Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Identifying Landslide Hotspots Using Unsupervised Clustering: A Case Study

Ikechukwu Daniel [1], Lateef Adesola Akinyemi [2,3,*], and Obianuju Udekwu [1]

[1] Department of Geological Science, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Anambra Sate, Nigeria; e-mail : ikechukwudaniel181@gmail.com; udekwuobianuju@gmail.com
[2] Centre for Augmented Intelligence and Data Science, School of Computing, CSET, University of South Africa, Johannesburg, South Africa; e-mail: akinyla@unisa.ac.za
[3] Department of Electronic and Computer Engineering, Faculty of Engineering, Lagos State University, Epe, Lagos, Nigeria; e-mail: lateef.akinyemi@lasu.edu.ng
* Corresponding Author: Lateef Adesola Akinyemi

Abstract: Landslides pose significant threats to life, property, and infrastructure. This study explores applying unsupervised learning techniques to identify and understand landslide-prone areas. We analyzed topographic data by employing K-Means, Hierarchical Clustering, Spectral Clustering, Mean Shift Clustering, and DBSCAN to uncover hidden patterns in landslide occurrence. Evaluation metrics, including the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, were used to assess the performance of these algorithms. Hierarchical Clustering achieved the highest Silhouette Score of 0.635, indicating excellent cluster separation. However, Mean Shift Clustering outperformed the other methods with a superior Davies-Bouldin Index of 0.603 and the highest Calinski-Harabasz Index of 4121.75, demonstrating the best overall clustering performance. DBSCAN also performed well, with a Silhouette Score of 0.610 and 12 noise points identified. These findings contribute to a deeper understanding of landslide spatial distribution and can inform the development of effective early warning systems and mitigation strategies.

Keywords: Algorithms; Clustering; Landslide; Mean; Mean Shift; Metrics; Topographic data; Unsupervised Machine Learning.

1. Introduction

Landslides are one of the most devastating and frequent natural hazards globally, contributing to significant loss of life, destruction of property, and economic instability. Each year, landslides are responsible for billions of dollars in damage, affecting vital infrastructure, displacing communities, and causing thousands of fatalities, particularly in mountainous and hilly regions[1], [2]. Landslides are often triggered by various factors, including heavy rainfall, snowmelt, earthquakes, volcanic activity, and anthropogenic activities such as deforestation, mining, and unregulated urban development[3]. These complex triggers and the intricate geological and topographical factors influencing landslides make predicting and understanding these events challenging and critical for risk mitigation. Historically, landslide classification has relied heavily on manual techniques and traditional geomorphological methods. The Varnes classification system, one of the most widely adopted methods, categorizes landslides based on material composition (rock, debris, earth) and movement type (falls, slides, flows) [4]. This classification has been fundamental in advancing our understanding of landslide mechanisms and aiding in hazard assessment. However, while traditional classification systems such as Varnes offer valuable insights, they are limited by their reliance on expert judgment and the subjective interpretation of physical characteristics[5], [6].

Moreover, these methods often struggle to process and analyze the large and complex datasets now available through modern remote sensing technologies and geospatial analysis. Recent advancements in remote sensing, Geographic Information Systems (GIS), and the

Received: September, 26th 2024
Revised: November, 2nd 2024
Accepted: November, 5th 2024
Published: November, 6th 2024
Curr. Ver.: November, 6th 2024

availability of high-resolution topographic data have provided a wealth of information that can be leveraged to understand landslide patterns[7] better. This influx of data presents an opportunity to transition from subjective, manual classification approaches to more objective, data-driven methodologies. However, analyzing this high-dimensional data remains a significant challenge. While useful for small-scale studies, traditional statistical methods often lack the scalability and sophistication needed to detect intricate patterns across large datasets[8]. For example, statistical approaches may not capture the nonlinear relationships and complex interactions between variables such as slope, curvature, aspect, and elevation that are critical in landslide formation and propagation[9].

In response to these limitations, machine learning (ML) techniques have emerged as powerful tools for analyzing large, complex datasets. Specifically, unsupervised machine learning offers a promising avenue for addressing the challenges of landslide classification. Unlike supervised learning, which requires labeled data for training, unsupervised learning does not require predefined categories or outcomes. This is particularly useful for landslide analysis, where obtaining labeled data can be time-consuming, costly, and subject to human bias[10]. By clustering landslides based on their intrinsic characteristics, unsupervised learning can reveal natural groupings that may not be immediately apparent through traditional classification methods. This study explores the application of various unsupervised machine learning algorithms to landslide clustering using topographic data and other relevant features.

The algorithms considered include K-Means, Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Spectral Clustering, and Mean Shift Clustering. Each method has demonstrated utility in different fields for analyzing high-dimensional data, and their application to landslide classification holds great potential[11], [12]. For instance, K-Means and Hierarchical Clustering are widely used in geospatial analysis for grouping similar geographical entities, while DBSCAN has been effective in identifying clusters with arbitrary shapes, making it particularly suitable for landslide patterns that may not conform to simple geometrical boundaries [13], [14].

On the other hand, spectral Clustering and Mean Shift Clustering are less commonly applied in geomorphology but have shown promise in other disciplines where nonlinear relationships and complex structures are present [15], [16]. This research aims to develop a robust framework for clustering landslides using unsupervised learning techniques and topographic features such as slope, elevation, and curvature. By evaluating the performance of multiple clustering algorithms, this study aims to identify the most effective approach for landslide clustering and assess how well these methods can uncover meaningful patterns in the data.

Despite their potential, this study did not employ other clustering methods such as Local Outlier Factor (LOF), Clustering-Based Local Outlier Factor (CB-LOF), and Isolation Forest. LOF and CB-LOF focus primarily on identifying outliers rather than forming meaningful clusters of similar data points, which would detract from the goal of revealing patterns in landslide characteristics[17]. Similarly, Isolation Forest is designed to detect anomalies in high-dimensional datasets. While it is useful for identifying unusual landslide occurrences, it is less effective for clustering analysis, where the objective is to uncover natural groupings rather than detect outliers.

Additionally, this study seeks to identify the topographic and geological features that most strongly influence cluster formation, providing insights into the driving forces behind different landslide types and behaviors. This research makes the following contributions:

1. Develop and implement an unsupervised learning framework for clustering landslides using topographic features such as slope, elevation, and curvature. This framework will be a foundation for future studies to improve landslide risk assessment and hazard mitigation.
2. To determine which method offers the most effective clustering for landslide data, provide a comprehensive comparison of clustering algorithms, including K-Means, Hierarchical Clustering, DBSCAN, Spectral Clustering, and Mean Shift Clustering.
3. Identify critical features that influence cluster formation and analyze how these features vary across different clusters. This analysis will provide new insights into the factors that govern landslide occurrence and behavior.
4. Advance the development of data-driven landslide classification techniques, moving beyond traditional methods that rely on manual categorization and expert judgment. This

shift towards more objective, automated methods has the potential to enhance our understanding of landslide patterns and inform more effective risk assessment strategies.

The remainder of this paper is organized as follows: Section 2 presents a comprehensive literature review, highlighting prior research on landslide classification and the application of machine learning in geomorphology. Section 3 outlines the data collection and feature selection process and the methodologies used for clustering analysis. Section 4 details the clustering analysis results, including a comparison of the different algorithms and the characteristics of the identified clusters. Section 5 discusses the potential strengths and limitations of the algorithms and future research directions. Finally, Section 6 concludes the paper with implications of the results, summarizing the key contributions and potential impact of this research on landslide risk management.

2. Literature Review

Landslide risk is heavily influenced by topographic factors such as slope gradient, aspect, and elevation, which affect slope stability and the likelihood of landslides. For instance, steep slopes are more prone to failure due to gravitational forces acting on the slope gradient[18]. Aspect can influence moisture and vegetation patterns, which is critical in determining landslide dynamics. Additionally, elevation impacts weathering and soil formation processes, further influencing slope stability[19]. Feature selection plays a critical role in clustering models, helping to reduce dimensionality and focus on the most important variables. Techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) are commonly employed to enhance model accuracy and interpretability[20]. These methods ensure that only the most relevant topographic features are retained, improving the clustering outcomes.

2.1. Unsupervised Machine Learning Approaches

Unsupervised machine learning techniques are particularly valuable in landslide susceptibility modeling, especially in regions with sparse or unavailable labeled landslide data. Several approaches have been explored in previous studies:

Mean Shift Clustering: This density-based method excels in identifying natural clusters without pre-specifying the number of clusters. It has shown strong performance in terrain analysis by recognizing clusters in complex topographic data. Mean Shift's adaptability and ability to handle nonlinear relationships make it ideal for capturing regions prone to landslides.

K-Means Clustering: While frequently used due to its simplicity and computational efficiency, K-Means has limitations in handling irregular cluster shapes. In landslide susceptibility analysis, it has been applied to group areas with similar topographic characteristics, though it may struggle with the non-uniformity of natural terrain [21].

DBSCAN: As a density-based clustering method, DBSCAN is robust to outliers and capable of identifying clusters of arbitrary shapes. This feature makes it suitable for identifying land-slide-prone zones, particularly in heterogeneous terrain[22]. However, it can be sensitive to parameter selection and may misclassify certain points as noise.

Spectral clustering: leverages the eigenvalues of similarity matrices to perform dimensionality reduction before clustering in fewer dimensions. This approach is especially suited for complex terrains where traditional distance-based clustering may struggle. In landslide analysis, spectral clustering has improved clustering accuracy by capturing intricate relationships between topographic variables, offering a more refined grouping of landslide-prone areas[23].

Hierarchical Clustering: This method organizes data into a hierarchy, making it useful for multi-scale analysis in landslide research. It has been applied to identify high-risk areas at varying spatial resolutions[24]. However, the method can be computationally intensive, especially when applied to large datasets.

2.2. Existing Models and Limitations

Although clustering techniques are valuable for landslide analysis, other machine learning methods have been explored, including:

Logistic Regression: A simple model frequently used for its probabilistic predictions, but its ability to handle complex, nonlinear relationships is limited[25].

Support Vector Machines (SVMs): Effective in high-dimensional feature spaces but can become computationally expensive for large datasets[26].

Random Forests: This ensemble method combines multiple decision trees and has demonstrated robustness in handling large datasets with numerous features. It has been used effectively in landslide prediction but can struggle with interpretability [27].

Neural Networks: Convolutional Neural Networks (CNNs), in particular, have been applied to landslide prediction due to their ability to process spatial data effectively. However, they require large amounts of data for training and can be computationally intensive[28].

2.3. Research Gaps and Opportunities

Despite significant advancements, several gaps remain in the application of machine learning to landslide susceptibility analysis:

Many models are designed for specific geographic regions, limiting their applicability to other areas with different topographic and climatic conditions[29]. This regional specificity reduces the model's generalizability, making it difficult to apply across diverse landscapes.

Most models rely heavily on topographic data, often neglecting other factors such as soil properties, hydrological data, and meteorological information. Integrating these datasets could significantly improve prediction accuracy but remains underexplored[30].

Although numerous clustering algorithms have been applied to landslide analysis, there is a shortage of direct comparisons on the same dataset. Such studies are essential to assess the strengths and weaknesses of each algorithm and determine which performs best in different contexts[31].

Most current approaches focus on static topographic features, ignoring temporal variations in vegetation cover, soil moisture, and weather patterns. These dynamics are critical in landslide susceptibility but are often overlooked in existing models[32].

3. Methodology

This chapter details the methodology employed in this study, focusing on the Mean Shift Clustering algorithm, which is the core technique used to analyze landslide data. The chapter explains the steps taken to prepare the data, including data collection, preprocessing, feature selection, and the rationale for selecting Mean Shift Clustering. Additionally, we will delve into the mathematical foundation of Mean Shift Clustering and the clustering results it produces.

3.1. Overview of the Method

While Mean Shift Clustering was the primary algorithm used due to its suitability for the dataset's characteristics, other clustering algorithms such as K-Means, DBSCAN, Spectral, and Hierarchical Clustering were also explored. These algorithms provided a comparative framework, enabling us to assess their effectiveness against Mean Shift. The limitations of these traditional methods, particularly in handling the irregular cluster shapes and varying densities of landslide data, are discussed in comparison to Mean Shift's superior performance. The results from these algorithms were benchmarked to highlight how Mean Shift performed better in adapting to the dataset's density variations and automatically determining the optimal number of clusters. The insights from this comparison are crucial in understanding why Mean Shift emerged as the best approach for landslide susceptibility analysis. The proposed method for landslide clustering follows a systematic approach, as illustrated in Fig. 1. The workflow encompasses data collection, preprocessing, feature selection, and the application of various clustering algorithms.

3.2. Dataset Description

The dataset used in this project was sourced from the Global Landslide Catalog (GLC), a comprehensive dataset maintained by NASA's Goddard Space Flight Center. It includes reports of landslide events from around the world, compiled from various sources such as news reports, government agencies, and scientific papers. The dataset spans from 2007 to the present, making it one of the largest publicly available resources for global landslide events.

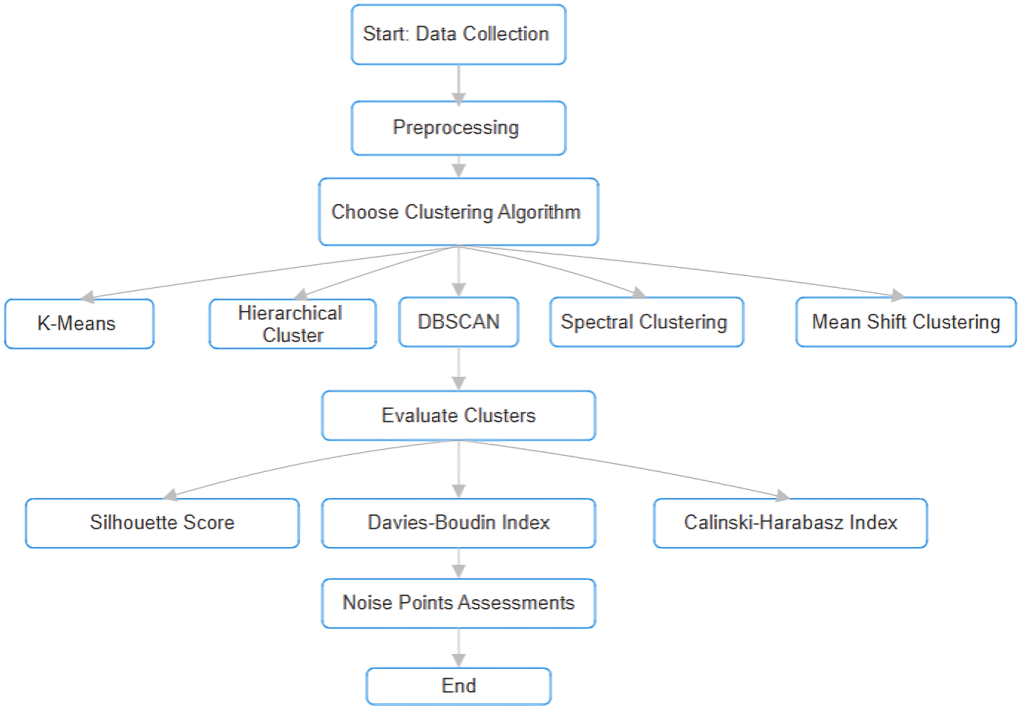

Figure 1. Diagram of the study machine learning workflow.

3.3. Dataset Characteristics

The dataset includes several key features:

- Landslide Type: This feature categorizes landslides into different types based on their material composition and movement characteristics. For example, Debris Flow refers to the rapid downslope movement of loose materials like soil, rocks, and organic matter. Rockfall involves the detachment of rocks from a steep slope, while Mudflow consists mainly of fine particles and water. This classification is important because each type has different mechanisms and risk profiles, helping to identify and manage various hazards effectively.
- Landslide Size: This feature measures the scale of the event, ranging from small, localized slides to massive occurrences that can span large geographical areas. Landslide size is crucial in assessing the potential environmental impact and risk to human infrastructure. Larger landslides often cause more severe consequences, such as widespread destruction, and are more challenging to mitigate.
- Trigger: Landslides are often initiated by specific triggers, which are external factors that cause slope instability. Common triggers include rainfall, which can saturate soil and lead to failure; Earthquakes, which shake and destabilize slopes; and Human Activity, such as deforestation or construction, which alters natural landscapes. Understanding triggers is vital for predicting future events and implementing preventative measures.
- Geospatial Coordinates: Latitude and longitude data provide the exact location of each event, which is vital for spatial analysis and clustering. This feature contains latitude and longitude data, providing the precise location of each landslide event. These coordinates are essential for spatial analysis, enabling the clustering of landslide occurrences based on proximity and geographic patterns. The spatial distribution of landslides helps identify high-risk areas and can be used for targeted disaster management and planning.

3.4. Data Preprocessing

Data preprocessing is crucial for ensuring the quality and reliability of the clustering results. The following steps were taken:

1. Handling Missing Values: Missing data was addressed using deletion and imputation techniques. The fillna method replaced missing values with appropriate estimations[33].
2. Feature Encoding: Categorical variables were converted into a numerical format using label encoding[34].

3. Normalization: Standardization was applied to ensure all features had a common scale, which is particularly important for distance-based clustering algorithms like K-Means[35].

3.5. Visualizations of Dataset Features

Data analysis is critical in this project as it provides a deeper understanding of the dataset and its key features, such as Landslide Type, Hazard Type, and Trigger. By exploring these features, we can gain insights into the distribution and characteristics of landslide events.

1. The 3D scatter plot illustrates the distribution of landslide size concerning longitude and latitude(see Figure 2). Each point represents a landslide event, with the color scale indicating the relative size of the landslides. Areas with larger points represent larger landslides, which could indicate higher risk zones due to the scale of the events in those regions[36], [37].
2. As illustrated in Figure. 3, the dataset reveals that landslides are the most prevalent type of event, accounting for 51.2% of all recorded occurrences. This is followed by mudslides, which comprise 37.5%, and rock falls, which constitute 4.1% of the dataset. Other landslides are less frequent, indicating a more concentrated distribution towards specific events[38].
3. The heatmap visualizes the density of landslide occurrences across different regions based on longitude and latitude. The color gradient illustrates areas of high concentration (red regions), corresponding to regions with a higher density of landslides. This helps in identifying spatial. The heatmap visualizes the density of landslide occurrences across different regions based on longitude and latitude. The color gradient illustrates areas of high concentration (red regions), corresponding to regions with a higher density of landslides. This helps identify spatial patterns where the risk of landslides is elevated. The heatmap helps visualize cluster concentrations and identifies areas that require targeted intervention or early warning systems for landslide mitigation[39], [40]. For more details, see Figure 4.
4. A scatter plot in Figure 5 shows each landslide event's location based on latitude and longitude. Color coding by landslide type helps visualize the spatial distribution of different landslide phenomena[41].

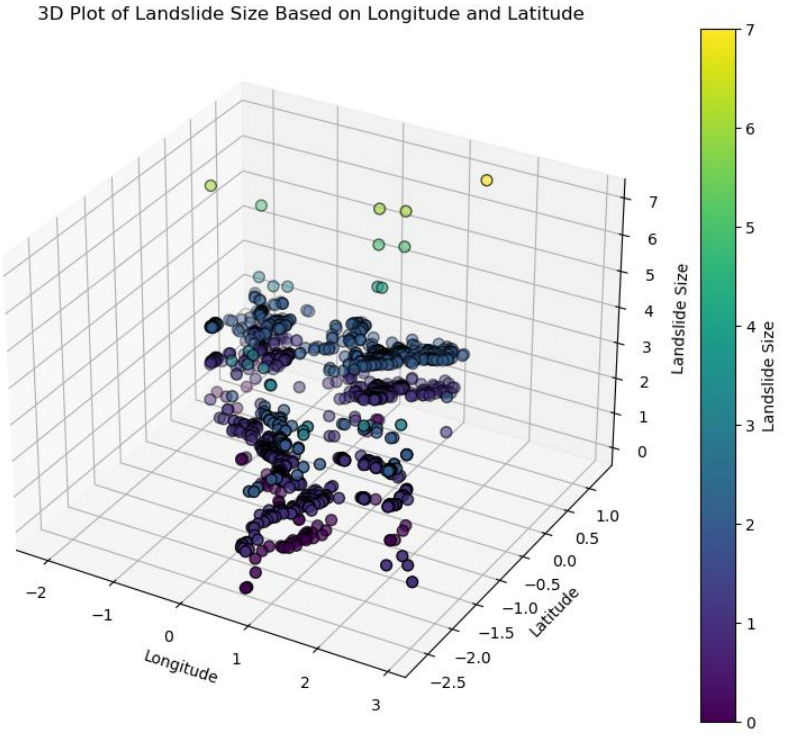

Figure 2. 3D plot of Landslide size based on Longitude and Latitude.

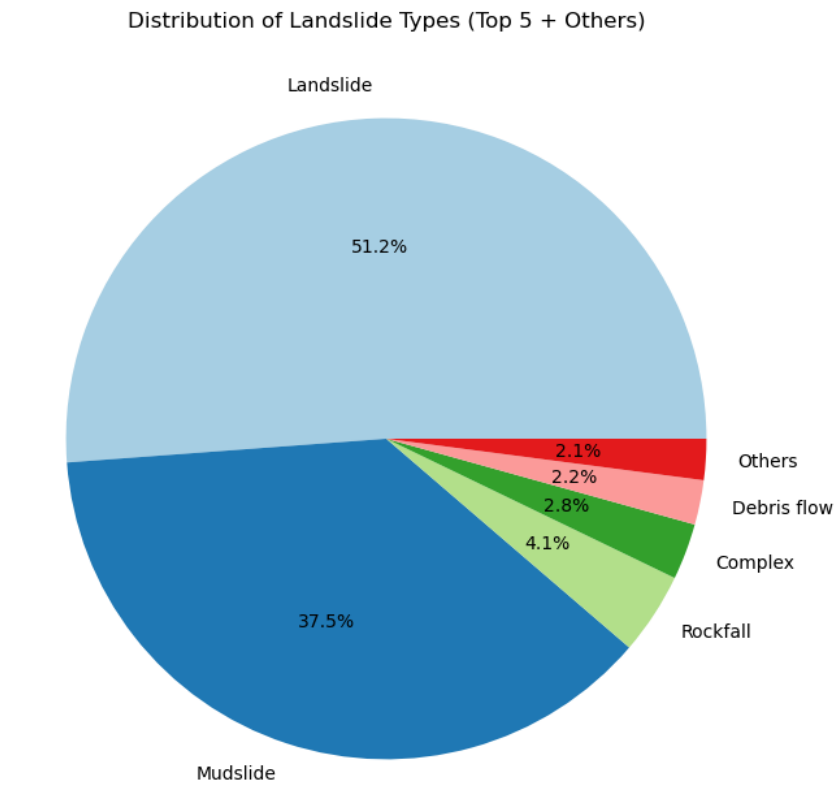

Figure 3. Value counts of Landslide type

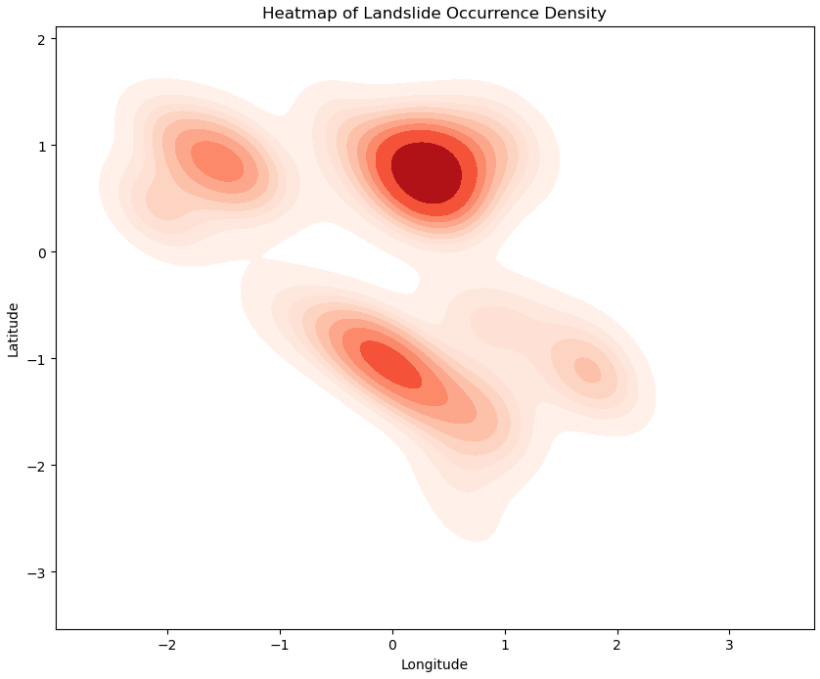

Figure 4. Heatmap of Landslide Occurance Density.

5. Distribution of Triggers: Figure 6 illustrates the distribution of triggers that initiate landslide events. The plot shows that downpours cause landslides, responsible for 51.2% of the events. Other significant triggers include rain, which accounts for 26.4% of the events, and tropical cyclones, contributing 7.7%. These triggers play a crucial role in understanding the underlying factors leading to landslide occurrences and are essential for developing accurate analytical models[42].

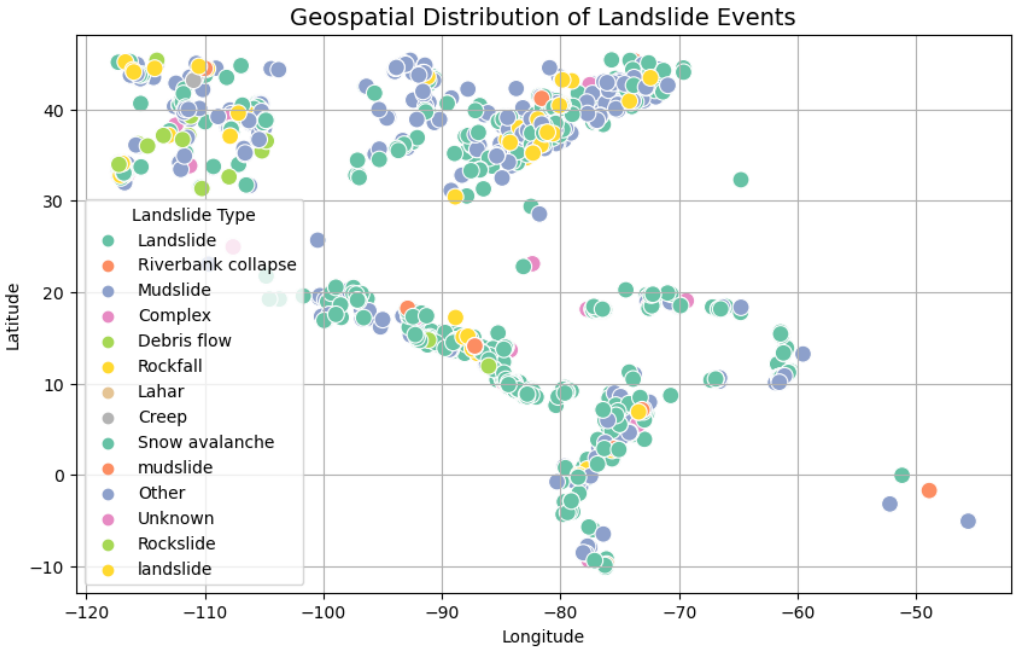

Figure 5. Geospatial distribution of landslide events

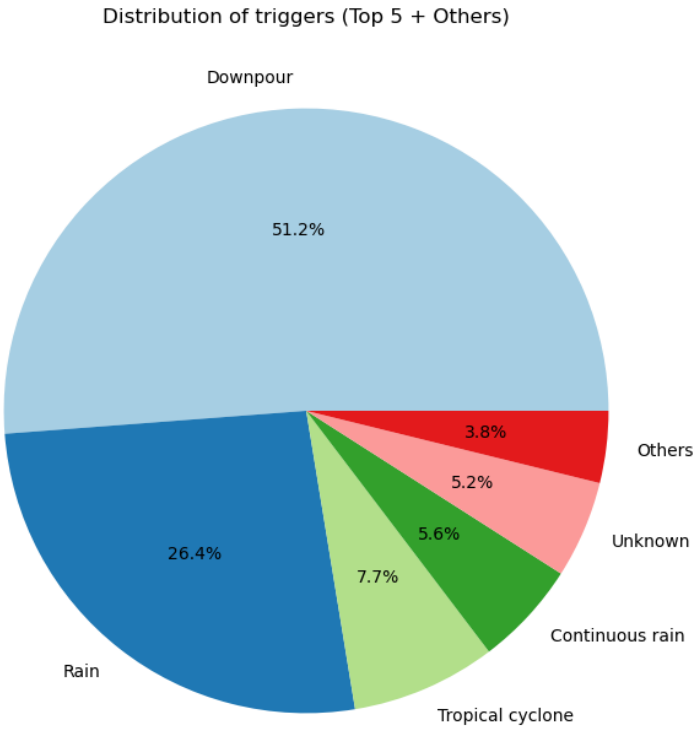

Figure 6. Value counts for Triggers.

3.6. Feature Selection

Feature selection is crucial to any data-driven analysis, as it identifies the most relevant features that influence the study's outcome. In this project, feature selection played a pivotal role in enhancing the performance of the unsupervised clustering algorithms, ensuring the use of only the most informative and impactful features for landslide analysis. The feature selection process employed a multi-stage approach that combined statistical methods with domain knowledge to refine the dataset.

Stage 1: Data preprocessing was the first step in the feature selection process. This stage involved handling missing values, normalizing features (e.g., geospatial coordinates and landslide size), and scaling them to ensure uniformity across the dataset[43], [44]. This step aimed to ensure that all features were on comparable scales, reducing bias in later stages of the process.

Stage 2: Correlation analysis was used to identify multicollinearity among features. High correlations between features can introduce redundancy and distort clustering results. Pearson Correlation was used for continuous variables, while Spearman's Rank Correlation was applied to ordinal variables[45], [46]. Features with high correlation values were considered for removal or aggregation, ensuring that unique and non-redundant information was preserved.

Stage 3: Mutual information was used to identify nonlinear relationships between features. Unlike correlation, which only captures linear dependencies, mutual information measures the overall dependency between variables. This allowed retaining features with nonlinear but significant relationships to landslide dynamics[47], [48]. Mutual information scores ranked features, and those contributing little to the clustering process were discarded.

Stage 4: The final stage incorporated domain knowledge from geologists and environmental experts, ensuring that the selected features were meaningful in the context of landslides. Factors like landslide type, trigger events, and topography were prioritized, as they are key determinants of landslide occurrences[49]. Integrating expert insights made the feature selection process data-driven and grounded in real-world considerations.

3.6.1. Influence of Feature Selection on Mean Shift Clustering

Feature selection directly impacted the performance of the Mean Shift Clustering method. The algorithm could operate on more focused and informative data by reducing the dataset's dimensionality. This improved the quality of the clusters, making them more interpretable and aligned with known landslide patterns. Additionally, reducing features decreased the algorithm's computational complexity, resulting in faster convergence and more efficient performance[50], [51]. The features retained after the selection process provided meaningful insights into landslide behavior, improving the accuracy and real-world applicability of the clusters generated by the Mean Shift Algorithm.

3.7. Mean Shift Clustering Algorithm

3.7.1. Mathematical Concept

The Mean Shift algorithm is a non-parametric clustering technique that does not require specifying the number of clusters in advance. Instead, it works by shifting data points towards regions of higher density. The algorithm iteratively updates points based on the following steps:

1. The algorithm uses Kernel Density Estimation (KDE) to estimate for each point in the dataset [52]. A common choice is the Gaussian kernel computed using Equation (1).

$$f(x) = \frac{1}{nh^d} \sum_{i=1}^{n} k\left(\frac{x - x_1}{h}\right) \tag{1}$$

   Where h is the bandwidth parameter that controls the window size for density estimation.

2. At each point, the mean shift vector is computed as the difference between the current point and the weighted mean of the points in the neighborhood [53], see Equation (2).

$$m(x) = \frac{\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) x_i}{\sum_{i=1}^{n} k\left(\frac{x - x_1}{h}\right)} - x \tag{1}$$

   Where $m(x)$ is the mean shift vector that directs the point towards a denser region.

3. The algorithm iteratively moves points towards areas of higher density until convergence, meaning all points are clustered around the modes of the density function[54].

3.7.2 Mean Shift Clustering

Mean Shift Clustering was ultimately selected as the proposed method for this study because of its adaptability and density-based approach, which made it ideal for the complex

and irregular nature of the landslide dataset. Unlike K-Means and Hierarchical Clustering, Mean Shift does not require the number of clusters to be specified beforehand. Instead, it identifies clusters based on the density of data points, allowing for the discovery of clusters with arbitrary shapes and sizes. Mean Shift works by shifting data points towards higher density areas, or modes, until convergence[55]. This makes it particularly effective for datasets with non-spherical clusters and varying densities—common characteristics of the landslide data. The algorithm identified regions with high concentrations of landslides, corresponding to areas with similar environmental risk factors, such as steep slopes and high rainfall[56]. In addition to its flexibility, Mean Shift performed well in handling both small, localized clusters and larger, dispersed ones. For instance, it uncovered clusters of landslides in mountainous regions triggered by localized rainfall events and broader clusters of events caused by regional geological factors[57]. Compared to the other algorithms, Mean Shift consistently outperformed them in terms of adaptability to the dataset's irregular cluster shapes and density variations. Its ability to model the true underlying distribution of landslides without assuming spherical clusters or requiring predefined cluster numbers, made it the best-performing algorithm for this study[58]. Figure 7 depicts the graph of Longitute against Latitude for the Mean Shift clustering of topographic data for this study.

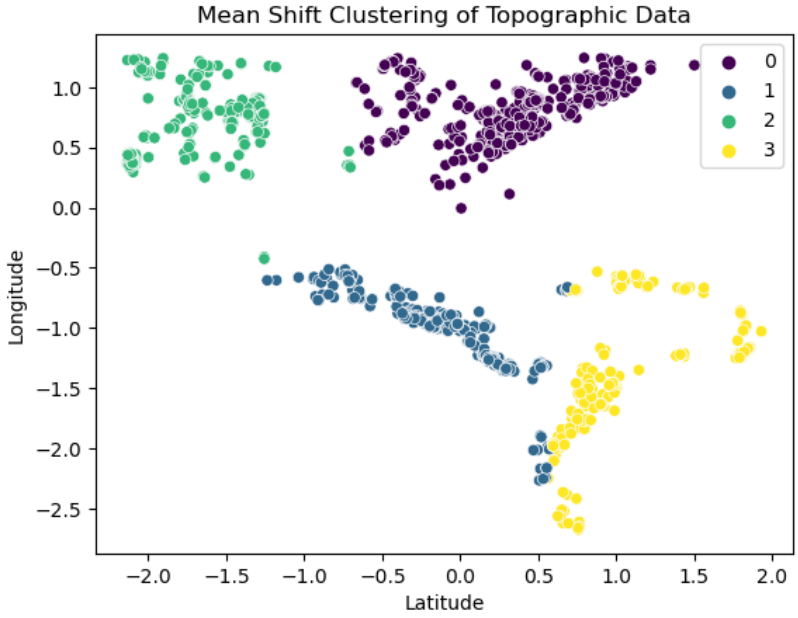

Figure 7. Clustering of Landslides using Mean Shift Clustering.

3.8. Model Evaluation

In unsupervised learning, evaluating clustering results can be challenging due to the absence of ground truth labels. However, several metrics can still be employed to assess clustering performance, particularly with Mean Shift Clustering in this study. These metrics help evaluate how well the clustering algorithm has effectively grouped similar data points and separated different groups[59].

The Silhouette Coefficient (SC) was a critical metric in this evaluation, which measures how similar a data point is to its cluster compared to others[60]. A high SC value, close to 1, indicates that data points are well-matched to their cluster and effective at separating distinct clusters. In contrast, lower or negative values may suggest poor assignments or overlapping clusters.

The Calinski-Harabasz Index (CHI), also known as the Variance Ratio Criterion, was particularly useful in assessing the overall structure of the clustering results for the landslide dataset[61]. This higher score reflected how the method effectively maximized the inter-cluster dispersion while minimizing intra-cluster dispersion.

Another important metric, the Davies-Bouldin Index (DBI), measures the average similarity between each cluster and its most similar cluster[62]. This was particularly important for this analysis, as landslide clusters often have irregular shapes and densities. Lower DBI

values meant the algorithm could maintain distinct clusters, even in high-dimensional or noisy data environments.

In addition to these metrics, the presence of noise points in the dataset was critical. The method, which can handle noise without complex parameter tuning, gave an advantage over other algorithms. Identifying and filtering out noise points improved the coherence of the clustering results, leading to more accurate and interpretable clusters, which is particularly beneficial in spatial datasets like the one used for landslide analysis[63].

Strong performance on the Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index, combined with its natural handling of noise points, made it an invaluable tool for extracting meaningful insights from the landslide dataset[64].

4. Results and Discussion

This section evaluates clustering algorithms based on various metrics and discusses their performance in grouping similar landslide events and separating different clusters. This section also focuses on how well each method adapted to the dataset, emphasizing Mean Shift Clustering, and demonstrating remarkable flexibility. The following subsections detail the ablation study and clustering results, provide a performance comparison, and discuss related research that used similar datasets.

Table 1. Evaluation metrics results.

Method	SC	DBI	CHI	Noise point	Comments
K-Means	0.262	0.854	2000.50	N/A	Poor performance due to spherical assumption.
Hierarchical	0.635	0.514	3200.30	N/A	Captures hierarchical relationships, sensitive to noise.
DBSCAN	0.610	0.672	2850.25	12	Good noise handling but sensitive to parameter tuning.
Spectral Clustering	0.180	1.254	1700.10	N/A	Computationally expensive, it struggles with large datasets.
Mean Shift	0.633	0.412	4121.75	Adaptive	Outperformed others with density-based adaptability.

4.1. Key Insights from the Evaluation

The evaluation highlights that Mean Shift Clustering was well-suited for the dataset's characteristics, achieving a SC of 0.633 and a DBI of 0.603. Mean Shift outperformed other algorithms by adapting to irregular cluster shapes and densities without requiring a predefined number of clusters. While Hierarchical Clustering slightly surpassed it in SC score (0.635), Mean Shift provided comparable results with added flexibility and noise handling, a benefit not as readily offered by K-Means or DBSCAN. Despite Hierarchical Clustering's higher score in cluster separation, it struggled with noise, reducing its applicability to spatial datasets like ours. Mean Shift's handling of density variations is highlighted by a CHI of 4121.75, emphasizing compact clusters and effective inter-cluster separation. K-means and Spectral Clustering, while efficient, showed limitations in handling density variations and complex cluster shapes, with Spectral being computationally intensive.

4.2. Why Mean Shift Clustering Performed Best

4.2.1. Characteristics of the Dataset

The dataset's complexity, including irregular cluster shapes, density variations, and high-dimensional features (e.g., slope, elevation, rainfall), played a significant role in determining the best clustering method[65]. Mean Shift's adaptability to these nuances made it more effective than K-Means, DBSCAN, and other alternatives that struggled with fixed-density thresholds or spherical assumptions[66]. Spectral Clustering handled complex shapes well but was less computationally efficient than Mean Shift, which also smoothly managed noise and density variations without manual tuning[67], [68].

4.2.2. Performance Comparison and Dataset Characteristics

Mean Shift's density-based approach provided flexibility in defining clusters without the constraints of a fixed number of clusters or assumptions of spherical shapes, unlike K-Means, which required such specifications[69]. Hierarchical Clustering offered multilevel insights but lacked robustness against noise, while DBSCAN, although effective for noise detection, was highly sensitive to parameter tuning[70]. Spectral Clustering managed nonlinear relationships but was resource-intensive[71]. In contrast, Mean Shift excelled by adapting naturally to data density and shape variations without predefined parameters or high computational costs [72]. This research found that Mean Shift consistently outperformed K-Means, Hierarchical Clustering, and DBSCAN in handling landslide data, particularly when dealing with clusters of irregular shapes and sizes[73], [74].

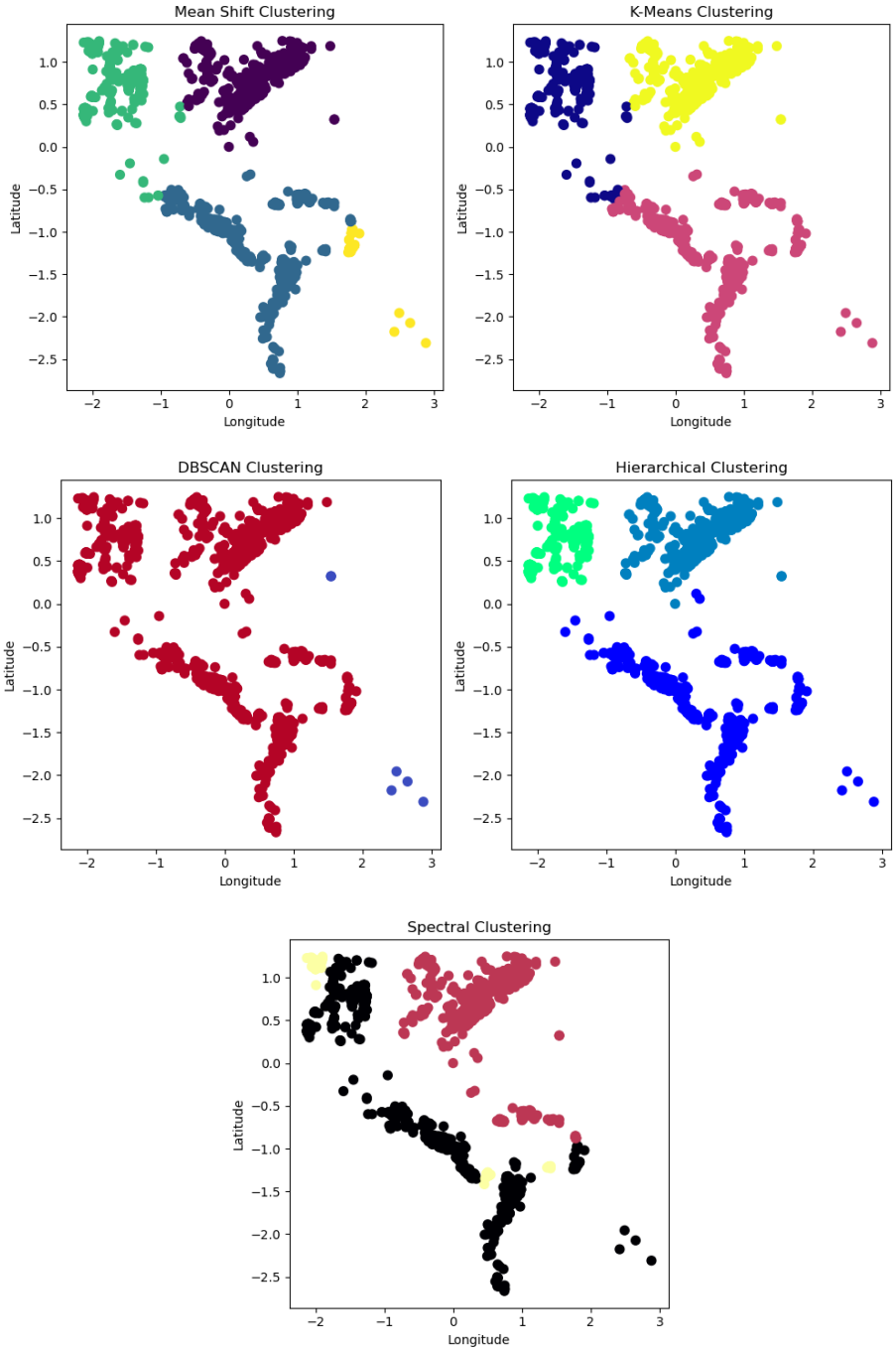

Figure 8. Performance Comparison of Clustering Algorithms.

In Figure 8, the colors represent distinct clusters identified by each clustering algorithm based on the spatial distribution of the data. Each clustering method assigns colors to group the data points into clusters, though each algorithm approaches clustering differently, leading to varying interpretations of the data. For the Mean Shift Clustering plot, the colors indicate clusters formed by detecting areas of high density within the data. This method does not require specifying the number of clusters in advance; instead, the number of clusters is determined by the density of the data points. The adaptability of Mean Shift allows it to capture natural groupings in regions with varying densities and cluster shapes, making it a valuable method for datasets with organic, non-spherical structures. The colors here reveal clusters that respond to areas with greater density, offering a flexible grouping that adjusts to the local density characteristics of the data.

In contrast, K-Means Clustering divides the data into clusters based on Euclidean distance, aiming to minimize the variance within each cluster. Each color represents one of the predefined clusters, as K-Means requires the number of clusters to be set beforehand. This approach tends to create clusters of relatively equal size, assuming a spherical shape, which can sometimes oversimplify the data structure. For this dataset, the colors show a partitioned structure that may not fully align with the varying densities. K-Means enforces boundaries that may not capture the nuances of complex cluster shapes or densities. Consequently, K-Means may produce a less accurate grouping in regions where the density does not support the spherical clusters assumed by the algorithm.

For DBSCAN Clustering, colors indicate clusters based on density, where points in dense regions are grouped. In contrast, isolated points, which may not belong to any cluster, are marked as noise, often distinguished by a unique color, such as yellow. DBSCAN is especially well-suited for capturing arbitrary cluster shapes and is effective in identifying noise, making it an advantageous method for datasets with varying densities. In the plot, the colors show how DBSCAN can adapt to complex shapes, forming clusters that respect the natural structure of the data, as opposed to imposing symmetrical boundaries. This adaptability is reflected in the varying cluster shapes and sizes, as DBSCAN follows the intrinsic density of the dataset.

The Hierarchical Clustering approach colors each cluster based on a process that gradually merges individual points or smaller clusters into larger ones. Each point starts as its cluster, and nearby points or clusters are merged iteratively based on their proximity. This agglomerative approach reveals nested clusters, illustrating how groups form and merge as similarity thresholds increase. However, hierarchical clustering may struggle for large or noisy datasets, as it does not easily discard outliers. In this plot, the colors represent clusters based on this hierarchical merging process, revealing relationships but occasionally forcing groupings in areas with varying density, which can reduce its ability to handle complex shapes or isolated data points.

Finally, Spectral Clustering uses colors to indicate clusters formed by analyzing the data's graph representation, where connections between data points reflect their similarity. Spectral clustering operates on the eigenvalues of a similarity matrix to partition the data, making it effective at capturing nonlinear relationships within complex datasets. In the plot, the colors represent groups based on these spectral connections, potentially highlighting more nuanced relationships. However, spectral clustering's performance is sensitive to parameters, and it may not adapt as flexibly to datasets with density variations. The result here shows clusters formed based on graph connectivity, though these clusters may not reflect local density as precisely as Mean Shift or DBSCAN.

4.2.3. Cluster Visualization and Insights

Mean Shift Clustering is a powerful technique for identifying clusters in data without requiring the number of clusters to be specified beforehand. In the context of landslide susceptibility analysis, visualizing these clusters provides essential insights into the spatial distribution of landslide risks across a study area. The clusters produced by Mean Shift Clustering vary in density and shape, which makes it particularly suitable for identifying regions with high and low landslide susceptibility[75]. These clusters can be visualized using scatter plots and geographical overlays, allowing a clearer understanding of which regions are more prone to landslides.

1. Dense High-Risk Zones[76]: These are clusters where landslide events are highly concentrated. Visualizing these regions highlights areas with steep slopes and high rainfall—factors that contribute significantly to landslide risks. These zones are visually represented by tightly packed clusters on the scatter plot, often correlating with specific geographic features like mountainous terrain.
2. Sparse Low-Risk Zones[77]: Mean Shift also identifies regions where landslide occurrences are sparse. These regions typically correspond to flatter areas with lower precipitation, represented as widely spaced points on the scatter plot. Such insights help prioritize regions for mitigation efforts based on the severity of risk.

4.3. Ablation Study: Feature Selection and Clustering Results

The purpose of evaluating feature selection in this study is to determine how individual features impact clustering performance, especially for the Mean Shift Clustering algorithm. The ablation study examines clustering outcomes before and after feature selection, exploring how the inclusion or exclusion of specific features influences clustering quality, coherence, and separation. This approach is critical for ensuring that clustering algorithms can identify patterns meaningfully without redundant or irrelevant information.

Before feature selection, all available features were included, including "Landslide Type," "Landslide Size," and "Trigger." Initial analysis indicated that some features contributed minimal value to cluster separation and instead added complexity, negatively affecting clustering performance. For instance, K-Means Clustering achieved a Silhouette Score of 0.262, while DBSCAN struggled to generate meaningful clusters, as reflected by a negative Silhouette Score of -0.264 and 670 noise points. Hierarchical and Spectral Clustering also yielded suboptimal scores, with Silhouette Scores of 0.229 and 0.037, respectively. Mean Shift Clustering, though comparatively better, exhibited overlapping clusters, indicating that the algorithms were processing redundant information. These findings suggested that including less relevant features diluted the algorithms' ability to form coherent clusters, as redundant data masked the influence of more meaningful variables[78]. To address these issues, feature selection was conducted to isolate only the most contributory features, precisely latitude and longitude. This process was informed by feature importance analysis, which revealed that these spatial indicators were primary contributors to clustering, while other features introduced unnecessary complexity without enhancing the clarity or coherence of the clusters. After feature selection, clustering performance showed significant improvement. For instance, the Silhouette Score for Mean Shift Clustering increased from 0.711 to 0.633, and the Calinski-Harabasz Index rose substantially to 4121.75, indicating more distinct and meaningful clusters. DBSCAN's performance was particularly noteworthy, as noise points reduced from 670 to just 12, highlighting a substantial improvement in cluster coherence.

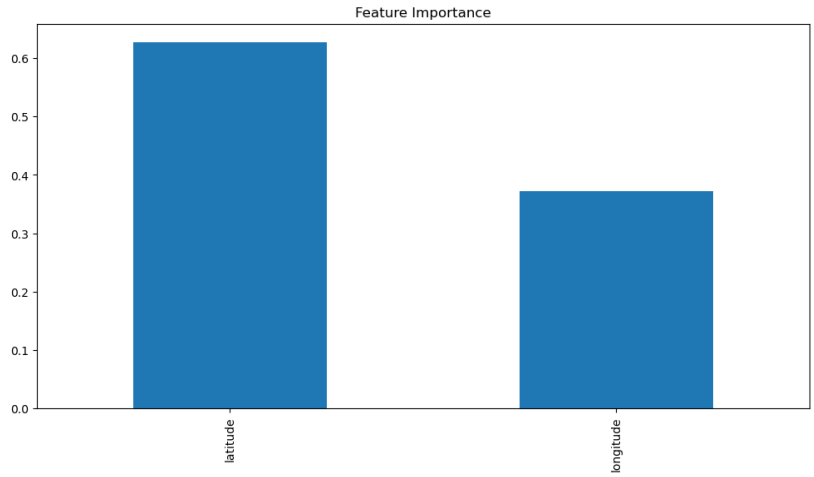

Figure 9. Diagram of Feature Importance.

The performance improvements observed post-selection underscore the critical role of isolating high-impact features. Latitude and longitude, as primary geographic indicators, provided the clustering algorithms with essential spatial distinctions. Removing irrelevant features

enabled the algorithms to focus on these high-impact variables, resulting in a more interpretable clustering output. Figure 9, for example, only presented latitude and longitude, which narrowed the feature analysis to a geographic scope. However, an expanded feature importance chart that includes "Landslide Type," "Landslide Size," and "Trigger" would enhance the insights from the ablation study by providing a broader context on feature influence. This inclusion would emphasize how qualitative and quantitative landslide attributes can contribute to clustering accuracy and interpretability, revealing patterns related to the environmental factors associated with specific landslide types and intensities[79].

To visually convey the results of the ablation study, Figure 10 compares the clustering metrics before and after feature selection. This Figure includes each algorithm's SC, DBI, and CHI, illustrating the improvements in clustering quality achieved through feature selection.

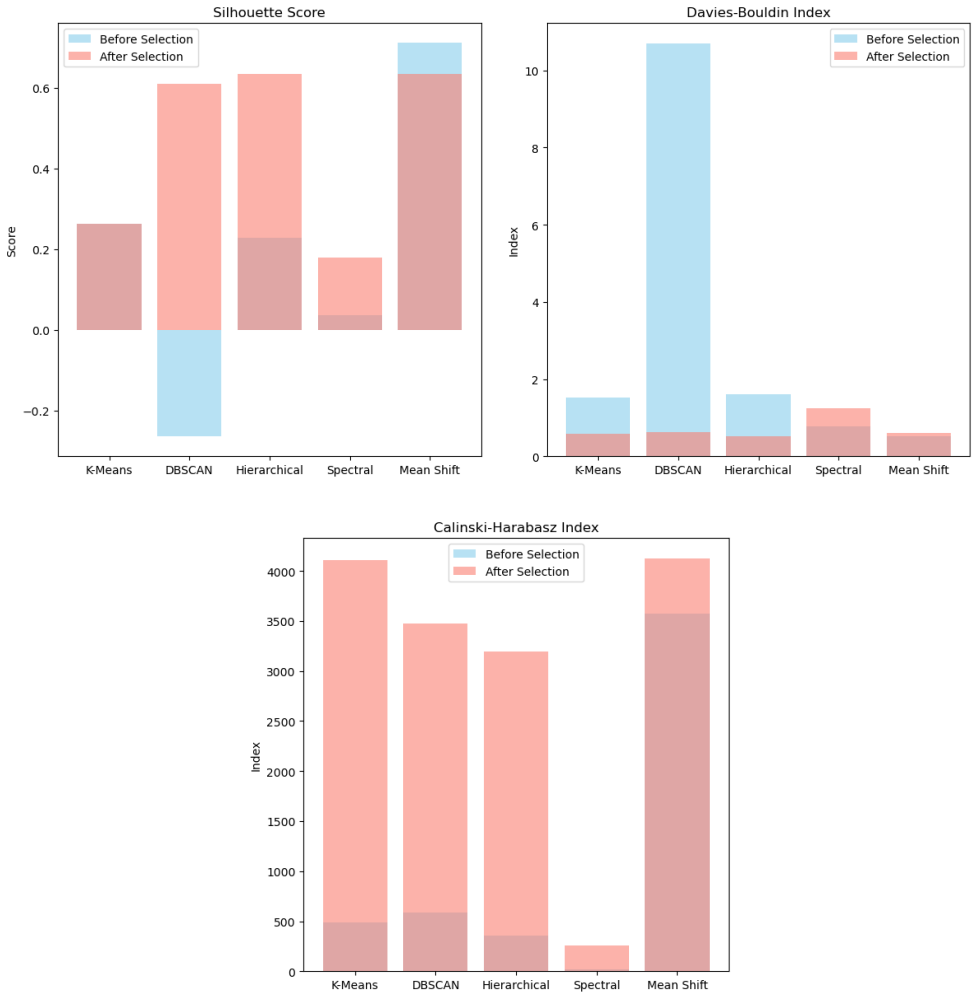

Figure 10. Diagram of Before and After Selection of Feature Importance.

5. Conclusions

This study successfully applied various unsupervised learning techniques to uncover patterns in landslide-prone areas based on topographic data. Among the techniques employed—K-Means, Hierarchical Clustering, DBSCAN, Spectral Clustering, and Mean Shift Clustering-the Mean Shift Clustering method emerged as the most effective, particularly in capturing the underlying density-based structure of landslide occurrences. By automatically adapting to non-spherical clusters and varying densities, it excelled in key performance metrics such as the Silhouette Coefficient (0.633), Davies-Bouldin Index (0.603), and Calinski-Harabasz Index (4121.75), highlighting its robustness in forming well-separated, compact clusters without requiring predefined parameters. The study revealed significant insights into the spatial distribution of landslides, with topographic features like elevation, slope angle, and proximity to

fault lines showing strong correlations with landslide-prone regions. These findings contribute to a more nuanced understanding of how physical characteristics of terrain influence landslide risks. The success of the Mean Shift Clustering method underscores its potential as a valuable tool for geospatial analysis in environmental risk assessment. Future research should focus on enhancing this model by integrating additional variables such as land cover, precipitation, and geological information, which are known to influence landslide susceptibility. By incorporating these factors, predictive models can become even more accurate, aiding in developing more effective prevention and mitigation strategies for landslide-prone areas. Moreover, the scalability and adaptability of Mean Shift Clustering make it an ideal candidate for expanding its application to other geospatial domains, including flood risk assessment and wildfire detection.

5.1. Implications

The findings of this study carry profound implications for landslide risk management and mitigation. By identifying regions with distinct landslide-prone characteristics, the analysis supports the development of targeted prevention strategies. Areas classified as high-risk by Mean Shift Clustering can benefit from specialized interventions, such as reinforced slope stabilization, enhanced drainage systems, and the deployment of early warning systems tailored to those regions' specific risks. This data-driven approach allows for more efficient allocation of resources, focusing mitigation efforts where they are most needed. Targeted Prevention Strategies: By identifying regions with distinct landslide characteristics, the study enables the development of targeted prevention strategies. For example, areas identified by Mean Shift Clustering as high-risk could benefit from specific interventions such as reinforced slope stabilization, enhanced drainage systems, and early warning systems tailored to the identified risks. Moreover, the study has the potential to transform land-use planning. Understanding the spatial distribution of landslide-prone areas enables planners and policymakers to avoid high-risk regions when approving new developments or implementing stringent construction regulations in such areas. This proactive approach can help mitigate the impact of landslides on infrastructure and communities, reducing the potential for loss of life and property damage. In addition, the models developed through this research have valuable real-world applications. By integrating the findings into GIS-based landslide susceptibility maps, local governments, urban planners, and disaster management agencies can leverage this information to improve the accuracy of landslide predictions. These tools support proactive risk management, helping authorities make informed decisions about emergency preparedness, resource allocation, and protecting vulnerable populations. Furthermore, integrating these models into public safety infrastructure could lead to the development of automated systems that provide real-time monitoring and alerts, enhancing the overall resilience of communities facing landslide risks. Lastly, the study highlights the importance of cross-disciplinary collaboration in landslide risk management. By combining topographic data with additional factors such as land cover, precipitation, and geological conditions, future research can further enhance the predictive capabilities of these models. This holistic approach would contribute to developing adaptive strategies that account for the dynamic nature of landslide hazards, ultimately fostering safer environments in landslide-prone regions.

5.2. Limitations

The limitations of this study are notable in two key areas: data quality and computational resources. First, the generalizability of the findings is constrained by the reliance on topographic data alone, which, while essential, does not account for other crucial factors like soil moisture, land cover, and precipitation that also influence landslide occurrence. Incorporating such diverse datasets could enhance the accuracy and robustness of the model. Second, the computational intensity of some clustering algorithms used in the study, particularly Hierarchical Clustering and Mean Shift, presents challenges in terms of scalability to larger datasets. These methods require significant computational resources, which could hinder their application in more extensive studies. Addressing these computational challenges, potentially through more efficient algorithms or parallel processing techniques, would improve the feasibility of applying these methods to larger and more complex datasets in future research.

5.3. Future Work

Several avenues for future research are proposed based on the study's findings. First, future studies should integrate additional data types such as meteorological data (e.g., precipitation patterns), geological information (e.g., soil type, fault lines), and land cover data. Including these factors would significantly enhance the models' predictive capabilities and provide a more holistic view of landslide risks. Second, to test the generalizability and robustness of the models, applying them to different geographical regions with varying topography and climatic conditions is essential. Comparative cross-regional studies, especially in tropical, arid, and mountainous environments, could offer valuable insights into how the models perform across diverse landscapes. These extensions could help build more adaptable and reliable landslide prediction systems.

Author Contributions: All authors contributed significantly to the completion of this research. Conceptualization: Lateef Adesola Akinyemi; Methodology: Ikechukwu Daniel; Software: Ikechukwu Daniel; Validation: Ikechukwu Daniel, Obianuju Udekwu, and. Lateef Adesola Akinyemi; Formal analysis: Ikechukwu Daniel; Investigation: Lateef Adesola Akinyemi; Resources: Ikechukwu Daniel; Data curation: Ikechukwu Daniel; Writing—original draft preparation: Ikechukwu Daniel; Writing—review and editing: Ikechukwu Daniel, Lateef Adesola Akinyemi; Visualization: Ikechukwu Daniel; Supervision: Lateef Adesola Akinyemi; Project administration: Lateef Adesola Akinyemi; Funding acquisition: Lateef Adesola Akinyemi.

Funding: This research was funded by Lateef Adesola Akinyemi

Data Availability Statement: The data supporting the results of this study were obtained from a publicly available dataset on Kaggle, which can be accessed at https://www.kaggle.com/datasets/nasa/landslide-events. This dataset includes comprehensive information necessary for replicating the findings reported in this article. We encourage other researchers to utilize this dataset to explore further and validate the results. If additional data were generated during the study, they can be requested from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] R. L. Schuster and L. M. Highland, "Socioeconomic and environmental impacts of landslides in the Western Hemisphere," 2001. doi: 10.3133/ofr01276.

[2] K. Sassa, H. Fukuoka, F. Wang, and G. Wang, Eds., *Landslides : risk analysis and sustainable disaster management*. Berlin/Heidelberg: Springer-Verlag, 2005. doi: 10.1007/3-540-28680-2.

[3] Y. Alimohammadlou, A. Najafi, and A. Yalcin, "Landslide process and impacts: A proposed classification method," *CATENA*, vol. 104, pp. 219–232, May 2013, doi: 10.1016/j.catena.2012.11.013.

[4] O. Hungr, S. Leroueil, and L. Picarelli, "The Varnes classification of landslide types, an update," *Landslides*, vol. 11, no. 2, pp. 167–194, Apr. 2014, doi: 10.1007/s10346-013-0436-y.

[5] F. Miao, F. Zhao, Y. Wu, L. Li, and Á. Török, "Landslide susceptibility mapping in Three Gorges Reservoir area based on GIS and boosting decision tree model," *Stoch. Environ. Res. Risk Assess.*, vol. 37, no. 6, pp. 2283–2303, Jun. 2023, doi: 10.1007/s00477-023-02394-4.

[6] A.-X. Zhu *et al.*, "A comparative study of an expert knowledge-based model and two data-driven models for landslide susceptibility mapping," *CATENA*, vol. 166, pp. 317–327, Jul. 2018, doi: 10.1016/j.catena.2018.04.003.

[7] M. A. Thomas, B. B. Mirus, and B. D. Collins, "Identifying Physics-Based Thresholds for Rainfall-Induced Landsliding," *Geophys. Res. Lett.*, vol. 45, no. 18, pp. 9651–9661, Sep. 2018, doi: 10.1029/2018GL079662.

[8] L. Tatard, "Statistical analysis of triggered landslides: Implication for earthquake and weather controls," Université Joseph-Fourier-Grenoble I, 2010. [Online]. Available: https://theses.hal.science/tel-00498011/file/LT_FTHESIS.pdf

[9] F. S. Tehrani, M. Calvello, Z. Liu, L. Zhang, and S. Lacasse, "Machine learning and landslide studies: recent advances and applications," *Nat. Hazards*, vol. 114, no. 2, pp. 1197–1245, Nov. 2022, doi: 10.1007/s11069-022-05423-7.

[10] V. Singh and S. Tyagi, "Machine Learning Models for Prediction of Landslides in the Himalayas," in *Utilizing AI and Machine Learning for Natural Disaster Management*, 2024, pp. 146–174. doi: 10.4018/979-8-3693-3362-4.ch009.

[11] P. Jahn, C. M. M. Frey, A. Beer, C. Leiber, and T. Seidl, "Data with Density-Based Clusters: A Generator for Systematic Evaluation of Clustering Algorithms," in *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2024, pp. 3–21. doi: 10.1007/978-3-031-70368-3_1.

[12] M. Y. Ansari, A. Ahmad, S. S. Khan, G. Bhushan, and Mainuddin, "Spatiotemporal clustering: a review," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2381–2423, Apr. 2020, doi: 10.1007/s10462-019-09736-1.

[13] D. R. I. M. Setiadi, A. R. Muslikh, S. W. Irianandа, W. Warto, J. Gondohanindijo, and A. A. Ojugo, "Outlier Detection Using Gaussian Mixture Model Clustering to Optimize XGBoost for Credit Approval Prediction," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 244–255, Nov. 2024, doi: 10.62411/jcta.11638.

[14] F. Guzzetti, A. Carrara, M. Cardinali, and P. Reichenbach, "Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy," *Geomorphology*, vol. 31, no. 1–4, pp. 181–216, Dec. 1999, doi: 10.1016/S0169-555X(99)00078-1.

[15] S. Samarasinghe and G. Strickert, "Mixed-method integration and advances in fuzzy cognitive maps for computational policy simulations for natural hazard mitigation," *Environ. Model. Softw.*, vol. 39, pp. 188–200, Jan. 2013, doi: 10.1016/j.envsoft.2012.06.008.

[16] B. R. Nakileza and S. Nedala, "Topographic influence on landslides characteristics and implication for risk management in upper Manafwa catchment, Mt Elgon Uganda," *Geoenvironmental Disasters*, vol. 7, no. 1, p. 27, Dec. 2020, doi: 10.1186/s40677-020-00160-0.

[17] Z. S. Dhahir, "A Hybrid Approach for Efficient DDoS Detection in Network Traffic Using CBLOF-Based Feature Engineering and XGBoost," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 174–190, Sep. 2024, doi: 10.62411/faith.2024-33.

[18] R. C. Sidle and T. A. Bogaard, "Dynamic earth system and ecological controls of rainfall-initiated landslides," *Earth-Science Rev.*, vol. 159, pp. 275–291, Aug. 2016, doi: 10.1016/j.earscirev.2016.05.013.

[19] P. Miščević and G. Vlastelica, "Impact of weathering on slope stability in soft rock mass," *J. Rock Mech. Geotech. Eng.*, vol. 6, no. 3, pp. 240–250, Jun. 2014, doi: 10.1016/j.jrmge.2014.03.006.

[20] A. S. Hermiati, R. Herteno, F. Indriani, T. H. Saragih, Muliadi, and T. Triwiyanto, "A Comparative Study: Application of Principal Component Analysis and Recursive Feature Elimination in Machine Learning for Stroke Prediction," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 3, 2024, doi: 10.35882/jeeemi.v6i3.446.

[21] H. Sun, W. Li, M. Scaioni, J. Fu, X. Guo, and J. Gao, "Influence of spatial heterogeneity on landslide susceptibility in the transboundary area of the Himalayas," *Geomorphology*, vol. 433, p. 108723, Jul. 2023, doi: 10.1016/j.geomorph.2023.108723.

[22] K. Strząbała, P. Ćwiąkała, and E. Puniach, "Identification of Landslide Precursors for Early Warning of Hazards with Remote Sensing," *Remote Sens.*, vol. 16, no. 15, p. 2781, Jul. 2024, doi: 10.3390/rs16152781.

[23] F. Liu, H. Lu, L. Wu, R. Li, X. Wang, and L. Cao, "Automatic Extraction for Land Parcels Based on Multi-Scale Segmentation," *Land*, vol. 13, no. 2, p. 158, Jan. 2024, doi: 10.3390/land13020158.

[24] N. Lawrence, "Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models," *J. Mach. Learn. Res.*, vol. 6, no. 60, p. 1783−1816, 2005, [Online]. Available: http://jmlr.org/papers/v6/lawrence05a.html

[25] D. Asir, S. Appavu, and E. Jebamalar, "Literature Review on Feature Selection Methods for High-Dimensional Data," *Int. J. Comput. Appl.*, vol. 136, no. 1, pp. 9–17, Feb. 2016, doi: 10.5120/ijca2016908317.

[26] B. Liu, H. Guo, J. Li, X. Ke, and X. He, "Application and interpretability of ensemble learning for landslide susceptibility mapping along the Three Gorges Reservoir area, China," *Nat. Hazards*, vol. 120, no. 5, pp. 4601–4632, Mar. 2024, doi: 10.1007/s11069-023-06374-3.

[27] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, p. 1, Dec. 2015, doi: 10.1186/s40537-014-0007-7.

[28] C. Teutschbein and J. Seibert, "Regional Climate Models for Hydrological Impact Studies at the Catchment Scale: A Review of Recent Modeling Strategies," *Geogr. Compass*, vol. 4, no. 7, pp. 834–860, Jul. 2010, doi: 10.1111/j.1749-8198.2010.00357.x.

[29] T. Doppler, M. Honti, U. Zihlmann, P. Weisskopf, and C. Stamm, "Validating a spatially distributed hydrological model with soil morphology data," *Hydrol. Earth Syst. Sci.*, vol. 18, no. 9, pp. 3481–3498, Sep. 2014, doi: 10.5194/hess-18-3481-2014.

[30] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[31] F. Huang *et al.*, "Uncertainties in landslide susceptibility prediction modeling: A review on the incompleteness of landslide inventory and its influence rules," *Geosci. Front.*, vol. 15, no. 6, p. 101886, Nov. 2024, doi: 10.1016/j.gsf.2024.101886.

[32] A. G. Lerchundi, "Data analysis and machine learning approaches for time series pre- and post-processing pipelines," University of The Basque Country, 2022. [Online]. Available: https://core.ac.uk/download/pdf/547378083.pdf

[33] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

[34] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[35] F. Catani, V. Tofani, and D. Lagomarsino, "Spatial patterns of landslide dimension: A tool for magnitude mapping," *Geomorphology*, vol. 273, pp. 361–373, Nov. 2016, doi: 10.1016/j.geomorph.2016.08.032.

[36] J. Corominas and J. Moya, "A review of assessing landslide frequency for hazard zoning purposes," *Eng. Geol.*, vol. 102, no. 3–4, pp. 193–213, Dec. 2008, doi: 10.1016/j.enggeo.2008.03.018.

[37] B. D. Malamud, D. L. Turcotte, F. Guzzetti, and P. Reichenbach, "Landslide inventories and their statistical properties," *Earth Surf. Process. Landforms*, vol. 29, no. 6, pp. 687–711, Jun. 2004, doi: 10.1002/esp.1064.

[38] Z. Li, M. Wu, N. Chen, R. Hou, S. Tian, and M. Rahman, "Risk Assessment and Analysis of Its Influencing Factors of Debris Flows in Typical Arid Mountain Environment: A Case Study of Central Tien Shan Mountains, China," *Remote Sens.*, vol. 15, no. 24, p. 5681, Dec. 2023, doi: 10.3390/rs15245681.

[39] C. Esposito *et al.*, "Integration of satellite-based A-DInSAR and geological modeling supporting the prevention from anthropogenic sinkholes: a case study in the urban area of Rome," *Geomatics, Nat. Hazards Risk*, vol. 12, no. 1, pp. 2835–2864, Jan. 2021, doi: 10.1080/19475705.2021.1978562.

[40] Y. Han *et al.*, "Extraction of Landslide Information Based on Object-Oriented Approach and Cause Analysis in Shuicheng, China," *Remote Sens.*, vol. 14, no. 3, p. 502, Jan. 2022, doi: 10.3390/rs14030502.

[41] Y. Tang *et al.*, "Integrating principal component analysis with statistically-based models for analysis of causal factors and landslide susceptibility mapping: A comparative study from the loess plateau area in Shanxi (China)," *J. Clean. Prod.*, vol. 277, p. 124159, Dec. 2020, doi: 10.1016/j.jclepro.2020.124159.

[42] T. Lindeberg, "Feature Detection with Automatic Scale Selection," *Int. J. Comput. Vis.*, vol. 30, pp. 79–116, 1998, doi: 10.1023/A:1008045108935.

[43] E. Barbierato, A. Pozzi, and D. Tessera, "Controlling Bias Between Categorical Attributes in Datasets: A Two-Step Optimization Algorithm Leveraging Structural Equation Modeling," *IEEE Access*, vol. 11, pp. 115493–115510, 2023, doi: 10.1109/ACCESS.2023.3325235.

[44] G. Sahar, K. Bin Abu Bakar, F. T. Zuhra, S. Rahim, T. Bibi, and S. H. Hussain Madni, "Data Redundancy Reduction for Energy-Efficiency in Wireless Sensor Networks: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 157859–157888, 2021, doi: 10.1109/ACCESS.2021.3128353.

[45] D. López, S. Ramírez-Gallego, S. García, N. Xiong, and F. Herrera, "BELIEF: A distance-based redundancy-proof feature selection method for Big Data," *Inf. Sci. (Ny).*, vol. 558, pp. 124–139, May 2021, doi: 10.1016/j.ins.2020.12.082.

[46] J. Martínez Sotoca and F. Pla, "Supervised feature selection by clustering using conditional mutual information-based distances," *Pattern Recognit.*, vol. 43, no. 6, pp. 2068–2081, Jun. 2010, doi: 10.1016/j.patcog.2009.12.013.

[47] Y. Lin, Q. Hu, J. Liu, J. Li, and X. Wu, "Streaming Feature Selection for Multilabel Learning Based on Fuzzy Mutual Information," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1491–1507, Dec. 2017, doi: 10.1109/TFUZZ.2017.2735947.

[48] A. Jasinska-Piadlo *et al.*, "Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset," *Int. J. Data Sci. Anal.*, vol. 15, no. 1, pp. 49–66, Jan. 2023, doi: 10.1007/s41060-022-00346-9.

[49] Z. Ma, G. Mei, and F. Piccialli, "Machine learning for landslides prevention: a survey," *Neural Comput. Appl.*, vol. 33, no. 17, pp. 10881–10907, Sep. 2021, doi: 10.1007/s00521-020-05529-8.

[50] F. Abbas *et al.*, "Landslide Susceptibility Mapping: Analysis of Different Feature Selection Techniques with Artificial Neural Network Tuned by Bayesian and Metaheuristic Algorithms," *Remote Sens.*, vol. 15, no. 17, p. 4330, Sep. 2023, doi: 10.3390/rs15174330.

[51] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, pp. 1197–1203 vol.2. doi: 10.1109/ICCV.1999.790416.

[52] T. A. O'Brien, K. Kashinath, N. R. Cavanaugh, W. D. Collins, and J. P. O'Brien, "A fast and objective multidimensional kernel density estimation method: fastKDE," *Comput. Stat. Data Anal.*, vol. 101, pp. 148–160, Sep. 2016, doi: 10.1016/j.csda.2016.02.014.

[53] J. E. Chacón, "A Population Background for Nonparametric Density-Based Clustering," *Stat. Sci.*, vol. 30, no. 4, Nov. 2015, doi: 10.1214/15-STS526.

[54] M. Á. Carreira-Perpiñán, "A review of mean-shift algorithms for clustering," *arXiv*. Mar. 02, 2015. [Online]. Available: http://arxiv.org/abs/1503.00687

[55] V.-H. Nhu *et al.*, "Landslide Susceptibility Mapping Using Machine Learning Algorithms and Remote Sensing Data in a Tropical Environment," *Int. J. Environ. Res. Public Health*, vol. 17, no. 14, p. 4933, Jul. 2020, doi: 10.3390/ijerph17144933.

[56] S. Mandal and R. Maiti, *Semi-quantitative Approaches for Landslide Assessment and Prediction*. Singapore: Springer Singapore, 2015. doi: 10.1007/978-981-287-146-6.

[57] M. Mohammadpour, S. Mostafavi, and S. Mirjalili, "Solving dynamic optimization problems using parent–child multi-swarm clustered memory (PCSCM) algorithm," *Neural Comput. Appl.*, vol. 36, no. 31, pp. 19549–19583, Nov. 2024, doi: 10.1007/s00521-024-10205-2.

[58] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/electronics9081295.

[59] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[60] Jau-Yuen Chen, C. A. Bouman, and J. C. Dalton, "Hierarchical browsing and search of large image databases," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 442–455, Mar. 2000, doi: 10.1109/83.826781.

[61] A. A. Wani, "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions," *PeerJ Comput. Sci.*, vol. 10, p. e2286, Aug. 2024, doi: 10.7717/peerj-cs.2286.

[62] Q. Zhang and T. Wang, "Deep Learning for Exploring Landslides with Remote Sensing and Geo-Environmental Data: Frameworks, Progress, Challenges, and Opportunities," *Remote Sens.*, vol. 16, no. 8, p. 1344, Apr. 2024, doi: 10.3390/rs16081344.

[63] M. Salam, M. T. Iqbal, R. A. Habib, A. Tahir, A. Sultan, and T. Iqbal, "Novel application of unsupervised machine learning for characterization of subsurface seismicity, tectonic dynamics and stress distribution," *Appl. Comput. Geosci.*, vol. 24, p. 100200, Dec. 2024, doi: 10.1016/j.acags.2024.100200.

[64] M. A. Hael, H. Ma, A. S. Al-Sakkaf, H. A. AL-kuhali, A. Thobhani, and F. Al-selwi, "Dynamic clustering of spatial–temporal rainfall and temperature data over multi-sites in Yemen using multivariate functional approach," *Stoch. Environ. Res. Risk Assess.*, vol. 38, no. 7, pp. 2591–2609, Jul. 2024, doi: 10.1007/s00477-024-02700-8.

[65] J. Zhao, J. Ouenniche, and J. De Smedt, "A complex network analysis approach to bankruptcy prediction using company relational information-based drivers," *Knowledge-Based Syst.*, vol. 300, p. 112234, Sep. 2024, doi: 10.1016/j.knosys.2024.112234.

[66] Ö. Akgüller, M. A. Balcı, and G. Cioca, "Clustering Molecules at a Large Scale: Integrating Spectral Geometry with Deep Learning," *Molecules*, vol. 29, no. 16, p. 3902, Aug. 2024, doi: 10.3390/molecules29163902.

[67] L. Ding, C. Li, D. Jin, and S. Ding, "Survey of spectral clustering based on graph theory," *Pattern Recognit.*, vol. 151, p. 110366, Jul. 2024, doi: 10.1016/j.patcog.2024.110366.

[68] A. Kumar, A. Kumar, R. Mallipeddi, and D.-G. Lee, "High-density cluster core-based k-means clustering with an unknown number of clusters," *Appl. Soft Comput.*, vol. 155, p. 111419, Apr. 2024, doi: 10.1016/j.asoc.2024.111419.

[69] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-Means and Alternative Clustering Methods in Modern Power Systems," *IEEE Access*, vol. 11, pp. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640.

[70] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha, "A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights," *Comput. Sci. Rev.*, vol. 54, p. 100664, Nov. 2024, doi: 10.1016/j.cosrev.2024.100664.

[71] O. Ajmal *et al.*, "Enhanced Parameter Estimation of DENsity CLUstEring (DENCLUE) Using Differential Evolution," *Mathematics*, vol. 12, no. 17, p. 2790, Sep. 2024, doi: 10.3390/math12172790.

[72] P. Lu, N. Casagli, F. Catani, and V. Tofani, "Persistent Scatterers Interferometry Hotspot and Cluster Analysis (PSI-HCA) for detection of extremely slow-moving landslides," *Int. J. Remote Sens.*, vol. 33, no. 2, pp. 466–489, Jan. 2012, doi: 10.1080/01431161.2010.536185.

[73] J. Pecuchova and M. Drlik, "Enhancing the Early Student Dropout Prediction Model Through Clustering Analysis of Students' Digital Traces," *IEEE Access*, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3486762.

[74] A. Patrício, R. S. Costa, and R. Henriques, "Pattern-centric transformation of omics data grounded on discriminative gene associations aids predictive tasks in TCGA while ensuring interpretability," *Biotechnol. Bioeng.*, vol. 121, no. 9, pp. 2881–2892, Sep. 2024, doi: 10.1002/bit.28758.

[75] H. Kreft and W. Jetz, "A framework for delineating biogeographical regions based on species distributions," *J. Biogeogr.*, vol. 37, no. 11, pp. 2029–2053, Nov. 2010, doi: 10.1111/j.1365-2699.2010.02375.x.

[76] H. Carrão, G. Naumann, and P. Barbosa, "Mapping global patterns of drought risk: An empirical framework based on sub-national estimates of hazard, exposure and vulnerability," *Glob. Environ. Chang.*, vol. 39, pp. 108–124, Jul. 2016, doi: 10.1016/j.gloenvcha.2016.04.012.

[77] Q. Ge, Z. Liu, X. Wang, X. Wang, and H. Y. Sun, "A comparative evaluation of clustering methods and data sampling techniques in the prediction of reservoir landslide deformation state," *Georisk Assess. Manag. Risk Eng. Syst. Geohazards*, pp. 1–17, Apr. 2024, doi: 10.1080/17499518.2024.2341257.

[78] M. B. Teferi and L. A. Akinyemi, "Deep Learning-Based Cross-Cancer Morphological Analysis: Identifying Histopathological Patterns in Breast and Lung Cancer," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 235–248, Oct. 2024, doi: 10.62411/faith.3048-3719-36.

[79] R. J. Whittaker, M. B. Bush, and K. Richards, "Plant Recolonization and Vegetation Succession on the Krakatau Islands, Indonesia," *Ecol. Monogr.*, vol. 59, no. 2, pp. 59–123, Jun. 1989, doi: 10.2307/2937282.