early warning system must ingest imagery, run inference, and issue alerts within a time window that is meaningful for evacuation potentially 6 to 12 hours in the case of rainfall-triggered failures [35]. State-of-the-art transformer models that achieve the best benchmark numbers often have inference latencies and memory footprints that are incompatible with this requirement at scale, particularly in regions with limited cloud computing infrastructure. The trade-off between model performance and operational feasibility has not been sufficiently examined in the published literature.

Finally, the temporal dimension of slow-moving landslides is almost entirely unaddressed by mainstream ML approaches. InSAR-derived surface displacement time series from Sentinel-1 can detect millimetre-scale precursory deformation months before catastrophic failure, but integrating these signals with ML-based prediction models requires time-series architectures, multi-year data records, and failure event labels that are rarely assembled in publicly available form. This gap represents a major missed opportunity, given that slow-moving failures are precisely the type where early detection could most reliably enable preventive action.

8. WHERE THE FIELD SHOULD GO

Several directions seem to us most likely to produce genuine advances in operational capability rather than incremental benchmark improvements.

Foundation models and self-supervised learning represent probably the most significant near-term opportunity. The success of SAM [32] and geospatial foundation models [33] in zero-shot and few-shot settings suggests that it may become possible to deploy useful landslide detection without regional labeled data a capability that would transform the field's applicability in data-sparse high-risk regions. Realising this potential will require foundation models trained specifically on multi-spectral and SAR data at the spatial resolutions relevant for landslide detection, which is an active area of investment in the broader Earth observation ML community.

Multimodal fusion combining optical, SAR, elevation, and time-series rainfall inputs within a single learnable framework remains underexplored relative to its obvious physical motivation. Landslides are driven by the interaction of static terrain factors and dynamic triggering inputs; a model that sees both simultaneously should in principle outperform one that sees either alone. Cross-modal attention mechanisms and contrastive multiview learning offer architectural tools for this, but producing the multi-source aligned training datasets required to train such models at scale is a non-trivial data engineering challenge [44].

Physics-informed neural networks (PINNs) that embed slope stability equations or hydrological process models as constraints on the learning objective offer a principled path toward models that generalise better across geologic settings. If a model is required to produce predictions consistent with the Mohr-Coulomb failure criterion, it cannot easily learn spurious correlations from the

training distribution. Early explorations of this idea in landslide susceptibility mapping show promise [10], and extension to dynamic rainfall-triggered scenarios is a natural next step.

Standardised global benchmarks are a prerequisite for progress that the community can measure honestly. A globally representative dataset with consistent annotation protocols, standardised spatial cross-validation splits, and multi-temporal coverage would allow direct comparison of methods in a way that currently is not possible. Assembling such a dataset requires coordinated international effort similar to what produced the COOLR catalog and sustained data curation beyond individual research projects. Funding agencies with a mandate in natural hazard reduction would be natural hosts.

Uncertainty quantification is another under-served need, particularly for operational applications. A susceptibility map that communicates only a point estimate provides less decision-relevant information than one that also expresses confidence bounds. Bayesian neural networks, Monte Carlo dropout, and conformal prediction methods all provide routes to calibrated uncertainty estimates from DL models, but their application to landslide ML remains limited [10]. Given the life-safety context of early warning applications, this seems like an important gap to close.

9. CONCLUSION

The past decade has seen genuine and substantial progress in ML-based landslide detection. From the early demonstrations that Random Forest could outperform logistic regression on susceptibility mapping, through the establishment of U-Net as a practical tool for automated inventory generation, to current explorations of transformer architectures and foundation models, the field has moved quickly and productively. The

combination of freely available Sentinel imagery, accessible cloud processing through platforms like Google Earth Engine, and open-source DL frameworks has lowered barriers to entry in ways that have broadened geographic participation in the research.

What the benchmark numbers do not always reveal is how much unfinished work remains between a well-performing model on a held-out test set and a system that reliably serves emergency managers in regions that look nothing like the training data. The generalisation problem is real and underappreciated. Slow-moving failures are nearly invisible in the literature relative to their societal importance. Operational integration of ML-based detection into warning systems remains a niche activity relative to the volume of pure modelling research. And the lack of standardised evaluation frameworks makes it genuinely difficult to know which algorithmic advances represent durable improvements versus local optimisation on particular benchmark conditions.

None of this diminishes the value of what has been achieved, or the promise of the directions now opening up particularly around foundation models and physics-informed learning. But the field would benefit from a clearer collective focus on the problems that matter most operationally, and from greater investment in the data infrastructure and evaluation standards that would make progress measurable in terms that go beyond the next decimal place on AUC-ROC.

REFERENCES

1. Froude, M.J. and Petley, D.N. (2018). Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, 18(8), pp.2161–2181.
2. Gariano, S.L. and Guzzetti, F. (2016). Landslides in a changing climate. *Earth-Science Reviews*, 162, pp.227–252.

3. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., et al. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120, pp.25–36.
4. Prakash, N., Manconi, A. and Loew, S. (2020). Mapping landslides on EO data: performance of deep learning models vs. traditional machine learning models. *Remote Sensing*, 12(3), p.346.
5. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), pp.5–32.
6. Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), pp.273–297.
7. Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S.R., Tiede, D. and Aryal, J. (2019). Evaluation of different machine learning methods and deep learning convolutional neural networks for landslide detection. *Remote Sensing*, 11(2), p.196.
8. Long, J., Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3431–3440.
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *ICLR 2021*.
10. Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M. and Guzzetti, F. (2018). A review of statistically-based landslide susceptibility models. *Earth-Science Reviews*, 180, pp.60–91.
11. Hungr, O., Leroueil, S. and Picarelli, L. (2014). The Varnes classification of landslide types, an update. *Landslides*, 11(2), pp.167–194.
12. Petley, D. (2012). Global patterns of loss of life from landslides. *Geology*, 40(10), pp.927–930.
13. Segoni, S., Piciullo, L. and Gariano, S.L. (2018). A review of the recent literature on rainfall thresholds for landslide occurrence. *Landslides*, 15(8), pp.1483–1501.
14. Joyce, K.E., Belliss, S.E., Samsonov, S.V., McNeill, S.J. and Glassey, P.J. (2009). A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography*, 33(2), pp.183–207.
15. Morishita, Y., Lazecky, M., Wright, T.J., Weiss, J.R., Elliott, J.R. and Hooper, A. (2020). LiCSBAS: An open-source InSAR time series analysis package integrated with the LiCSAR automated Sentinel-1 InSAR processor. *Remote Sensing*, 12(3), p.424.
16. Turner, D., Lucieer, A. and De Jong, S.M. (2015). Time series analysis of landslide dynamics using an unmanned aerial vehicle (UAV). *Remote Sensing*, 7(2), pp.1736–1757.
17. Pourghasemi, H.R., Pradhan, B. and Gokceoglu, C. (2012). Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. *Natural Hazards*, 63(2), pp.965–996.
18. Youssef, A.M. and Pourghasemi, H.R. (2021). Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha Basin, Saudi Arabia. *Geoscience Frontiers*, 12(2), pp.639–655.
19. Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., et al. (2018). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, 151, pp.147–160.
20. LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), pp.436–444.

21. Ronneberger, O., Fischer, P. and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. MICCAI 2015, LNCS 9351, pp.234–241.
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
23. Merghadi, A., Yunus, A.P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D.T., et al. (2020). Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. *Earth-Science Reviews*, 207, p.103225.
24. Ji, S., Yu, D., Shen, C., Li, W. and Xu, Q. (2020). Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. *Landslides*, 17(6), pp.1337–1352.
25. Ding, H., Xu, Q., Dong, X., Zhang, S., Li, W. and Peng, D. (2022). A deep learning model for fast detection and classification of landslides from SAR imagery. *Remote Sensing of Environment*, 274, p.113000.
26. Bromley, J., Guyon, I., LeCun, Y., Sackinger, E. and Shah, R. (1994). Signature verification using a 'Siamese' time delay neural network. *Advances in Neural Information Processing Systems*, 6.
27. Chen, H. and Shi, Z. (2021). A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 13(7), p.1428.
28. Mondini, A.C., Guzzetti, F., Chang, K.T., Monserrat, O., Martha, T.R. and Manconi, A. (2021). Deep learning forecast of rainfall-induced shallow landslides. *Nature Communications*, 12(1), p.5462.
29. Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE CVPR*, pp.779–788.
30. Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W. and Zhao, T. (2020). Landslide detection and segmentation using remote sensing images and deep neural networks. *IEEE Access*, 8, pp.85920–85942.
31. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS 2021*.
32. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.4015–4026.
33. Wang, D., Zhang, J., Du, B., Xia, G.S. and Tao, D. (2023). An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61, pp.1–20.
34. Intrieri, E., Carlà, T. and Gigli, G. (2019). Forecasting the time of failure of landslides at slope-scale: A literature review. *Earth-Science Reviews*, 193, pp.333–349.
35. Fang, Z., Wang, Y., Peng, L. and Hong, H. (2022). Predicting flood susceptibility using LSTM neural networks. *Journal of Hydrology*, 594, p.125734.
36. Kirschbaum, D., Stanley, T. and Zhou, Y. (2015). Spatial and temporal analysis of a global landslide catalog. *Geomorphology*, 249, pp.4–15.
37. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, pp.18–27.
38. Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., et al. (2015). The climate hazards infrared precipitation with stations a

- new environmental record for monitoring extremes. *Scientific Data*, 2, p.150066.
39. Powers, D.M.W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), pp.37–63.
40. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp.321–357.
41. Pan, S.J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp.1345–1359.
42. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE CVPR*, pp.2117–2125.
43. Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
44. Zhao, W., Chen, J., Chen, W. and Hu, X. (2023). Multi-source remote sensing data fusion for landslide susceptibility evaluation. *IEEE Transactions on Geoscience and Remote Sensing*, 61, pp.1–15.
45. Huang, F., Yan, J., Fan, X., Yao, C., Huang, J., Chen, W. and Hong, H. (2023). Uncertainty pattern in landslide susceptibility prediction modelling: effects of different landslide boundaries and spatial shape expressions. *Geoscience Frontiers*, 13(2), p.101317.