Harnessing Machine Learning for Rain Induced Landslide Detection and Analysis

Shelly Gupta[1,3(✉)], Hussain Otudi[1,2,3], Ameen Abdel Hai[1,3], Rafaa Aljurbua[1,3], Jovan Andjelkovic[1,3], Abdulrahman Alharbi[1,2,3], and Zoran Obradovic[1,3]

[1] Center for Data Analytics and Biomedical Informatics, Computer and Information Science Department, Temple University, Philadelphia, PA, USA
{shelly.gupta,hussain.otudi,aabdelhai,rafaa.aljurbua, jovan.andjelkovic,adalharbi,zoran.obradovic}@temple.edu

[2] Department of Computer Science, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia

[3] Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia

Abstract. Landslides pose significant risks to infrastructure, ecosystems, and human lives, making accurate prediction crucial for disaster preparedness and mitigation. We integrate multimodal environmental data to enhance landslide prediction using machine learning. Specifically, we combine temporal weather data from ASOS, static vegetation data from NLCD, static soil composition data from SOLUS100, and temporal soil attributes from ERA5-Land to estimate landslide probability within a 5 km radius of ASOS weather stations across six U.S. states. We frame this as a multiclass classification problem, predicting high, low, or no landslide probability. Given the inherent imbalance in landslide occurrence, we explore various techniques such as SMOTE oversampling, class-weighted training, and dimensionality reduction to improve model performance. Our results indicate that XGBoost trained on SMOTE-balanced, PCA-reduced data incorporating all four datasets achieves the highest macro F1-score of 0.70. Analysis of feature importance reveals that significant predictors span all datasets, highlighting the necessity of integrating diverse environmental variables. Additionally, we conduct state-wise and seasonal comparisons to assess regional variations in model effectiveness. This research demonstrates the potential of multimodal data fusion and machine learning in landslide forecasting, paving the way for more robust and interpretable predictive models for natural hazard assessment.

Keywords: Landslides · Multimodal · Machine Learning · Feature selection

1 Introduction

Landslides triggered by heavy rainfall pose a significant threat to public infrastructure, resulting in substantial human, material, and economic losses world-

L. Iliadis et al. (Eds.): EANN 2025, CCIS 2582, pp. 94–108, 2025.
https://doi.org/10.1007/978-3-031-96199-1_8

wide [12]. Detecting these shallow, rainfall-induced landslides is critical for effective disaster management and risk mitigation. These events are influenced by a complex interplay of natural factors, including topography, geology, soil composition, and climatic conditions. Understanding the multimodal mechanisms driving landslide formation is, therefore, essential for developing predictive models and mitigation strategies.

Machine learning and deep neural networks have been shown to be useful in many geo-based applications, specifically landslides. Many works have focused on change detection-based algorithms [7], but it is resource-intensive. It requires both pre-event and post-event images, which can be difficult to obtain in a timely manner, particularly in regions with persistent cloud cover or adverse weather conditions. Similar aerial images have been used to detect landslides, but they have proven ineffective in areas with heavy vegetation. In Minnesota, with 35% forest cover, and in Wisconsin, with 48.98% forest cover [1], a significant portion of the land is hidden from aerial view, thus making imaging ineffective. Even with high-resolution imagery, the pixel size may not be sufficient to detect smaller landslides, which tree canopies can hide. Some studies have utilized ground-based measurements to assess landslide occurrences. These approaches leverage environmental and geotechnical data, such as soil properties, precipitation levels, and terrain characteristics, to predict landslide susceptibility [13]. However, most of these studies focus on regions outside the United States and often lack a detailed analysis of the factors influencing their models' decision-making processes. Understanding which environmental variables contribute the most to landslide predictions is crucial for improving the interpretability and reliability of the model, particularly in diverse regions.

In this study, we integrate multimodal data sources to enhance landslide prediction by leveraging both static and temporal environmental variables. Specifically, we incorporate temporal weather data from the Automated Surface Observing Systems (ASOS) [4], static vegetation data from the National Land Cover Database (NLCD) [11], static soil composition data from the Soil Landscapes of the United States (SOLUS) [16], and temporal soil attributes from ERA5-Land [15]. By combining these diverse datasets, we aim to develop a robust machine learning framework to estimate landslide probability within a 5 km radius of ASOS weather station coordinates across six U.S. states: Minnesota, North Dakota, South Dakota, Iowa, Michigan, and Wisconsin. We frame this as a multiclass classification problem, where each instance is categorized into one of three classes: high probability of landslide, low probability of landslide, or no landslide. However, since landslides are rare events, our dataset is highly imbalanced, which presents challenges for predictive modeling. To address this, we explore various imbalance mitigation techniques and analyze their impact on classification performance. In this work, we make the following key contributions:

- We integrate diverse data sources, combining meteorological, soil, and vegetation attributes to improve landslide prediction accuracy.
- We employ multiple traditional and deep learning models to classify landslide probability while addressing challenges posed by data imbalance.

- We conduct a dimension reduction analysis to assess whether reducing feature space improves classification performance.
- We analyze model explainability by identifying the most influential features contributing to landslide prediction.
- We compare model performance across different states and seasons to evaluate whether localized models outperform general predictive models.

Through this comprehensive approach, we aim to provide insights into the key environmental factors influencing landslides and assess the effectiveness of different modeling strategies for handling highly imbalanced geospatial data.

2 Related Works

Landslide prediction has been explored using various methodologies, including traditional machine learning, deep learning, and heuristic-based approaches. However, many of these studies rely solely on image-based models, overlooking the integration of multimodal environmental data or focusing on geographically constrained regions, limiting their generalizability.

Several studies have employed machine learning techniques on aerial imagery for landslide detection. For instance, traditional machine learning models were used but primarily focused on images, overlooking other environmental factors such as soil composition, vegetation, and meteorological data [10]. While effective in certain regions, this approach lacks robustness in areas with dense vegetation or limited image availability due to cloud cover or adverse weather conditions.

Other studies have focused on heuristic-based models rather than machine learning. In particular, Weather and Research Forecast (WRF) models were employed alongside geomorphological features to detect landslides [20]. However, that approach only considered two classes (landslide and no landslide) and did not incorporate machine learning techniques, limiting the ability to adapt to complex, nonlinear relationships within the data.

Deep learning models have also been used for landslide prediction, particularly in image-based approaches. PlanetScope, Sentinel-2 imagery, and ALOS-PALSAR2 elevation models were leveraged to predict landslides using convolutional neural networks (CNNs) [19]. However, this method is both costly and ineffective in regions with dense forest cover, such as Wisconsin, where tree canopies obscure terrain features critical for landslide detection.

Hydrological models have also been explored, but their applicability remains uncertain across different geographic regions. A hydrological model was developer for landslide prediction in Chile [9], but the findings may not generalize well to other regions, including the Midwestern United States, due to differences in terrain, climate, and geological features. The challenges of generalizing landslide prediction models across different geographical areas were further emphasized, highlighting the need for localized datasets and models tailored to specific environmental conditions [8]. Additionally, the prior study lacked soil moisture measurements and relied on traditional data collection methods instead of real-time IoT sensors, which could improve predictive accuracy. Multimodal studies have been shown to help in various weather-based problems [2,17,18]. Future advancements in landslide modeling, such as those proposed by [14], suggest integrating IoT sensor networks to enhance data granularity and timeliness.

Some research has explored alternative approaches, such as clustering and early warning systems. [6] investigated landslide clustering, but their work had limited application as a real-time warning system and did not analyze the explainability of contributing factors. Similarly, an initial landslide detection model was developed and provided insights into contributing factors [5]. Still, the study was confined to Seattle, Washington, and did not account for vegetation data, which plays a critical role in stabilizing slopes and influencing landslide susceptibility. Our work bridges the gap between traditional and deep learning approaches by proving that multimodal data fusion enhances prediction accuracy and generalizability across different geographic regions.

3 Data

In this study, we integrate multimodal data from multiple sources to enhance the accuracy of landslide detection. Specifically, we incorporate temporal weather data from the Automated Surface Observing Systems (ASOS), static vegetation data from the National Land Cover Database (NLCD), static soil composition data from the Soil Landscapes of the United States (SOLUS100), and temporal soil attributes data from ERA5-Land. Each of these datasets contributes unique and complementary information relevant to landslide susceptibility.

Our objective is to predict the probability of landslides using labels derived from the NASA Global Landslide Nowcast database, which provides high-resolution landslide hazard predictions. By integrating diverse data sources, we aim to capture environmental factors influencing landslide occurrence. The following subsections detail each data source, including its attributes and spatial and temporal resolution.

3.1 NASA Global Landslide Nowcast Dataset

The NASA Global Landslide Nowcast, generated by the Landslide Hazard Analysis for Situational Awareness (LHASA) model, identifies regions with a high likelihood of landslide occurrence on a daily basis [3]. LHASA integrates satellite-based precipitation estimates with a landslide susceptibility map to assess risk. While the model has the capability to run every 30 min, the dataset used in this study consists of a daily archive derived from past runs, covering latitudes from 60°N to 60°S.

For this analysis, we retrieved data from 2018 to 2020. Given a specific coordinate, we extracted relevant raster files and identified all data points within a 5 km radius, using the Haversine distance formula to account for Earth's curvature. The Haversine formula is given by:

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right) \tag{1}$$

where d is the great-circle distance between two points, R is Earth's radius (approximately 6371 km), φ_1, φ_2 are the latitudes of the two points in radians, λ_1, λ_2 are the longitudes of the two points in radians, $\Delta\varphi = \varphi_2 - \varphi_1$ and $\Delta\lambda = \lambda_2 - \lambda_1$. After obtaining the relevant points, we calculated the landslide probability by dividing the total number of labeled landslide occurrences by the total number of labels. A probability greater than 0.5 was classified as a high probability of landslide occurrence, while a probability between 0 and 0.5 was considered low. If no occurrences were detected, the area was classified as having no landslide risk.

3.2 Automated Surface Observing Systems Dataset

This study seeks to detect landslides occurring within a 5 km radius of an Automated Surface Observing Systems (ASOS) weather station. ASOS is the primary surface weather network in the United States, providing critical meteorological data that are essential for aviation safety [4]. The network consists of over a thousand weather stations distributed nationwide. For this analysis, we identified ASOS sensors located in Minnesota, Wisconsin, North Dakota, South Dakota, Iowa, and Michigan. We extracted various meteorological variables, including air temperature, dew point temperature, wind speed (in knots), one-hour precipitation, pressure altimeter, wind gust speed (in knots), sky altitude at various levels, ice accretion over 1, 3, and 6 h, and peak wind gust speed.

The ASOS data is recorded at varying temporal resolutions, ranging from 1-minute to 5-minute intervals. Given the focus of this study on landslide detection at the end of each day, it was necessary to aggregate the data to a daily resolution. This aggregation was performed by calculating each attribute's mean, standard deviation, minimum, and maximum values. The daily summary statistics provide a standardized representation of the weather conditions over time, aligning the temporal scale of the weather data with that of the landslide occurrences.

3.3 ERA5-Land Dataset

ERA5-Land is a high-resolution reanalysis dataset that consistently represents land surface variables over time with a spatial resolution of approximately 9 km [15]. The dataset offers vertical coverage extending from 2 m above the surface to a soil depth of 289 cm and is available at an hourly temporal resolution. For this study, we collected 28 meteorological and land surface attributes per location. These attributes include forecast albedo, snow layer temperature, soil temperature at various depths, surface temperature, skin reservoir content, dew point temperature, and air temperature at 2 m above the surface. Additionally, we extracted wind speed in different directions, surface pressure, leaf area index, soil water volume at multiple depths, and snow-related variables such as cover, mass, and depth. Furthermore, we incorporated lake-related parameters, including depth, temperature, and the thickness of the uppermost lake layer.

For this study, we acquired ERA5-Land data for the six selected states and refined the dataset by retaining only the data points located within a 5 km radius of ASOS weather stations. To ensure consistency with the temporal scale of the analysis, the hourly data was aggregated to a daily resolution. This was achieved by computing each attribute's mean, standard deviation, minimum, and maximum values, providing a comprehensive statistical representation of daily variations in land surface and meteorological conditions.

3.4 Soil Landscapes of the United States 100-Meter Dataset

The Soil Landscapes of the United States (SOLUS) dataset provides high resolution (100 m) maps documenting 20 key soil properties [16]. These properties include bulk density (oven dry), calcium carbonate content, cation exchange capacity, clay content, coarse sand, electrical conductivity (saturated paste), effective cation exchange capacity, fine sand, gypsum, medium sand, pH, rock content, total sand content, sodium adsorption ratio, silt content, soil organic carbon, very coarse sand, very fine sand and depth to bedrock.

The dataset is stored as raster files and covers the conterminous United States. Soil properties are documented at multiple depths, specifically at 0, 5, 15, 30, 60, 100, and 150 cm. For this study, we extracted soil property data for locations within a 5 km radius of ASOS weather stations. To maintain consistency with other datasets, we aggregated these attributes using the same approach, computing each property's mean, standard deviation, minimum, and maximum values.

3.5 National Land Cover Database

The National Land Cover Database (NLCD) provides comprehensive land cover classification and tree canopy cover data for the United States, offering area coverage statistics at the county level for 20 land cover categories [11]. These categories include open water, perennial ice/snow, developed land (ranging from open space to high intensity), barren land, various forest types (deciduous, evergreen, and mixed), shrubland, grassland, pasture/hay, cultivated crops, and wetland types (woody wetlands and emergent herbaceous wetlands). To integrate NLCD data into this study, we identified the county corresponding to each ASOS weather station and mapped the respective land cover attributes accordingly.

4 Experimental Setup

We integrate multiple datasets, including soil composition data (SOLUS100), vegetation data (NLCD), soil attributes (ERA5), and weather data (ASOS), to detect landslides within a 5 km radius of an ASOS weather station. The weather stations are distributed across six states: Minnesota, North Dakota, South Dakota, Wisconsin, Michigan, and Iowa. Initially, we identified 360 weather stations; however, we narrowed our analysis to locations that experienced at least one landslide event between 2018 and 2020, ultimately reducing the dataset to 43 stations.

We performed a temporal split, using all records from 2018 and 2019 for training and those from 2020 for testing. We also partitioned the training set, allocating 80% for model training and 20% for validation. We frame this task as a multiclass classification problem, where each instance is assigned one of three labels: '0' (no landslide), '1' (low probability of landslide), and '2' (high probability of landslide). A key challenge in this study is the class imbalance in the dataset: out of a total of 46,904 samples, 43,864 instances (93.51%) are labeled as '0', 2,970 instances (6.33%) are labeled as '1', and only 70 instances (0.14%) are labeled as '2'.

In this study, we experiment with various machine learning models to assess their performance in detecting landslides. The models tested include ensemble methods such as Random Forest, LightGBM, and XGBoost, as well as simpler linear models like logistic regression and Support Vector Classifier (SVC). Random Forest, an ensemble learning technique, is known for its robustness and ability to handle non-linear relationships, while LightGBM and XGBoost are gradient-boosting methods that are effective in handling large datasets with complex interactions. Logistic regression, a widely used linear classifier, provides a benchmark for performance in simpler, less complex datasets. SVC is considered for performing well with high-dimensional data and small sample sizes.

In addition to these traditional machine learning algorithms, we also train deep learning models, specifically neural networks, to explore their ability to capture non-linear patterns in the data. These neural networks are regularized using kernel methods to prevent overfitting, dropout techniques are employed to randomly drop units during training, improving generalization. The ReLU activation function is used for introducing non-linearity, which helps the model learn complex representations, and batch normalization is applied to accelerate training by normalizing the inputs to each layer. This combination of regularization and activation functions allows the model to learn from the data effectively while preventing overfitting and ensuring convergence.

Given the inherent class imbalance in our dataset, where landslides are rare events, we employ several techniques to address this issue and improve the model's ability to predict the minority classes correctly. We first utilize the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic instances for the minority class by interpolating between existing minority class samples. This method helps balance the distribution of the classes without losing any information from the majority class. We also experiment with SMOTEEN, a hybrid approach that combines both oversampling of the minority class and undersampling of the majority class. This hybrid technique aims to provide a more balanced dataset while avoiding the loss of potentially important majority class data. Finally, we incorporate class weights into the training process, where the each class's weight is inversely proportional to its frequency in the training data. This method adjusts the model's learning process, ensuring that it places greater emphasis on the minority classes during training.

To further improve model performance and computational efficiency, we explore dimensionality reduction techniques due to the diverse nature of the data sources and the high number of features involved. Principal Component Analysis (PCA) is used to reduce the dimensionality of the data by projecting it onto a lower-dimensional space that captures the most variance in the dataset. Similarly, Singular Value Decomposition (SVD) is employed to decompose the data matrix and reduce its rank, which can help capture the essential information while discarding noise. We train models using both PCA- and SVD-reduced data, and the original feature set to compare the impact of dimensionality reduction on model performance. These techniques help mitigate overfitting and computation time issues, especially in high-dimensional datasets.

4.1 Evaluation Metrics

Given the imbalanced nature of our dataset, where most samples are labeled as 'no landslide' (class 0) and the landslides are much less frequent, we assess the performance of our models using the macro F1 score. The macro F1 score is a valuable metric in imbalanced classification problems, as it computes the F1 score for each class independently and then averages them without considering the class distribution. This allows us to equally evaluate the performance across all classes (no landslide, low probability of landslide, and high probability of landslide), ensuring that the model's performance on the minority classes is adequately reflected.

In addition to the macro F1 score, we also report the false alarm rate and hit rate. The false alarm rate (FAR) is defined as the proportion of non-landslide instances (class 0) that are incorrectly predicted as landslides (class 1 or class 2). It can be calculated using the following formula:

$$\text{False Alarm Rate} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}} \quad (2)$$

The hit rate (HR), on the other hand, quantifies the proportion of actual landslides (classes 1 and 2) that the model correctly identifies. It is calculated using the formula:

$$\text{Hit Rate} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (3)$$

Both metrics provide valuable insight into the model's ability to detect landslides while minimizing the number of incorrect predictions.

5 Results

We conducted extensive experiments to assess how dataset combinations, imbalance handling, dimensionality reduction, and model choice affect landslide prediction. We tested each setup with and without oversampling and dimension reduction and reported the best model and configuration for each dataset combination. As shown in Table 1, integrating all four datasets (ASOS, ERA5, NLCD,

Table 1. Landslide detection in six states (2020) using 2018–2019 data. Performance comparison with oversampling, dimensionality reduction, and ML models.

Data	Oversampling Technique	Dimension Reduction	Best Model	Macro F1 Score
ASOS	SMOTE	-	Random Forest	0.63
ERA5	-	-	XGBoost	0.62
ASOS+NLCD	SMOTE	-	Random Forest	0.67
ERA5+NLCD	-	SVD	XGBoost	0.64
ASOS+SOLUS	-	-	XGBoost	0.66
ERA5+SOLUS	SMOTE	PCA	XGBoost	0.64
ASOS+ERA5	-	-	XGBoost	0.64
ASOS+ERA5+NLCD	-	PCA	XGBoost	0.66
ASOS+NLCD+SOLUS	SMOTE	PCA	XGBoost	0.68
ERA5+NLCD+SOLUS	SMOTE	-	Neural Net	0.61
ASOS+ERA5+NLCD +SOLUS	SMOTE	PCA	XGBoost	**0.70**

SOLUS) achieved the highest Macro F1 Score of 0.70, highlighting the value of multimodal environmental data. Our findings indicate that as the number of datasets increases, it leads to a higher number of input features. In these cases, SMOTE and PCA become particularly useful in improving classification performance. XGBoost without PCA achieved an F1 macro of 0.68, whereas without oversampling, it reached 0.64. Furthermore, our results consistently show that XGBoost outperforms other models, making it the most effective choice for landslide prediction across different dataset configurations. The highest-performing model (XGBoost trained on the full dataset with SMOTE and PCA) achieves a Macro F1 Score of 0.70. XGBoost approach is able to capture complex interactions between environmental variables better than simpler models. Interestingly, neural networks underperformed in our experiments, likely due to high-class imbalance. Even though we used weighted cross entropy and trained on synthetically oversampled data, our experiments showed that neural networks could not generalize well. Our experiments demonstrate that combining multiple data sources, leveraging oversampling techniques, and applying dimensionality reduc-

Table 2. Individual class performance. Hit rate measures correctly detected positives, while false alarm rate measures incorrectly detected negatives as positives.

Metric	No landslide	Low probability	High probability
Hit Rate	96.5%	52.2%	76.5%
False Alarm Rate	2.7%	54.7%	43.5%

tion contribute to enhanced predictive accuracy. These findings underscore the importance of multimodal data integration and preprocessing to improve landslide prediction models.

To further evaluate the best model's performance (XGBoost with SMOTE and PCA trained on 4 datasets combined), we conduct a class-wise analysis of hit and false alarm rates, as shown in Table 2. For the "No Landslide" class, the model achieves a high hit rate (96.5%) and a low false alarm rate (2.7%), indicating strong performance in identifying non-landslide events. This suggests that the model is well-calibrated for detecting stable conditions but may be biased toward the majority class due to dataset imbalance. The hit rate for the "Low Probability" class is 52.2%, meaning nearly half of the low-probability landslides are missed. The false alarm rate (54.7%) is high, indicating that many events are incorrectly classified as low probability when they belong to other classes. For the "High Probability" class, the model performs better, with a hit rate of 76.5%, meaning it detects most of the high-risk landslide cases. However, the false alarm rate of 43.5% suggests that many non-landslide events are misclassified as high probability, leading to potential over-warning situations. From a disaster management perspective, these results are acceptable if the primary goal is to predict high-risk landslides, as the model successfully captures most major events. Additionally, the low detection of low-probability landslides (HR = 52.2%) means that smaller-scale events are often overlooked, which could be a concern in regions where even minor landslides pose significant risks.

Next, we investigate which features contribute the most to the landslide detection, and the results are shown in Table 4. For this experiment, we analyze the best-performing model, XGBoost with SMOTE and PCA. After training the model, we analyze the principal components that had the highest influence on predictions and backtrack their contributions to the original attributes. The results reveal that important features originate from all four datasets, demonstrating that an integrated approach leveraging multiple data modalities enhances predictive performance. Notably, land cover attributes from NLCD and soil composition features from SOLUS appear frequently, highlighting the strong influence of terrain and soil properties on landslide susceptibility. Additionally, meteorological attributes from ERA5 and ASOS, such as horizontal wind speed, lake bottom temperature, and cloud layer height, emphasize the role of atmospheric conditions in triggering landslides. These findings reinforce the importance of combining static environmental features and dynamic weather patterns to improve landslide risk assessment, ultimately supporting the development of more accurate and reliable predictive models (Table 3).

To assess individual feature contributions without dimensionality reduction, we train XGBoost on the original dataset with class weights. The results, shown in Table 4, reveal that while some features remain consistent with the PCA-transformed model, such as evergreen forest area (NLCD) and coarse sand percentage (SOLUS), others differ. Notably, features like precipitation amount (ASOS), soil moisture content (ERA5), and soil chemistry such as pH and gypsum content (SOLUS) emerges, indicating that training on raw data high-

Table 3. Top features ranked by importance from the model trained on the PCA-reduced dataset, listed in decreasing order.

Feature	Source
Area of Evergreen Forest	NLCD
Lake bottom temperature	ERA5
Fine sand% (5cm depth)	SOLUS
Horizontal speed of air (mean)	ERA5
Horizontal speed of air (max)	ERA5
Very Coarse sand% (15cm depth)	SOLUS
Developed, Low Intensity	NLCD
Areas dominated by shrubs	NLCD
Area of Barren Land (Rock/Sand/Clay)	NLCD
Thickness of the uppermost layer of inland water bodies	ERA5
Height of the lowest cloud layer	ASOS
Coarse sand% (5cm depth)	SOLUS
Area of Open Water	NLCD
Area of Pasture/Hay	NLCD

lights additional factors that PCA de-emphasized. Despite variations in individual features, key environmental patterns persist. Water availability in soil remains critical, as reflected in soil moisture, precipitation, and soil water volumecfeatures, while soil composition plays a significant role. Additionally, soil temperature variations at different depths (ERA5) emerge as important factors, further emphasizing the role of subsurface conditions in landslide susceptibility. Crucially, significant features continue to originate from all four datasets, reinforcing the importance of an integrated approach. While PCA enhances feature abstraction, training on original data with class weights preserves explicit environmental variables, ensuring interpretability in landslide risk assessment. These results highlight how different modeling strategies influence feature selection while maintaining consistency in broader predictive patterns.

In the next experiment, we evaluate landslide detection performance across six states using the macro F1 score, comparing three model setups: state-specific models trained individually for each state, an all-states model trained on data from all six states, and a model trained on all states except Minnesota (MN). Minnesota is the only state with three classes, while all other states have two classes. The results in Table 5 show that state-specific models generally perform well, as they are tailored to the unique characteristics of each region. However, the all-states model improves performance in MN (0.68 vs. 0.66 in the state specific model) by providing additional training examples for the underrepresented class. In contrast, states with only two classes experience a decline in performance with the all-states model, as the introduction of an additional class

Table 4. Top features ranked by importance from the model trained on the original dataset, listed in decreasing order.

Feature	Source	PCA Top Feature?
Area of Evergreen Forest	NLCD	Yes
1 h precipitation amount (mean)	ASOS	No
Volume of water in soil layer 1 (0–7 cm)	ERA5	No
Silt content% (30 cm depth)	SOLUS	No
Oven-dry bulk density (60 cm depth)	SOLUS	No
Amount of rain intercepted by foliage and water from dew (mean)	ERA5	No
Sodium Adsorption Ratio (5 cm depth)	SOLUS	No
Air temperature	ASOS	No
Temperature of the surface of the Earth (std)	ERA5	No
Soil temperature at 28–100 cm (std)	ERA5	No
Soil pH(60 cm depth)	SOLUS	No
Gypsum content (100 cm depth)	SOLUS	No
Coarse sand% (5 cm depth)	SOLUS	Yes
Soil pH(0 cm depth)	SOLUS	No
Coarse sand% (15 cm depth)	SOLUS	Yes

creates confusion. Notably, excluding MN from training enhances performance in certain two-class states, such as North Dakota, where the macro F1 increases from 0.76 to 0.81. These results indicate that class distribution differences across states impacts model performance and a single model trained on all states may not be optimal for every region.

Next, we compare the performance of seasonal models with that of the overall model. To achieve this, we categorize the data into four seasons: Summer (June–

Table 5. Performance comparison between state-specific models and an overall model trained on all states.

State	Classes	State Model	All-States Model	Model Excluding MN
Minnesota (MN)	3	0.66	**0.68**	-NA-
Iowa (IA)	2	0.72	0.70	**0.73**
North Dakota (ND)	2	0.76	0.68	**0.81**
South Dakota (SD)	2	0.63	**0.68**	0.65
Michigan (MI)	2	0.67	0.67	**0.68**
Wisconsin (WI)	2	**0.74**	0.73	0.73

Table 6. Performance comparison between season-specific all-states models and an overall all-states model.

Season	Time period	Seasonal model	Overall model
Summer	Jun-Aug	0.56	**0.72**
Fall	Sep-Nov	0.47	**0.49**
Winter	Dec-Feb	0.50	0.50
Spring	Mar-May	0.64	**0.65**

August), Winter (December–February), Fall (September–November), and Spring (March–May). The results of this comparison are presented in Table 6. Our analysis indicates that all seasonal models are affected by temporal splitting. Except for Winter, the overall model consistently outperforms the seasonal models. In the Winter months, performance remains unchanged due to the limited number of landslide occurrences in the testing year. Only two landslides were recorded in these months, and neither model successfully identified them.

6 Conclusion

In this study, we develop a machine learning framework for landslide prediction by integrating multimodal environmental datasets, combining static and temporal attributes from ASOS, NLCD, SOLUS, and ERA5-Land. By framing landslide detection as a multiclass classification problem, we evaluate various machine learning models while addressing the challenges of data imbalance through oversampling techniques such as SMOTE and class-weighted training. Our results demonstrate that integrating all four datasets enhances predictive performance, with XGBoost trained on SMOTE-balanced, PCA-reduced data achieving the highest macro F1-score.

Our analysis highlights the importance of both meteorological and land-based attributes, with features from all datasets contributing significantly to model predictions. But despite achieving strong overall performance, our class-wise evaluation reveals limitations in detecting low-probability landslides, emphasizing the trade-off between reducing false alarms and capturing rare events. We provide insights into how environmental conditions impact landslide susceptibility by comparing models across states and seasons. This work underscores the value of integrating diverse data sources and applying advanced machine learning techniques to predict natural disasters more effectively. Future work could explore further improvements in the model in generalization, including using additional remote sensing data and dynamic aggregation of features.

Acknowledgments. This research was sponsored by the U.S. Army Engineer Research and Development Center (ERDC) and was accomplished under Cooperative Agreement Number W9132V-23-2-0002. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing

the official policies, either expressed or implied, of the U.S. Army Engineer Research and Development Center (ERDC) or the U.S. Government.

References

1. Most Forested States 2024 worldpopulationreview.com. https://worldpopulationreview.com/state-rankings/most-forested-states
2. Aljurbua, R., Alshehri, J., Gupta, S., Alharbi, A., Obradovic, Z.: Early prediction of power outage duration through hierarchical spatiotemporal multiplex networks. In: Complex Networks and Their Applications XIII: Proceedings of The Thirteenth International Conference on Complex Networks and Their Applications: COMPLEX NETWORKS 2024-Volume 3, p. 320. Springer, Cham (2024)
3. Amatya, M., Medwedeff, W., Clark, M.K.: Data-driven landslide nowcasting at the global scale
4. ASOS Network: Iowa environmental mesonet, Iowa State University (2022). https://mesonet.agron.iastate.edu/ASOS/. Accessed 31 May 2022
5. Baum, R.L., Godt, J.W.: Early warning of rainfall-induced shallow landslides and debris flows in the USA. Landslides **7**, 259–272 (2010)
6. Benz, S.A., Blum, P.: Global detection of rainfall-triggered landslide clusters. Nat. Hazard. **19**(7), 1433–1444 (2019)
7. Coluzzi, R., et al.: Rapid landslide detection from free optical satellite imagery using a robust change detection technique. Sci. Rep. **15**(1), 4697 (2025)
8. Ebrahim, M., et al.: Exploring time series models for landslide prediction: a literature review. Geoenviron. Disast. **11**(1), 1–20 (2024)
9. Fustos, I., Abarca-del Río, R., Mardones, M., González, L., Araya, L.: Rainfall-induced landslide identification using numerical modelling: a southern Chile case. J. S. Am. Earth Sci. **101**, 102587 (2020)
10. Fustos-Toribio, I., Manque-Roa, N., Vásquez Antipan, D., Hermosilla Sotomayor, M., Letelier Gonzalez, V.: Rainfall-induced landslide early warning system based on corrected mesoscale numerical models: an application for the Southern Andes. Nat. Hazard. **22**(6), 2169–2183 (2022)
11. Homer, C., et al.: Completion of the 2011 national land cover database for the conterminous united states-representing a decade of land cover change information. Photogrammetric Eng. Remote Sens. **81**(5), 345–354 (2015)
12. Iizuka, S., Xuan, Y., Kondo, Y.: Impacts of disaster mitigation/prevention urban structure models on future urban thermal environment. Sustain. Urban Areas **19**, 414–420 (2015)
13. Kang, J., Wan, B., Gao, Z., Zhou, S., Chen, H., Shen, H.: Research on machine learning forecasting and early warning model for rainfall-induced landslides in Yunnan Province. Sci. Rep. **14**(1), 14049 (2024)
14. Kuradusenge, M., Kumaran, S., Zennaro, M.: Rainfall-induced landslide prediction using machine learning models: the case of Ngororero district, Rwanda. Int. J. Environ. Res. Publ. Health **17**(11), 4147 (2020)
15. Muñoz-Sabater, J., et al.: Era5-land: a state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data **13**(9), 4349–4383 (2021)
16. Nauman, T., et al.: Data from: soil landscapes of the united states 100-meter (solus100) soil property maps project repository (2024)

17. Otudi, H., Gupta, S., Albarakati, N., Obradovic, Z.: Classifying severe weather events by utilizing social sensor data and social network analysis. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 64–71 (2023)
18. Otudi, H., Gupta, S., Obradovic, Z.: Leveraging diverse data sources for enhanced prediction of severe weather-related disruptions across different time horizons. In: International Conference on Engineering Applications of Neural Networks, pp. 220–234. Springer, Cham (2024)
19. Salam, R., Pla, F., Ahmed, B., Painho, M.: A convolutional neural network-based approach for automatically detecting rainfall-induced shallow landslides in a data-sparse context. Nat. Hazards Res. (2024)
20. Wang, H., Zhang, L., Yin, K., Luo, H., Li, J.: Landslide identification using machine learning. Geosci. Front. **12**(1), 351–364 (2021)