LandsatQuake: A Large-Scale Dataset for Practical Landslide Detection

Vihaan Akshaay Rajendiran[1]*, **Amanda Roeliza Guilalas Hunt**[1]*, **Gen K. Li**[1]†, **Lei Li**[2]†
[1]University of California, Santa Barbara [2]Carnegie Mellon University
{vihaanakshaay,amandaroelizahunt,ligen}@ucsb.edu, leili@cs.cmu.edu

Abstract

Identifying landslides from remote imagery is critical for rapid responses after landslide hazards and for assessing their environmental impacts. Existing datasets for landslide detection models are mostly based on multi-sourced, high-resolution (e.g., 1-5 m) satellite imagery from commercial companies (e.g., Planet Labs) and ultra-high-resolution (e.g., < 1m) photos from unmanned aerial vehicle (UAV) surveys. However, obtaining those data is often economically expensive and labor-intensive, limiting their applicability. Those datasets may not fully characterize large, deep-seated landslides (>10,000 m^2) – a most damaging hazard – across broad mountain ranges (10s km x 100s km). Here we present 'LandsatQuake,' a benchmark dataset composed of 31 landslide inventories from 21 earthquake-prone regions across the world covering a total area of 5.56×10^7 km^2 and spanning the last 40 years. This dataset emphasizes practicality by using satellite images acquired by Landsat, the only satellite system that has recorded Earth's land surface for >40 years. The public availability, high coverage of the world, and longevity make the Landsat data ideal for developing historical and recent landslide inventories caused by known triggers (e.g., earthquakes or rainstorms). The moderate resolution (30 m) allows LandsatQuake to fully characterize large landslides (>10,000 m^2) at mountain range scales. Additionally, we demonstrate the challenges of applying existing computer vision algorithms to practical landslide detection problems by evaluating several baselines and multiple landslide datasets using metrics including the signal-to-area ratio. Finally, we provide digital elevation model (DEM) data for these regions and show the effectiveness of integrating DEM and slope data for landslide detection.

1 Introduction

Landslides represent a severe natural hazard. In the United States, landslides are estimated to cause 25-50 deaths annually and more than one billion dollars in damages [Froude and Petley, 2018, Schuster and Highland, 2001]. Landslides can induce other cascading hazards, such as aggradation of river channels and flooding [Peng et al., 2014]. Additionally, landslides are a primary agent of erosion, acting to drive long-term landform changes and biogeochemical cycles [Hilton et al., 2011, Emberson et al., 2016].

High-quality and timely landslide inventory maps are critical for mitigating and assessing landslide-induced damage. Most high-quality landslide inventories are produced from manual mapping, which

*Equal contribution.
†Corresponding authors.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

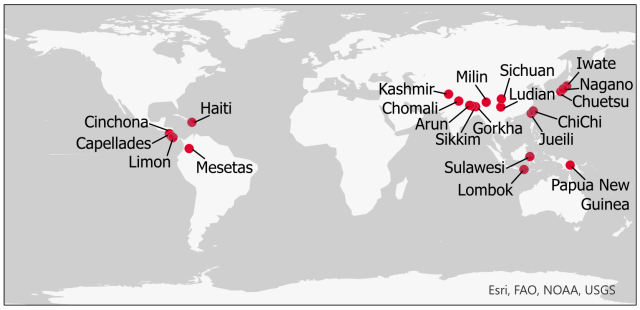


Figure 1: Map showing locations of all landslide inventories path of LandsatQuake

is labor-intensive and time-consuming [Galli et al., 2008]. We still lack ready open-source algorithms that allow efficient and rapid mapping of landslides after catastrophic landslide-triggering events. Automated landslide detection using machine learning is a rapidly growing field and provides a much faster alternative to manual mapping [Milledge et al., 2021].

Studies have shown that, in comparison to traditional machine learning methods, deep learning techniques can achieve superior performance when handling large-scale remote sensing data [qiang Yang et al., 2024]. Convolutional Neural Networks[Lecun et al., 1998] with the right architecture are shown to outperform methods like Artificial Neural Networks [Rosenblatt, 1958], Support Vector Machines [Cortes and Vapnik, 1995] and Random Forest [Breiman, 2001] for satellite images from the Himalayan region [Ghorbanzadeh et al., 2019]. Fully Convolutional Networks [Long et al., 2015], such as the U-Net [Ronneberger et al., 2015] are also commonly used for landslide detection as a semantic segmentation problem [Su et al., 2021]. Ullo et al. [2020] constructed landslide detection as an instance segmentation problem and used Mask-RCNN [He et al., 2018] with enhancements to achieve superior performance. Although there have been various attempts to provide new architectures and techniques for landslide detection, most of these focus only on specific regions and fail to generalize to a broader distribution of satellite images or regions.

In addition, current landslide detection models mainly rely on high-resolution satellite imagery (5-10 m resolution) and/or ultra-high-resolution aerial photos (<1 m resolution), which require purchase from commercial companies and/or labor-intensive field drone surveys [Xu et al., 2024]. High-resolution datasets only cover very recent timespans (2010-2015 onwards), limiting models to learn from existing landslide data from before 2010. Additionally, current landslide detection models are unable to separate amalgamated landslides, which heavily skews volume calculations and gives an inaccurate estimate of landslide hazard magnitudes [Larsen et al., 2010].

To address these challenges, we have developed a new dataset, **LandsatQuake (LQ)**. This dataset is specifically designed for practical landslide detection:

Broad Coverage and Historical Depth: LandsatQuake leverages imagery acquired by Landsat, the only satellite system that has consistently recorded Earth's land surface for over 40 years. This dataset provides extensive historical data across 21 landslide-active regions covering a total land area of 5.56×10^7 km^2. We focus on earthquake-impacted mountain ranges, given that earthquakes are a major trigger of landslides and can weaken rocks to induce landslides during storm events [Jones et al., 2021, Li et al., 2022a].

Practical Relevance: Landsate images have a moderate resolution (30 m) Landsat, allowing the LandsatQuake dataset to fully characterize large, deep-seated landslides (>10,000 m^2) over extensive mountain ranges (10s x 100s km).

Integration with DEM Data: We provide digital elevation model (DEM) data for the regions covered in LandsatQuake. Our experiments show the effectiveness of integrating DEM and slope data [Guzzetti et al., 2012] for improving landslide detection accuracy.

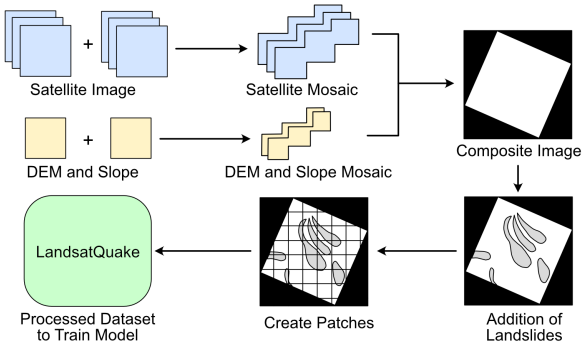


Figure 2: Workflow of LandsatQuake. Satellite images and DEMs are separately composited and converted to a mosaic. The DEM mosaic covers less area than the satellite mosaic. Both mosaics are then composited to a single image. Landslide polygons bounds are overlain on the image. Lastly, the composite is segmented into various patches before being read into the model.

Cost-Effective and Accessible Data: Unlike high-resolution commercial satellite images and ultra-high-resolution UAV photos, Landsat imagery is open-source and free from licensing constraints. This makes it economically feasible and easily accessible for researchers and practitioners.

Evaluation of Existing Models: Our evaluation of several traditional computer vision models on the LandsatQuake dataset demonstrates that their performance does not meet the practical requirements necessary for effective landslide detection. This highlights the necessity of developing new methodologies that address real-world challenges.

2 Related Work

In recent years, growing efforts have been spent in developing benchmark datasets for landsldie detection models. Here, we compile information about three recent landslide datasets:

(A) CAS Landslide Dataset is a collection of 3 band multi-sensor RGB images taken from nine distinct regions [Xu et al., 2024]. Satellite data were procured from Google Earth Engine (GEE), including images from Sentinel-2A and Landsat. This dataset also includes images from UAV drone surveys. Images have a ground resolution of 5 m or less. The landslide polygons were sourced from multiple open-source repositories.

(B) Landslide4Sense dataset [Ghorbanzadeh et al., 2022] was a competition setup created by the Institute of Advanced Research in Artificial Intelligence (IARAI). The dataset is comprised of images from Sentinel-2A and DEM and slope layers from the Advanced Land Observing Satellite (ALOS). The images used 14 bands, the first 12 corresponding to the bands from sentinel and the last 2 corresponding to DEM and Slope data from ALOS. Four regions were used in this study, and the dataset was split into training, test, and validation parts.

(C) HR-GLDD Meena et al. [2023] is a global dataset using generalized deep learning for landslide mapping on high-resolution satellite images. This dataset incorporates images from PlanetScope taken after 2016 with a ground resolution of 3 m. Images are composed of 4 bands: Red, Green, Blue, and Near Infrared. 13 landslide events from various triggers were incorporated into the study, half of which were triggered by earthquakes and the other half by rainfall.

These three studies use high-resolution (<10 m) images and cover the recent decade (2015 and onward). However, as pointed out in several of these studies, there is a lack of globally distributed datasets [Meena et al., 2023]. Satellite images with extensive global coverage tend to have a moderate resolution, whereas models trained on high-resolution data may not perform well on moderate-resolution data. Consequently, these models likely work best with a small, highly curated portion of available landslide inventories. These models will miss the nuances present in more realistic data. Another limitation is that two of these three datasets do not include topography data. Landslides

(a) Raw Image before Transformation

(b) Image after Contrast Stretch

Figure 3: Comparison of satellite image a) before and b) after histogram stretching. Several types of histogram stretches were tested (including linear, percent clip, and histogram equalized stretch). Best result was from a contrast stretch with a brightness-dampening factor.

generally occur on steep slopes and medium-to-high elevations, thus including digital topography data (slope and elevation) would give the model complementary information for landslide detection [Guzzetti et al., 2012].

3 Problem Definition & Dataset

3.1 Data Procurement

Satellite image data was downloaded from the USGS through EarthExplorer usg. Images were selected from $\pm$ 2 years from when an earthquake event occurred. Images with less than 20 percent cloud cover were chosen. Downloaded images were approximately 2 by 2 degrees in size depending on their location. Images came projected in local UTM coordinate zones. The study uses images from the Landsat 4, 5, and 8 missions, all of which have 30 m ground resolution. Landsat 7 was excluded from this study, as many images have artificial strips due to failure of its scan line corrector Storey et al. [2003]. Landsat 4 and 5 were formatted as 8 bit images, whereas Landsat 8 was packaged as 16 bit. To normalize our image format, we converted all 16-bit images to 8-bit.

DEM data was acquired from ALOS Global Digital Surface Model generated by the Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM) **?**. This dataset has a ground resolution of 30 m. DEM data were downloaded in 1-degree x 1-degree tiles. Downloaded images were in the EPSG 4326 coordinate system and were re-projected to local UTM coordinate zones.

The LandsatQuake dataset uses 31 landslide inventories from 21 major earthquake-impacted mountain regions (Table 1). The landslide inventories were sourced from the USGS earthquake-triggered landslide inventory and from literature Marc et al. [2016], Schmitt et al. [2017], Li et al. [2022a]. The polygons were processed to fix geometrical errors and projected to local UTM coordinates. Because the images were too coarse to see small landslides, we filtered landslides smaller than 10,000 m^2. In practice, the >10,000 m^2 landslides mostly correspond to deep-seated bedrock landslides that dominate total landslide volumes and have most significant impacts.

3.2 Dataset Construction

After compiling DEM and satellite imagery data for each region, we processed the data one region at a time. We first formed a mosaic of DEMs and, in parallel, composited the bands for each of the area patches in every region to form multi-band images. We then combined the DEM mosaics and satellite imagery into a single image file, with the landslide bounds overlain on top. The image was then segmented into patches of 224 pixel x 224 pixel. Lastly, we removed images with artifacts and high cloud coverage and input images with corresponding landslide coverage into the model.

The following Landsat bands were included in the dataset: Band 1—blue; Band 2—green; Band 3—red; Band 4 near-infrared (NIR); Band 5—short wave infrared 1 (SWIR 1); Band 7—short wave infrared 2 (SWIR 2). Band 6, the thermal band, was excluded from this study because it was not available for the Landsat 4 and 5 images.

Event Name	Satellite Mission	Total LS Polygon Area (m²)	Total Landslide-Impacted Landscape Area (m²)
1984 Nagano	Landsat 5	691,758.9	103,576,600.0
1987 Sichuan Pre-earthquake	Landsat 4 & 5	135,000,000.0	44,492,800,000.0
1991 Limon Earthquake	Landsat 4 & 5	3,191,348.0	1,890,711,000.0
1998 Jueili	Landsat 5	798,792.6	220,086,400.0
1999 Chamoli Earthquake	Landsat 5	13,925,520.0	7,271,280,000.0
1999 ChiChi Earthquake	Landsat 5	95,281,110.0	13,853,220,000.0
2004 Chūetsu Earthquake	Landsat 5	3,408,261.0	186,418,000.0
2005 Kashmir Earthquake	Landsat 7	105,000,000.0	4,624,910,000.0
2008 Iwate Earthquake	Landsat 5	7,400,152.0	550,880,400.0
2009 Cinchona	Landsat 7	13,774,950.0	315,402,500.0
2010a Haiti	Landsat 5	2,585,252.0	2,652,562,000.0
2010b Haiti	Landsat 5	8,082,003.0	3,636,516,000.0
2011 Sikkim Earthquake	Landsat 5	24,841,270.0	4,319,320,000.0
2014 Ludian Earthquake	Landsat 8	3,025,449.0	185,513,200.0
2015 Gorkha Earthquake	Landsat 8	47,225,690.0	33,912,650,000.0
2016 Capellades	Landsat 8	479,465.6	18,924,310.0
2017 Milin	Landsat 8	29,617,520.0	697,679,400.0
2018 Sulawesi (2021)	Landsat 8	14,237,000.0	3,561,889,000.0
2018 Papua New Guinea	Landsat 8	161,000,000.0	55,434,590,000,000.0
2018 Lombok (2019a)	Landsat 8	238,781.6	157,587,600.0
2019 Mesetas	Landsat 8	472,647.1	269,697,100.0
2011 Arun	Landsat 7	1,109,647.0	4,476,125,000.0
2012 Arun	Landsat 7	1,135,598.0	2,137,784,000.0
2013 Arun	Landsat 7 & 8	185,797.2	1,796,088,000.0
2014 Arun	Landsat 8	407,968.0	1,174,011,000.0
2015 Arun	Landsat 8	3,083,048.0	3,010,666,000.0
2016 Arun	Landsat 8	2,638,067.0	4,729,137,000.0
2017 Arun	Landsat 8	227,308.3	531,250,200.0
2018 Arun	Landsat 8	466,409.3	1,254,700,000.0
2019 Arun	Landsat 7 & 8	165,738.5	3,285,947,000.0
2020 Arun	Landsat 8	678,627.1	1,723,635,000.0

Table 1: LandsatQuake covers 31 inventories across 21 regions, with a total area of 5.56×10^7 km^2.

We then created natural-colored (or RGB) images to make them look realistic and closer to the original data distribution commonly used by human annotators to map landslides. Specifically, we composited Band 3 in channel 1, Band 2 in channel 2, and Band 1 in channel 3. After combining these channels, the image was still dark (Figure 2a). To enhance visibility, we processed each image with a transform function 3. The function applied a contrast stretch to the image, spreading the intensity values over a larger range. This allowed the mid-tone values to be more visible and reduced the strength of the lightest and darkest values. The images were still bright after the contrast stretch, so a dampening factor was applied to reduce the intensity of the brightest pixels (Figure 2b).

After processing the first three bands, we added bands 4,5,and 6 (Landsat Bands 4, 5, and 7). For the seventh band, we concatenated the DEM band crop obtained from the mosaic. We also added an eighth band as slope angles calculated from DEM in ArcGIS Pro, given that slope angles are a most common measure for landslide risks Guzzetti et al. [2012].

4 Experiments

4.1 Domain-specific Analysis

4.1.1 Single-class Object Detection/Instance Segmentation over Semantic Segmentation

In this section, we first convert all the datasets to object detection, similar to our dataset. Technically, it is possible to go from semantic segmentation to object detection, but this process is not lossless.

Multiple instances with overlaps will be considered as one object if we simply take the minimum and maximum bounds of a semantic cluster of pixels.

For landslide detection in particular, it is practically crucial and useful to separate instances explicitly. This is particularly important for the estimate of landslide volume, a basic measure of landslide hazard. The current approach to estimating landslide volume is based on empirical scaling between landslide area and landslide volume [Larsen et al., 2010]:

$$V = \alpha \times A^{\gamma} \tag{1}$$

where α and γ are empirical scaling coefficients, with γ being greater than 1 (around 1.5). For multiple landslides, their total volume is calculated as:

$$V_{\text{tot}} = \sum (\alpha \times A_i^{\gamma}) \tag{2}$$

However, landslides often occur in clusters where multiple landslides amalgamate [Marc and Hovius, 2015]. If the amalgamated landslides are mapped as one instance without being segregated into individual instances, the total volume would be wrongly estimated as:

$$V_{\text{tot\_w}} = \alpha \times \left(\sum A_i \right)^{\gamma} \tag{3}$$

Comparing Eqs. (2) and (3) indicates that without splitting instances of landslides (as in instance segmentation or object detection), the actual landslide volume would be largely biased, as demonstrated in field studies [Marc and Hovius, 2015, Li et al., 2014].

We emphasize that an accurate landslide volume estimate is important because it informs the amount of materials produced by landslides and the magnitude of landslide hazards, which are critical information for stakeholders and decision-makers and for dispatching rescue forces after landslide hazards occur. Landslide volumes also indicate the magnitude of landslide-induced erosion, which is important for assessing mountain growth, a fundamental topic in geoscience [Li et al., 2014].

4.1.2 Signal-to-Area Ratio (SAR)

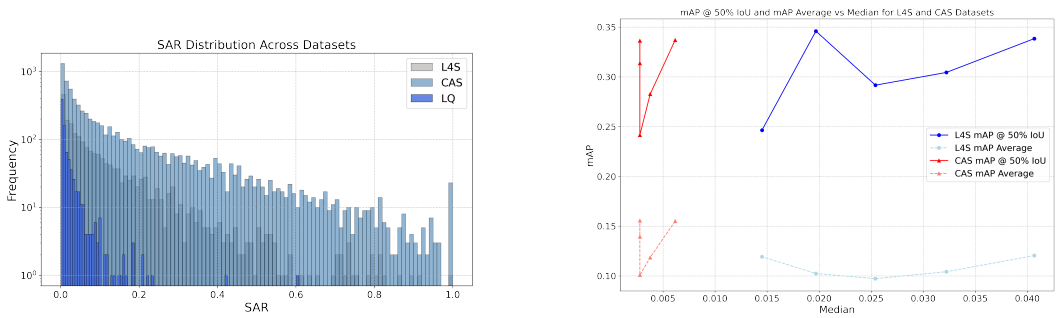
One way to capture information about landslide areas in computer vision applications is through the signal-to-area ratio (SAR), the ratio between landslide area and the total area of surrounding landscapes. Since instance-level information is not available in the previous datasets (L4S, CAS, and HR-GLDD), we calculate the SAR by summing the number of landslide pixels and dividing by the total area of the image (similar to the pixel ratio [Zhang et al., 2019]). This method allows us to provide a comprehensive comparison across all datasets. For CAS, we only consider images from satellite sources, excluding those from UAV surveys.

From Figure 4a, we see that LandsatQuake has narrow, clustered, and low SAR whereas L4S and CAS have more distributed SAR. LandsatQuake has significantly lower SAR than the other datasets.

We next prepared various object detection subsets and analyzed the mean Average Precision (mAP) by sampling the datasets to generate subsets with varying medians, [Everingham et al., 2010]]. The result mAP plotted against the median SAR of the datasets (Figure 4b). Although one would expect that models have better performance as the SAR increases, Figure 4b indicates that such a correlation does not show up clearly in these datasets.

Observations show that, for the same dataset, as the SAR increases, performance fluctuates but generally increases under high SAR. In addition, for CAS, subsets that have very similar median values nonetheless exhibit remarkably different mAP values, both on average and at the 50th percentile.

These observations suggest that factors other than SAR, such as color contrasts (e.g., landslides typically appear in lighter colors compared to the predominantly green surrounding vegetation), might have a more significant impact on model performance.



(a) SAR Distribution for different datasets

(b) mAP with Faster R-CNN (ResNet-50)

Figure 4: ((a) SAR distribution of three datasets: LQ (ours), CAS, and L4S, presented on a logarithmic scale for clearer comparison. (b) To align the number of images across datasets, L4S and CAS were divided into equi-distributional groups (2 and 7, respectively). By replacing the top x% of images from the primary group with the bottom x% from another group, we created subsets with varied medians, for which we report the mean Average Precision (mAP).

4.1.3 Expert Review

Identifying and quantifying all possible factors that affect landslide detection from satellite imagery may be challenging and impractical. To bypass this level of abstraction, we gathered direct input by showing both raw and annotated images to experts in landslide annotation. This approach ensures that any significant factors influencing landslide detection are inherently embedded in the feedback we receive from experts in the field, as they provide the primary landslide annotation data.

To conduct this study, we randomly sampled 100 images from each dataset (L4S, CAS, and LandsatQuake). After annotating these images based on the truth labels, we shuffled them to prepare for evaluation. We then presented these images to experts, asking them to assign a score between 1 and 10 for each image to represent the ease of identifying landslides, where 1 indicates the hardest/most likely to miss and 10 denotes the easiest/most likely to detect. This study was conducted with three experts, and the results are reported in Table 2 below.

Score	Expert A	Expert B	Expert C	Average
CAS	5.61	7.60	6.03	6.41
Landslide4Sense	6.9	8.39	4.78	6.69
LandsatQuake	**2.0**	**1.81**	**1.12**	**1.64**

Table 2: Scores of three GeoScience experts and the average scores for each dataset

The scores reveal drastic differences in the difficulty level of identifying landslides between LandsatQuake and other existing datasets. This highlights the challenges associated with using open-source Landsat imagery for landslide detection in contrast to high-resolution images from previous deep learning-based datasets.

4.2 Models

Acknowledging the difficulty in detecting landslides from Landsat imagery, we measured the performance of various existing computer vision models to assess how current CV methods struggle with realistic, moderate-resolution datasets such as LandsatQuake.

To quantify these challenges, we trained Faster R-CNN [Ren et al., 2016] object detection models using PyTorch with various feature-extracting backbones. The dataset was prepared by segregating images and labels into training, validation, and test datasets with a split of 80%, 10%, and 10%, respectively, with data configurations loaded from YAML files. Since our data is very different in

Model (Backbone)	mAP (Average)	mAP @ 50% IoU (mAP50)
ConvNeXt Small [Liu et al., 2022]	0.000018	0.000090
ConvNeXt Tiny [Liu et al., 2022]	0.000064	0.000319
EfficientNet [Tan and Le, 2020]	0.000081	0.000202
Darknet [Redmon, 2013–2016]	0.000990	0.004950
ResNet18 [He et al., 2015]	0.000021	0.000140
ResNet50 FPN [He et al., 2015]	**0.002975**	**0.010006**
ResNet101 [He et al., 2015]	0.000033	0.000083
ResNet152 [He et al., 2015]	0.000024	0.000066
SqueezeNet 1.0 [Iandola et al., 2016]	0.000073	0.000163
SqueezeNet 1.1 [Iandola et al., 2016]	0.000043	0.000163
Swin [Liu et al., 2021b]	0.000040	0.000146
ViTDet Tiny [Li et al., 2022b]	0.000017	0.000086
ViTDet [Li et al., 2022b]	0.000043	0.000143

Table 3: Comparison of mAP scores for Faster R-CNN with different backbones evaluated on LQ.

distribution from most standard pre-training datasets such as ImageNet [Krizhevsky et al., 2012] or Common Objects in Context [Lin et al., 2015], we evaluated our models by using backbones trained from scratch. We chose a batch size of 8 and 20 training epochs across all different configurations (backbones), and optimization was performed using the Stochastic Gradient Descent (SGD) [LeCun et al., 1998] optimizer with momentum 0.9 and a learning rate of 0.001. Model performance was evaluated based on mean Average Precision (mAP) at different Intersection-over-Union (IoU) [Everingham et al., 2010] thresholds (specifically from 0.5 to 0.95 in steps of 0.05), and models were saved according to the best validation mAP observed. We have reported the mAP average and mAP @ IoU=50 values for various transformer-based and convolution-based backbones in Table 4.

The results show that traditional computer vision models, which are optimized and improved iteratively for standard deep learning benchmarks, perform poorly on LandsatQuake. Thus, applying existing models to Landsat-based data remains challenging.

4.3 Effect of Extra Spectral Bands

All experiments conducted thus far have utilized only RGB bands to ensure consistency in the evaluation of datasets (since the sources are typically different satellites with different bands) and to align with models typically developed for RGB datasets. In this section, we assess the utility of additional bands from our dataset by employing a Faster R-CNN model with a ResNet-50 backbone, modified to accept 6 input channels instead of 3. The model was trained and validated using custom means and standard deviations for normalization as [0.0]s and [1.0]s, respectively, for a total of 20 epochs to maintain uniformity across experiments. We utilized an Adam [Kingma and Ba, 2017] optimizer with a learning rate of 0.0005 and included weight decay for regularization. A ReduceLROnPlateau learning rate scheduler was applied to adjust the learning rate based on validation loss, with a patience of 5 epochs and a reduction factor of 0.1. The dataset was split into training, validation, and test sets, each loaded with a batch size of 8. We initially mask Channels 4, 5, and 6 (i.e., Landsat Bands 4 (NIR), 5(SWIR 1), and 7 (SWIR 2)). Subsequently, we evaluate the model's performance by unmasking one channel at a time and reporting the mAP scores.

DEM and slope data are commonly considered useful for landslide detection [Guzzetti et al., 2012, Wang et al., 2021]. Therefore, we incorporated these as the 7th and 8th bands in our dataset. To evaluate the comprehensive performance using all these bands, we performed a similar assessment, utilizing 8 input bands instead of 6, and compared the performance with various configurations of masked data.

Our results (Table 4) indicate a general trend of improved performance with the inclusion of more bands and information. Although this trend is not consistently reflected in the mAP average, the

#Bands	mAP (Average)	mAP @ 50% IoU (mAP50)
3 bands	0.0009197	0.003822
4 bands	0.0010123	0.001817
5 bands	**0.0077761**	0.022527
6 bands	0.0053710	0.012381
DEM (8 bands)	0.0034852	**0.034486**

Table 4: Comparison of Faster R-CNN (ResNet-50) with different number of input bands.

mAP at 50% IoU provides a more practical metric for comparison. This metric is more forgiving and considers predictions that are sufficiently accurate to be practically meaningful. We note a stronger correlation between the number of bands and performance, with the addition of DEM bands significantly enhancing the model's performance measured by mAP at 50% IoU.

5 Limitations and Future Work

First, due to the dataset's complexity, most out-of-the-box computer vision models struggle to learn effectively from LandsatQuake. This difficulty hampers the ability to conduct analysis of learned models [Grün et al., 2016] for insights, making it challenging to understand the models and visualize the internal representation of features.

Second, despite improvements in data quality and satellite technology over time, our dataset relies on shapefiles for incidents from various years in the past. Consequently, the only available source of satellite images for those old events is often Landsat (which is open-source and has a timespan of over 40 years [Wulder et al., 2022]), either due to licensing restrictions or because it is the sole source available for older events.

Thirdly, the continuing evolution of satellite technology [García-Arenal and Fraile, 2017] makes it challenging to provide a universal format that can integrate our Landsat-based dataset with those using images from other sources.

Finally, although the inclusion of DEM bands improves performance, the typical mAP scores for small object detection problems [Liu et al., 2021a] and sparse datasets, including those in aerial imagery [Koyun et al., 2022], can be up to an order of magnitude higher. This emphasizes the need for more specific methods for landslide detection.

6 Conclusion

We present LandsatQuake, a large-scale dataset for landslide detection. This dataset covers 21 landslide-active regions and 31 landslide inventories from the last 40 years and utilizes the open-access Landsat imagery, offering historical data across various global locations. It is particularly valuable as it mirrors the real-world conditions of landslide detection, using only openly accessible and widely available imagery. We tested several traditional computer vision models on this dataset and showed that their performance does not meet the practical requirements necessary for effective implementation. This initiative highlights the necessity of developing methodologies and datasets that not only refine traditional performance metrics but also address practical and realistic challenges that are pressing and crucial, such as those in landslide detection.

References

Landsat-4, Landsat-5, and Landsat-8 images courtesy of the U.S. Geological Survey.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

Robert Emberson, Niels Hovius, Albert Galy, and Odin Marc. Chemical weathering in active mountain belts controlled by stochastic bedrock landsliding. *Nature Geoscience*, 9:42–45, 2016. URL https://api.semanticscholar.org/CorpusID:131225387.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

M. J. Froude and D. N. Petley. Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, 18:2161–2181, 2018. doi: 10.5194/nhess-18-2161-2018.

Mirco Galli, Francesca Ardizzone, Mauro Cardinali, Fausto Guzzetti, and Paola Reichenbach. Comparing landslide inventory maps. *Geomorphology*, 94(3):268–289, 2008. ISSN 0169-555X. doi: https://doi.org/10.1016/j.geomorph.2006.09.023. GIS technology and models for assessing landslide hazard and risk.

Fernando García-Arenal and Aurora Fraile. Chapter 56 - origin and evolution of satellites. In Ahmed Hadidi, Ricardo Flores, John W. Randles, and Peter Palukaitis, editors, *Viroids and Satellites*, pages 605–614. Academic Press, Boston, 2017. ISBN 978-0-12-801498-1. doi: https://doi.org/10.1016/B978-0-12-801498-1.00056-5. URL https://www.sciencedirect.com/science/article/pii/B9780128014981000565.

Omid Ghorbanzadeh, Thomas Blaschke, Khalil Gholamnia, Sansar Raj Meena, Dirk Tiede, and Jagannath Aryal. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, 11(2), 2019. ISSN 2072-4292. doi: 10.3390/rs11020196. URL https://www.mdpi.com/2072-4292/11/2/196.

Omid Ghorbanzadeh, Yonghao Xu, Pedram Ghamisi, Michael Kopp, and David Kreil. Landslide4sense: Reference benchmark data and deep learning models for landslide detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. doi: 10.1109/TGRS.2022.3215209.

Felix Grün, Christian Rupprecht, Nassir Navab, and Federico Tombari. A taxonomy and library for visualizing learned features in convolutional neural networks, 2016.

Fausto Guzzetti, Alessandro Cesare Mondini, Mauro Cardinali, Federica Fiorucci, Michele Santangelo, and Kang-Tsung Chang. Landslide inventory maps: New tools for an old problem. *Earth-Science Reviews*, 112(1):42–66, 2012. ISSN 0012-8252. doi: https://doi.org/10.1016/j.earscirev.2012.02.001. URL https://www.sciencedirect.com/science/article/pii/S0012825212000128.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

Robert G. Hilton, Patrick Meunier, Niels Hovius, Peter J. Bellingham, and Albert Galy. Landslide impact on organic carbon cycling in a temperate montane forest. *Earth Surface Processes and Landforms*, 36(12):1670–1679, 2011. doi: 10.1002/esp.2191.

Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, 2016.

J. N. Jones, S. J. Boulton, M. Stokes, G. L. Bennet, and M. R. Z. Whitworth. 30-year record of himalaya mass-wasting reveals landscape perturbations by extreme events. *Nature Communications*, 12(1):6701, 2021. doi: 10.1038/s41467-021-26964-8. URL https://doi.org/10.1038/s41467-021-26964-8.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Onur Can Koyun, Reyhan Kevser Keser, İbrahim Batuhan Akkaya, and Behçet Uğur Töreyin. Focus-and-detect: A small object detection framework for aerial images. *Signal Processing: Image Communication*, 104:116675, May 2022. ISSN 0923-5965. doi: 10.1016/j.image.2022.116675. URL http://dx.doi.org/10.1016/j.image.2022.116675.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 2012.

I. Larsen, D. Montgomery, and O. Korup. Landslide erosion controlled by hillslope material. *Nature Geoscience*, 3:247–251, 2010. doi: 10.1038/ngeo776. URL https://doi.org/10.1038/ngeo776.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 1998.

G. K. Li, S. Moon, and J. T. Higa. Residence time of over-steepened rock masses in an active mountain range. *Geophysical Research Letters*, 49(7):e2021GL097319, 2022a. doi: 10.1029/2021GL097319.

Gen Li, A. Joshua West, Alexander L. Densmore, Zhangdong Jin, Robert N. Parker, and Robert G. Hilton. Seismic mountain building: Landslides associated with the 2008 wenchuan earthquake in the context of a generalized model for earthquake volume balance. *Geochemistry, Geophysics, Geosystems*, 15(2):233–249, January 2014. doi: 10.1002/2013GC005067. URL https://doi.org/10.1002/2013GC005067.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection, 2022b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Yang Liu, Peng Sun, Nickolas Wergeles, and Yi Shang. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172:114602, 2021a. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.114602. URL https://www.sciencedirect.com/science/article/pii/S0957417421000439.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. URL https://arxiv.org/abs/2201.03545.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

O. Marc and N. Hovius. Amalgamation in landslide maps: effects and automatic detection. *Natural Hazards and Earth System Sciences*, 15:723–733, 2015. doi: 10.5194/nhess-15-723-2015. URL https://doi.org/10.5194/nhess-15-723-2015.

Odin Marc, Niels Hovius, Patrick Meunier, Tolga Gorum, and Taro Uchida. A seismologically consistent expression for the total area and volume of earthquake-triggered landsliding. *Journal of Geophysical Research: Earth Surface*, 121(4):640–663, 2016. doi: 10.1002/2015JF003732.

S. R. Meena, L. Nava, K. Bhuyan, S. Puliero, L. P. Soares, H. C. Dias, M. Floris, and F. Catani. Hr-gldd: a globally distributed dataset using generalized deep learning (dl) for rapid landslide mapping on high-resolution (hr) satellite imagery. *Earth System Science Data*, 15(7):3283–3298, 2023. doi: 10.5194/essd-15-3283-2023. URL https://essd.copernicus.org/articles/15/3283/2023/.

David G. Milledge, Dino G. Bellugi, Jack Watt, and Alexander L. Densmore. Automated landslide detection outperforms manual mapping for several recent large earthquakes. *Natural Hazards and Earth System Sciences*, 2021. doi: 10.5194/nhess-2021-168. URL https://doi.org/10.5194/nhess-2021-168.

M. Peng, L.M. Zhang, D.S. Chang, and Z.M. Shi. Engineering risk mitigation measures for the landslide dams induced by the 2008 wenchuan earthquake. *Engineering Geology*, 180:68–84, 2014. ISSN 0013-7952. doi: https://doi.org/10.1016/j.enggeo.2014.03.016. Special Issue on "The geological and geotechnical hazards of the 2008 Wenchuan earthquake, China: Part I".

Zhi qiang Yang, Wen wen Qi, Chong Xu, and Xiao yi Shao. Exploring deep learning for landslide mapping: A comprehensive review. *China Geology*, 7(2):330–350, 2024. ISSN 2096-5192. doi: https://doi.org/10.31035/cg2024032. URL https://www.sciencedirect.com/science/article/pii/S2096519224001137. Special issue on Landslide Monitoring, Early Warning, and Risk Assessment.

Joseph Redmon. Darknet: Open source neural networks in c. http://pjreddie.com/darknet/, 2013–2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. doi: 10.1037/h0042519.

Robert G. Schmitt, Hakan Tanyas, M. Anna Nowicki Jessee, Jing Zhu, Katherine M. Biegel, Kate E. Allstadt, Randall W. Jibson, Eric M. Thompson, Cees J. van Westen, Hiroshi P. Sato, David J. Wald, Jonathan W. Godt, Tolga Gorum, Chong Xu, Ellen M. Rathje, and Keith L. Knudsen. An open repository of earthquake-triggered ground-failure inventories. Data Series 1064, US Geological Survey, Reston, VA, 2017. URL https://doi.org/10.3133/ds1064.

Robert L. Schuster and Lynn M. Highland. Socioeconomic and environmental impacts of landslides in the western hemisphere. Open-File Report 2001-276, US Geological Survey, 2001. URL https://doi.org/10.3133/ofr01276.

James Storey, Pasquale Scaramuzza, Gail Schmidt, and Julia Barsi. LANDSAT 7 SCAN LINE CORRECTOR-OFF GAP-FILLED PRODUCT DEVELOPMENT, 2003. Work performed under USGS contract 03CRCN0001.

Zhi Su, Joe K. Chow, Ping-Seng Tan, and et al. Deep convolutional neural network–based pixel-wise landslide inventory mapping. *Landslides*, 18:1421–1443, 2021. doi: 10.1007/s10346-020-01557-6. Received 02 December 2019; Accepted 06 October 2020; Published 28 October 2020; Issue Date April 2021.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

Silvia Liberata Ullo, Amrita Mohan, Alessandro Sebastianelli, Shaik Ejaz Ahamed, Basant Kumar, Ramji Dwivedi, and G. R. Sinha. A new mask r-cnn based method for improved landslide detection, 2020.

Haojie Wang, Limin Zhang, Kesheng Yin, Hongyu Luo, and Jinhui Li. Landslide identification using machine learning. *Geoscience Frontiers*, 12(1):351–364, 2021. ISSN 1674-9871. doi: https://doi.org/10.1016/j.gsf.2020.02.012. URL https://www.sciencedirect.com/science/article/pii/S1674987120300542.

Michael A. Wulder, David P. Roy, Volker C. Radeloff, Thomas R. Loveland, Martha C. Anderson, David M. Johnson, Sean Healey, Zhe Zhu, Theodore A. Scambos, Nima Pahlevan, Matthew Hansen, Noel Gorelick, Christopher J. Crawford, Jeffrey G. Masek, Txomin Hermosilla, Joanne C. White, Alan S. Belward, Crystal Schaaf, Curtis E. Woodcock, Justin L. Huntington, Leo Lymburner, Patrick Hostert, Feng Gao, Alexei Lyapustin, Jean-Francois Pekel, Peter Strobl, and Bruce D. Cook. Fifty years of landsat science and impacts. *Remote Sensing of Environment*, 280:113195, 2022. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2022.113195. URL https://www.sciencedirect.com/science/article/pii/S0034425722003054.

Yezhou Xu, Chen Ouyang, Qiang Xu, et al. Cas landslide dataset: A large-scale and multisensor dataset for deep learning-based landslide detection. *Scientific Data*, 11:12, 2024. doi: 10.1038/s41597-023-02847-z. Received 11 August 2023; Accepted 11 December 2023; Published 02 January 2024.

R. Zhang, D. Yin, J. Ding, et al. A detection method for low-pixel ratio object. *Multimedia Tools and Applications*, 78:11655–11674, 2019. doi: 10.1007/s11042-018-6653-6. URL https://doi.org/10.1007/s11042-018-6653-6.