ISSN: 2994-6433

Research Article

Journal of Sensor Networks and Data Communications

Enhancing Landslide Prediction: A Comparative Study of Ensembled and Non-Ensembled Machine Learning Approaches with Dimensionality Reduction and Random Feature Selection to Showcase Entropy Management

Adeel Abbas[1], Farkhanda Abbas[2*], Fazila Abbas[3], Abdulwahed Fahad Alrefaei[4] and Mohammed Fahad Albeshr[5]

[1]The University Of Poonch, Rawalakot, Azad Kashmir 12350 Pakistan

[2]School of Computer Science, China University of Geosciences, Wuhan 430074, China

[3]Institute of Soil and Environmental Sciences, University of Agriculture Faisalabad, Faisalabad 38000, Pakistan

[4]Department of Zoology, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia

[5]Department of Zoology, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia

[*]Corresponding Author
Farkhanda Abbas, School of Computer Science, China University of Geosciences, Wuhan 430074, China.

Submitted: 2024, Oct 15; **Accepted:** 2024, Nov 27; **Published:** 2024, Nov 29

Citation: Abbas, A., Abbas, F., Abbas, F., Alrefaei, A. F., Albeshr, M. F. (2024). Enhancing Landslide Prediction: A Comparative Study of Ensembled and Non-Ensembled Machine Learning Approaches with Dimensionality Reduction and Random Feature Selection to Showcase Entropy Management. *J Sen Net Data Comm, 4*(3), 01-23.

Abstract

This research looks at how well different ensembled and non-ensembled machine learning algorithms perform both before and after dimensionality reduction and manual feature engineering using random feature selection. LightGBM, Extra Trees (EXT), XGBoost, Gradient Boosting Machine (GBM), Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree (DT) are among the algorithms that were assessed. With a computational time (CT) of 15.985 seconds prior to dimensionality reduction, LightGBM obtained an AUC/ROC score of 0.833, whereas Extra Trees (EXT), XGBoost, and GBM each obtained AUC/ROC scores of 0.832 with CTs of 15.892, 16.203, and 15.904 seconds, respectively. While Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree (DT) displayed decreasing AUC/ROC scores and varying CTs (RF: 0.784, 16.130s), NB: 0.740, 2.456s, KNN: 0.718, 1.897s, and DT: 0.689, 1.787s), CatBoost came in second with an AUC/ROC score of 0.816 and a CT of 17.121 seconds. The algorithms showed improved performance metrics following the reduction of dimensionality: LightGBM had the highest AUC/ROC score of 0.979 with a CT of 15.344 seconds, while CatBoost had a competitive AUC/ROC score of 0.977 with a CT of 15.235 seconds. Other methods also showed improvement. All computations were performed more efficiently as a result of the smaller feature space. AUC/ROC scores for LightGBM and XGBoost were 0.830 and 0.829, respectively, with CTs of 22.345 and 22.455 seconds, after manual feature engineering through random feature selection. CatBoost, on the other hand, had an AUC/ROC score of 0.814 with a CT of 24.587 seconds. These modifications revealed extra computational complexity brought about by feature engineering, which had an impact on calculation times as well as performance measures. The impact of preprocessing strategies on computing efficiency and model performance is highlighted in this work. By concentrating on pertinent features, dimensionality reduction dramatically improved AUC/ROC scores and shortened calculation times, whereas manual feature engineering offered more nuanced insights but frequently at the expense of more computational complexity. The trade-offs associated with maximizing the accuracy and efficiency of machine learning models are highlighted by these results.

Keywords: Machine Learning, Accuracy, Complexity, Entropy, Landslide Susceptibility Mapping, Dimensionality Reduction, Principal Component Analysis (PCA)

1. Introduction

The last few decades have seen incredible progress in the ability to gather and store data, which has led to an information explosion in many scientific fields. Experts in fields such engineering, biology, astronomy, remote sensing, economics, and consumer transactions now have to deal with daily challenges posed by constantly growing datasets and simulations [1]. These datasets bring new difficulties to the field of data analysis, in sharp contrast to smaller, more conventionally studied ones. Traditional statistical techniques have constraints, not only from the increase in the number of observations but also mainly from the growing number of factors associated with each observation. The number of variables measured for every observation is indicated by the data's scale. Highly multidimensional datasets present both special potential and complex mathematical problems. They have the potential to inspire new theoretical advancements in the area [2]. High-dimensional datasets frequently provide a problem in that not all of the variables that are captured are essential for understanding the underlying processes of interest. Even though there are computationally demanding methods that can create very accurate predictive models from this type of data, many applications still need that the original dataset be made less dimensional before beginning any modeling work [3]. The issue under examination can be expressed mathematically as follows: Given a p-dimensional random variable, $x = (x_1,\dots,x_p)^T$ the goal is to find a reduced-dimensional representations $s = (s_1,\dots,s_k)^T$ where $k \leq p$, that successfully captures the underlying information in the original data according to a particular criterion. Many times, these discrete elements within s are referred to as hidden or latent variables. Different disciplines utilize different nomenclature to refer to the p-dimensional multivariate vectors: "variable" is frequently utilized in statistics, but "feature" and "attribute" are prominent substitutes found in computer science and machine learning literature.Throughout this paper, we make the following assumptions: : We have a dataset with nobservations, each representing a realization of a p-dimensional random variable denoted as $x = (x_1,\dots,x_p)^T$, with a mean $E(x) =(\mu_1,\dots,\mu_p)^T$ and a covariance matrix $E\{(x\text{-}\mu)\ (x\text{-}\mu)^T\} = \sum p{\times}p$. We represent such an observation matrix as $X = \{x_{ij}{:}1 \leq i \leq p,\ 1 \leq j \leq n\}$. If μ_i and $\sigma_i = \sqrt{\sum(i,\imath)}$ denote the mean and standard deviation of the ith random variable, respectively,we frequently standardize the observations as follows:

$s_{ij} = ((x_{ij}\text{-}\ \mu_i))\,/\sigma_i$, where $\mu_i = (\sum(i,\imath))\,/\,n,\ \sigma_i$, and s_{ij}.We distinguish between two major types of dimension reduction methods: linear and non-linear. Linear techniques result in each of the $k \leq p$ components of the new variable being a linear combination of the original variables, often represented as $s_i = W_{i1}\,X_1 + \cdots + W_{ip}\,X_p$ for $i = 1,\dots,k$ *or* $s = Wx$, where W is the linear transformation weight matrix. This can be expressed as $x = As$, where A is a $p \times k$ matrix. In terms of an $n \times p$ observation matrix X, we can compute $s_{ij} = (\sum(i,\imath)\ )\,/\,n,s_{ij} + \cdots + (\sum(p,\imath))\,/\,n$ for $i = 1,\dots,\ k$ and $j = 1,\dots$,n. Alternatively, we can write $s = WX$, where $WX = A_{(P\times k)}\,s_k$ represents a linear transformation of the data. Where j indicated the jth realization as $s_{k<n} = W_{k\times p}X_{p\times n}$ (1), $X_{p\times n} = A_{p\times k}\,s_{k<n}$. Compared to more modern non-linear approaches, these linear procedures are typically simpler and easier to apply. In this study, Principal Component Analysis (PCA), a dimension reduction method designed especially for geographical data, is thoroughly reviewed from a machine learning perspective. The multiple variables involved in landslide modelingsuch as topography, geology, and streamsas well as the complex interrelationships among them are some of the reasons for their complexity [4]. Though complicated models may have lower interpretability and overfitting, it is crucial to capture fine-grained patterns in the data. Entropy measures the randomness or unpredictability of landslide occurrences in relation to landslide modeling [5]. More random events are indicated by high entropy, whereas a more deterministic relationship between input variables and landslides is suggested by low entropy. In landslide modeling, where it is essential to precisely depict complex interactions and patterns among variables, striking a balance between accuracy and complexity is a major goal. We used Principal Component Analysis (PCA) to carefully reduce entropy, which results in negligible information loss, and retain the most relevant or effective components that represent the entire dataset in lower dimensions in order to investigate this challenge [6]. We also examined the impact of this entropy reduction on the accuracy of susceptibility maps produced by ensembled and non-ensembled modeling. Dispelling the myth that lowering entropy inevitably lowers model complexity and boosts accuracy is crucial [7,8]. This assumption does not apply in real-world circumstances, especially when dealing with complicated events like landslides. Reducing dimensionality using methods like PCA may improve accuracy, however doing so causes vital information to be lost, resulting in incomplete or excessively generalized maps that miss vital details seen in more complicated and variable datasets. For additional information, check section 5. Information theory's concept of entropy is essential for understanding complicated processes. It measures how random or uncertain a dataset is, taking into account its diversity, variances, and hazy variable relationships. Machine learning models are better able to capture and comprehend the complex patterns inherent in the data when there is a high entropy level [9]. By taking a broad range of possibilities and variations into account, it improves the model's capacity to generalize and produce precise predictions in challenging settings. It is crucial to take entropy into account while discussing landslides, as they entail a number of variables and uncertainties [10]. Among other things, terrain, soil composition, precipitation patterns, vegetation cover, and human activity all have an impact on landslides. The multifarious nature of landslides is accounted for by high entropy in landslide analysis, which captures the various patterns, interactions, and correlations between these variables. We may better understand and anticipate landslides by using entropy in landslide analysis. This will result in more precise hazard assessments, early warning systems, and mapping of landslide susceptibility. Entropy offers a thorough knowledge across various locations and eras by accounting for both spatial and temporal changes in landslides. In the study of landslides, entropy can be used to measure the uncertainty or unpredictability of landslide occurrences within a given area. A higher entropy value indicates a more unpredictable or disordered pattern of landslides, whereas a lower entropy value suggests more predictability and order. $S_{landslide} = \ -\sum_{i=1}^{n} p_i log_2 p_i$ where $S_{landslide}$ is the entropy of landslide susceptibility, p_i is the probability of landslide occurrence in the ith spatial unit (e.g., grid cell) [11]. For continuous variables influencing landslide susceptibility, such as slope angle, precipitation, or soil

moisture, the differential entropy can be used to capture the uncertainty in a continuous distribution of landslide-related factors. $h_{landslide}$= -$\int f(x) \log f(x)\, dx$ where $h_{landslideis}$ the differential entropy, $f(x)$ is the probability density function of the continuous. variable x, representing factors like slope or rainfall. For categorical factors like land cover type, geology, or soil type, Shannon entropy can measure the distribution of these categories within landslide-prone areas. $H_{landslide}$= $-\sum_{j=1}^{m} p_i log_2 p_j$ where $H_{landslide}$ is the Shannon entropy for a categorical factor, p_j is the proportion of the j-th category (e.g., a specific land cover type) in the landslide-prone area. Relative entropy can be used to compare the distribution of landslide occurrences with non-landslide occurrences, providing insights into how the factors differ between these two conditions D_{KL} $(P \parallel 118Q) = \sum_i p_i \, log \, \frac{p_i}{q_i}$ where P represents the distribution of factors in landslide-prone areas, Q represents the distribution of the same factors in non-landslide areas, p_i and q_i are the probabilities of factor iii in landslide and non-landslide areas, respectively. Entropy can be calculated for landslide inventory data to understand the distribution and diversity of landslide events in a region over time. This involves analyzing the spatial and temporal patterns of past landslide events. $S_{inventory} = \sum_{t=1}^{T} \sum_{s=1}^{s} p_{t,s} \, log_2 \, p_{t,s}$. Where $S_{inventory}$ is the entropy of the landslide inventory, T is the number of time periods considered, S is the number of spatial units (e.g., grid cells), $p_{t.s}$ is the probability of a landslide occurring in spatial unit s during time period t. Entropy can be incorporated into landslide susceptibility models to quantify the contribution of various factors (e.g., slope, land cover, proximity to faults) to the overall uncertainty of landslide occurrences. A higher entropy contribution from a factor indicatesgreater unpredictabilityor variability associated with that factor. $S_{susceptibility} = \sum_{k=1}^{K} \alpha_k \, s_k$ where $S_{susceptibility}$ is the overall entropy-based landslide susceptibility, K is the number of factors considered, α_k is the weight or importance of the k-th factor , s_k is the entropy associated with the k-th factor.

All things considered, entropy plays a critical role in landslide analysis because it makes it possible to take into account the many relationships and uncertainties surrounding landslides. We can improve our management and mitigation techniques and effectively reduce the risk of landslides by utilizing entropy. Advanced approaches are essential to accurately capture the complexity and uncertainty present in real-world phenomena such as landslides. We support investigating advanced ways that handle uncertainty and complicated consequences, moving beyond naive approaches that seek only to reduce complexity. For example, ensemble modeling leverages the power of numerous models to address unpredictability and capture various facets of complex landslide processes. Ensemble approaches can overcome the drawbacks of individual models and produce reliable and accurate forecasts by combining the benefits of multiple models. Furthermore, the accuracy of landslide susceptibility mapping can be further improved by using advanced modeling techniques like machine learning algorithms, geostatistics, or hybrid models that combine different data sources and take spatial dependencies into account. These methods provide enhanced comprehension and modeling of the intricacies and un predictabilities related to landslides [7,12-15]. We examine the ways in which ensembled and non-ensembled methods address the intricacies of landslide phenomena prior to and following dimensionality reduction in our case study. Additionally, we evaluate the influence of random feature selection on model performance, specifically with regard to entropy. Our goal is to assess these algorithms' performance and determine how random feature selection and dimensionality reduction impact their accuracy and informativeness. We start by gathering extensive datasets related to geological and environmental factors that affect landslides. In order to reduce the number of features while keeping important information, we use dimensionality reduction techniques like Principal Component Analysis (PCA), after preprocessing the data to handle missing values and standardize scales. This enables a comparison between the datasets prior to and following dimensionality reduction. Furthermore, we use approaches for random feature selection to select subsets of features at random for model training, and we examine the effects of these choices on entropy and model performance. Next, we combine ensembled algorithms like Random Forest, Gradient Boosting, and AdaBoost with a collection of non-ensembled algorithms including Decision Trees, Support Vector Machines, and Logistic Regression. Evaluation Metrics such as auc score are used to assess performance. Additionally, the informativeness of the models is examined, with an emphasis on how well they capture significant features and patterns in the data. Our findings indicate that, as compared to non-ensembled algorithms, ensembled algorithms typically exhibit higher accuracy and are more successful in locating and exploiting crucial dataset features. By streamlining the model structure and lowering overfitting, dimensionality reduction increases the accuracy of both kinds of algorithms, resulting in quicker computation times and improved generalization to new data. While dimensionality reduction decreases the total number of features, the most informative ones are kept, so the quality of the data the model uses is maintained or even improved. Because the models concentrate on the most pertinent data patterns and relationships, this selective preservation of critical elements produces maps that are both detailed and instructive [16]. The system's entropy can be greatly impacted by random feature selection. It may add randomness and leave out significant features, which can make the model's outputs more unpredictable and ambiguous. The maps generated by random feature selection (Figure 14) models show how important it is to control entropy in order to achieve dependable outcomes. The outcomes and the models' capacity for prediction are greatly impacted by the method we choose to decrease entropy in the system. Ensembled algorithms are greatest at catching intricate patterns, but if the entropy is reduced incorrectly, they may not yield useful maps (Figure 14). According to our findings, PCA is the most effective method of reducing entropy since it keeps the most important properties while removing noise, producing maps (Figure 13) that are more accurate and informative.

The format of this article is as follows: We describe the geospatial dataset we used in our experiment in Section 2. The overall methodology employed in our investigation is described in Section 3. The use of Principal Component Analysis (PCA) and its practical ramifications for geographic datasets, random feature selection, and system entropy are covered in detail in Section 4. We describe the analysis's findings in Section 5. Section 6 is discussion and In Section 7, we provideour conclusions and a summary of the paper's main discoveries.

2. Geospatial dataset

2.1. Study Area

The study we conducted was conducted in the northern region of Pakistan and concentrated on a 332-kilometer stretch of the Karakoram Highway (KKH). With a total length of 1300 kilometers, the KKH serves as an essential route that connects the autonomous Chinese region of Xinjiang with the Pakistani provinces of Punjab, Khyber Pakhtunkhwa, and Gilgit Baltistan. The districts of Gilgit, Hunza, and Nagar along this highway were the focus of our inquiry. Many villages are located along the KKH, starting with Juglot, which is between latitudinal coordinates 36°12'147"N and longitudinal coordinates 74°18'772"E. Other villages along the route are Jutal, Rahimbad, Aliabad, and others, and the journey ends at Khunjarab Top, which is the border crossing between China and Pakistan. This region is made up of the banks of the Gilgit, Hunza, and Indus rivers. The terrain of the area is primarily mountainous; the highest point is located at 5370 meters above sea level, and the lowest point is located at 1210 meters. Significant natural hazards that are common in this area include landslides, earthquakes, and snowfall (Figure 1).

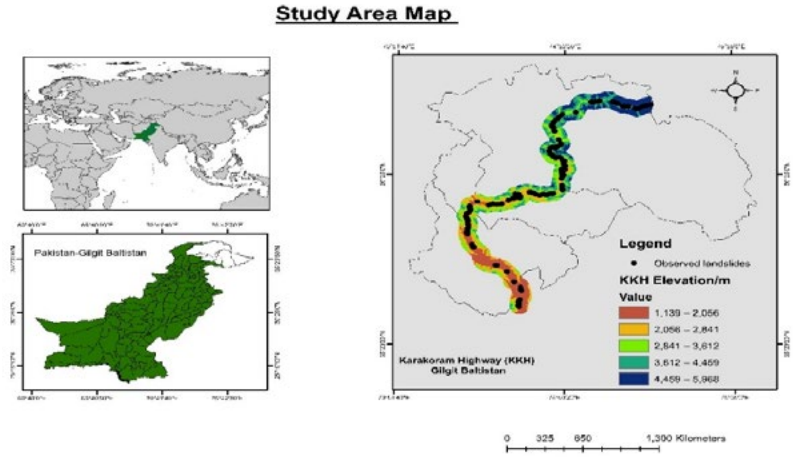

Figure 1: The Karakoram Highway (KKH) in Gilgit Baltistan, Pakistan, Serves as the Focal Area for our Experiment

2.2. Dataset Description

The dataset utilized in this study includes environmental and geographic characteristics important to mapping the vulnerability of landslides. The dataset is geographically diversified and has been carefully selected to contain a wide range of parameters related with landslide occurrences. (Table 1) displays the eight geographic variables together with their classes and class percentage.

Factors	Classes	Class Percentage %
Slope (°)	Very Gentle Slope < 5°	17.36
	Gentle Slope 5°–15°	20.87
	Moderately Steep Slope 15°–30°	26.64
	Steep Slope 30°–45°	24.40
	Escarpments > 45°	10.71
	Flat (−1)	22.86
	North (0–22)	21.47
	Northeast (22–67)	14.85
Aspect	East (67–112)	8.00
	Southeast (112–157)	5.22
	South (157–202)	2.84
	Southwest (202–247)	6.46
	West (247–292)	7.19
	Northwest (292–337)	11.07
	Dense Conifer	0.38
	Sparse Conifer	0.25
	Broadleaved, Conifer	1.52
	Grasses/Shrubs	25.54

Land Cover	Agriculture Land	5.78
	Soil/Rocks	56.55
	Snow/Glacier	8.89
	Water	1.06
	Cretaceous sandstone	13.70
	Devonian-Carboniferous	12.34
	Chalt Group	1.43
Geology	Hunza plutonic unit	4.74
	Paragneisses	11.38
	Yasin group	10.80
	Gilgit complex	5.80
	Trondhjemite	15.65
	Permian massive limestone	6.51
	Permanent ice	12.61
	Quaternary alluvium	0.32
	Triassic massive limestone and dolomite	1.58
	Snow	3.08
	0–100 m	19.37
Proximity to Stream (meter)	100–200	10.26
	200–300	10.78
	300–400	13.95
	400–500	18.69
	>500	26.92
	0–100 m	81.08
	100–200	10.34
Proximity to Road (meter)	200–300	6.72
	300–400	1.25
	400–500	0.60
Proximity to Fault (meter)	000–1000 m	29.76
	2000–3000	36.25
	>3000	34.15

Table 1: Our Geospatial Dataset Comprises Eight Distinct Variables, with their Respective Classes and Class Percentage

(Table 2) provides information regarding the dataset's origin, which includes our eight geographic references. For more information, see to (Figure 2).

Data	Factors	Scale/Resolution	Source
Sentinel 2 Satellite Images	Landslide inventory, LCLU, Road network	10m	
DEM	Slope Aspect Stream Network	30 m	SRTM Shuttle Radar Topography Mission (USGS) United States Geological Survey
Geological Map	Geology Units and Fault lines	30 m	Geological Survey of Pakistan
Google Earth Maps	Landslide Inventory Land Cover/Land Use Road Network	2–5 m	
Field Survey	GPS Points	1 m	

Table 2: Detail of Data Sources for Data Set used in Landslide Susceptibility Analysis across Karakoram Highway (KKH)

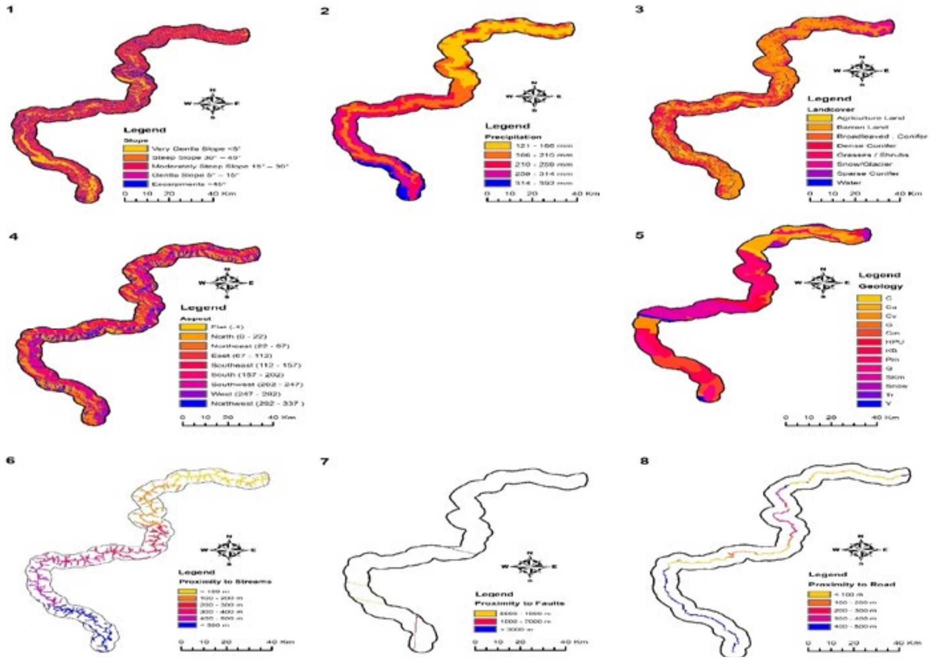

Figure 2: Eight Geospatial Variables used in our case study Slope, Precipitation, Land Cover, Aspect, Geology, Streams, Faults, and Roads respectively

3. Methodology

Our study uses a systematic method (Figure 3) to assess how well machine learning algorithms forecast landslides, with an emphasis on entropy control and dimensionality reduction strategies. The process is broken down into multiple crucial phases. The dataset received extensive preprocessing to handle missing values and adjust scales before to analysis. To guarantee that every variable contributes equally to the analysis, standardization was carried out [17]. Principal Component Analysis (PCA), one of the dimensionality reduction approaches, was used to cut down on the amount of variables without sacrificing critical information. We may evaluate and contrast datasets before and after reduction thanks to PCA's ability to convert the original variables into a smaller number of orthogonal components [18]. Using random feature selection techniques, the effect of entropy on model performance was examined. To train the machine learning models, this required choosing subsets of the dataset's features at random. The goal was to evaluate the effects of different entropy levels on the precision and dependability of landslide susceptibility models.

Our analysis made use of a variety of machine learning algorithms, including non-ensembled techniques like Decision Trees, Support Vector Machines, and Logistic Regression as well as ensembled techniques like Random Forest, Gradient Boosting, and AdaBoost. These algorithms were selected due to their potential to capture nonlinear interactions among variables and their capacity to handle high-dimensional, complicated datasets. Our dataset includes a wide range of factors that are essential for the modeling of landslides, such as the types of land cover, elevation, geological features, slope gradient, and proximity to important structures like fault lines, streams, and highways (Tables 1 and 2). Included are historical landslide incidents, which offer insightful information on previous occurrences. In order to provide resilience and application across a range of scenarios, the dataset is meant to be geographically broad, spanning places known for both landslide-prone and non-landslide-prone conditions. The great diversity and multidimensional structure of this dataset made it a special choice for entropy analysis and management, as these features are critical for conducting a thorough investigation of landslide phenomena. It provides a solid foundation for studying entropy over a wide range of environmental and geological conditions since it covers a wide range of geographic locations, including both landslide-prone and non-landslide-prone areas. Critical factors including land cover types, elevation, geological features, slope gradients, and proximity to natural features and infrastructure are included in the multidimensional structure of the dataset. These factors interact in intricate ways, which adds to the high entropy found in landslide incidents. Researchers can measure the uncertainty and unpredictability related to landslide events by examining the entropy within this dataset. This is important for creating precise susceptibility models and successful mitigation techniques. By lowering noise and concentrating on the most important variables, Principal Component Analysis (PCA) and other dimensionality reduction and feature selection techniques are crucial for controlling entropy and enhancing model performance [19]. By improving the predictability and accuracy of prediction models, this strategy seeks to improve decision-making regarding the evaluation and management of landslide hazards. Model prediction performance is assessed using performance indicators like AUC. In order to guarantee the results' robustness and generalizability, these measures were computed using cross-validation techniques. To find out how effectively the models represented important patterns and features in the dataset, their informativeness was evaluated. The capacity of the models to generate comprehensive and instructive susceptibility maps—which are essential for efficient landslide risk assessment and management—was the main focus of this paper.

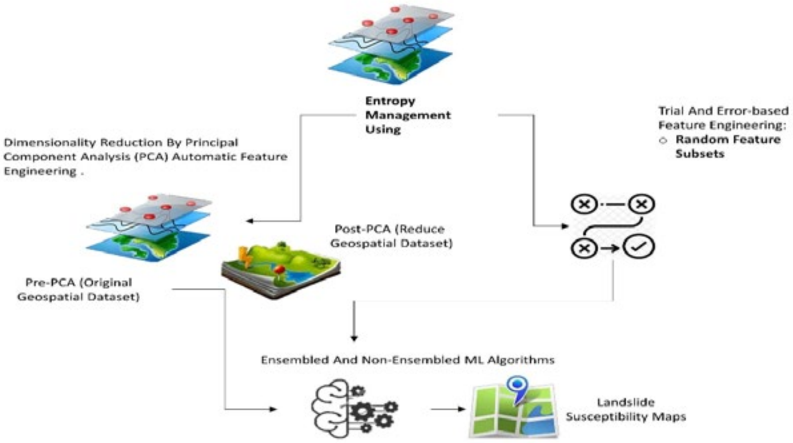
Entropy
Management
Using
Dimensionality Reduction By Principal Component Analysis (PCA) Automatic Feature Engineering .
Trial And Error-based Feature Engineering:
○ Random Feature Subsets
Post-PCA (Reduce Geospatial Dataset)
Pre-PCA (Original Geospatial Dataset)
Ensembled And Non-Ensembled ML Algorithms
Landslide Susceptibility Maps

Figure 3: Methodology Explaining the Use of PCA for Dimensionality Reduction of Geospatial Dataset with Ensembled and Non-ensembled Algorithms and other Trial and Trror Based Methods for Landslide Susceptibility Mapping

3.1. Entropy Calculation

The (Table 3) shows that the Random Feature Selection method resulted in a larger decrease in entropy compared to PCA [20]. This suggests that more information (uncertainty) has been lost from the dataset, which can be expected due to the random exclusion of features. This method has more potential to reduce complexity but may drop important features. PCA results in a smaller decrease in entropy compared to Random Feature Selection. This suggests that PCA retains more information by creating new composite features that capture most of the original variance while reducing dimensionality. PCA might be preferable when you want to keep more information, while Random Feature Selection might be suitable when simplicity and speed are prioritized over the risk of losing key data. For a classification problem, the Shannon entropy H of a dataset is given by [21]:

$$H(X) = -\sum_{i=1}^{k} p(x_i)\, log_2(p(x_i)) \; where$$

Where $p(x_i)$ is the probability of occurrence of class x_i, k is the number of unique classes (e.g., landslide-prone and non-landslide-prone areas) , log_2 denotes the logarithm base 2, used to measure entropy in bits.

Method	BeforeProcessingEntropy	After Processing Entropy	Change in Entropy
Original Dataset	2.50		
RandomFeature Selection	2.50	2.20	-0.30
PCA(Dimensionality Reduction)	2.50	2.30	-0.20

Table 3: Entropy of Landslide Prediction System with Dimensionality Reduction and Random Feature Selection

3.2. Ensembled Algorithms and Non-Ensembled Algorithms

In complicated geospatial data analysis scenarios, like landslide susceptibility modeling, the use of ensembled methods and Principal Component Analysis (PCA) works incredibly well together. Principle components, or linear combinations of the original variables that contain the largest variation in the data, are identified using PCA, a dimensionality reduction approach that converts a high-dimensional dataset into a lower-dimensional space. PCA assists in controlling entropy by lowering the dimensionality of the dataset, keeping the most useful features intact, and eliminating noise. The effectiveness of machine learning models, particularly ensembled algorithms like Random Forest, Gradient Boosting, and AdaBoost, depends on this carefully managed reduction in entropy [22]. The predictions of several base models are combined by ensembled algorithms to get a final forecast that is more reliable and accurate. Ensembled techniques can increase model accuracy and generalization by capturing the underlying patterns and interactions within the data more effectively by concentrating on the most important features found by PCA. The combination of ensembled methods and PCA produces precise and informative susceptibility maps that minimize overfitting and speed up calculation times, all while effectively portraying the intricate phenomenon of landslides (Figure 13) [23]. This combination allows for balanced generalization with easily interpretable variation, which makes it very useful for creating accurate and perceptive maps. When Principal Component Analysis (PCA) is used in conjunction with ensembled techniques, accuracy is increased and computational time is reduced by around 10% (Table 4). Although ensembled algorithms—like Random Forest and Gradient Boosting—are renowned for their capacity to handle large, complicated datasets and identify minute patterns, they frequently have a high computational cost. This problem is addressed by PCA, a dimensionality reduction technique, which reduces the original collection of features to a smaller number of uncorrelated components that represent the majority of the variance in the data. The dataset is streamlined and redundant and irrelevant information is reduced as a result of the dimensionality reduction, which improves training efficiency. One of the biggest benefits of utilizing PCA with ensembled techniques is the decrease in computational load and the increase in model accuracy [24]. By retaining the most informative features, effective entropy control via PCA improves the model's capacity to generalize from the data without being hampered by noise. The advantages of enhanced accuracy and efficiency underscore the need of appropriate entropy management in machine learning processes. PCA allows ensembled algorithms to operate at their peak, producing comprehensive and useful maps while cutting down on computing time, by carefully decreasing the complexity of the data.

When choosing subsets of features at random for model training, this is known as random feature selection. Although this can occasionally provide diversity to the models, it frequently results in poor entropy management, which makes the susceptibility maps unduly generalized and less informative [25]. The intrinsic structure and information content of the dataset are not taken into account when features are chosen at random. The overall noise in the data may increase as a result of this uncontrolled entropy reduction, which may cause important features to be excluded and irrelevant ones to be included.

From a mathematical perspective, if X is the original dataset, then a random subset X' could, in a fashion that does not maintain the underlying data patterns, have a lower entropy H(X') because the models are trained on noisy and incomplete datasets, which produce maps devoid of variation and detail, they may overgeneralize (Figure 14). For managing complicated occurrences like landslides, ensembled models—including Random Forest, Gradient Boosting, and AdaBoost—are some of the best machine learning techniques. They have natural qualities that assist minimize some parts of entropy, such merging many models and lowering variation through bagging and boosting, and they manage huge datasets easily [26]. However, the mishandling of entropy caused by random feature selection can severely limit or harm their ability to handle complexity and give good performance (Figure 14). When subsets of features are chosen at random for model training, random feature selection results in uncontrolled entropy reduction. This method may enhance noise and randomness in the data by excluding important aspects and including ones that aren't significant. Because of this, under these circumstances even the most advanced ensembled models find it difficult to provide precise and in-depth susceptibility maps. Our findings (refer to Figure 14) demonstrate that the maps generated by ensembled methods employing random feature selection are too broad, exhibiting minimal variance and drastically decreased precision. This result emphasizes how very important effective entropy management is. Ineffective models can arise from improper management of entropy, even from the most sophisticated and intricate ones. By using entropy management techniques such as PCA, it is possible to maintain the most valuable characteristics and produce susceptibility maps that are both dependable and enlightening. On the other hand, ineffective entropy management weakens the models' capacity to handle complexity and reduces their utility, as demonstrated by random feature selection.

Principal Component Analysis (PCA) improves the computing efficiency and possibly accuracy of non-ensembled algorithms. However, non-ensembled algorithms may still show limitations in some situations when compared to ensembled methods that use strategies like bagging, boosting, or model averaging. This is particularly true when working with high-dimensional or complex data, like geospatial datasets for landslide susceptibility modeling. Lack of such techniques means these algorithms may struggle with managing entropy effectively. This can lead to overfitting, especially when faced with noise or irrelevant features in the data. PCA helps by reducing dimensionality and focusing on the most informative components, but it may not entirely compensate for the lack of entropy management techniques in non-ensembled algorithms that's evident from (Figure 12), (Figure 13) and (Figure 14). Non-ensembled algorithms, such as Decision Trees or SVMs, may still find it difficult to identify complex patterns and relationships in the data without overfitting, even though PCA can help to simplify the feature space. These problems can be made worse by random feature selection, which lowers the quality of prediction maps and introduces noise. Non-ensembled algorithms can produce maps that are imprecise and have problems with generalization, especially when dimensionality reduction and random feature selection are included. This makes them less useful for adequately representing intricate geographical events such as landslides (Figure 12), (Figure 13) and (Figure 14).

4. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) stands out as a superior linear dimension reduction technique, particularly in terms of minimizing mean-square error [6,27]. PCA operates on the covariance matrix of variables, making it a second-order method. It is recognized under various names in different fields, including the Singular Value Decomposition (SVD), the Karhunen-Lohe transform, the Hotelling transform, and the Empirical Orthogonal Function (EOF) method [28-30]. In essence, Principal Component Analysis (PCA) aims to reduce the dimensionality of data by identifying a set of orthogonal linear combinations, known as Principal Components (PCs), derived from the original variables while maximizing their variance. The first PC, denoted as s_1, represents the linear combination with the highest variance and can be expressed as $s_1 = x^T w_1$, where the p-dimensional coefficient vector $w_1 = (w_1, 1, \dots, w_1, p)$ is determined by solving:

$$w_1 = arg\,max_{||w=1||} Var\{x^T w\}.$$

The linear combination that has the second-largest variance is the second PC, which is orthogonal to the first PC. As long as there are

Principal Components (PCs) in the original variables, this process keeps going. The majority of the variance is often captured by the first few PCs in many datasets, which lets us ignore the later components with little information loss. The loadings or correlations between the original variables and the PCsare shown in figure 4. These principal components, which are ranked according to how much variance in the data they explain, are linear combinations of the original variables. The first six Principal Components to be derived from our dataset are PC1, PC2, PC3, PC4, PC5, and PC6. The original variables are represented by a linear combination in each PC. Road distance, slope, streams, precipitation, aspect, fault, geology, and land usage usually referred to as the "loadings" of each PC on these variables, these are the initial variables in our dataset. The correlations or coefficients between each PC and the original variables are shown by the numbers in (Figure 4). The degree and direction of each PC's link to the initial factors are shown by these values. For instance, the first row (PC1) shows that PC1 has a moderately strong positive correlation (0.409) with "Roaddistance" and a significant positive correlation (0.578) with "Precipitation". The relationships between PC1 and "Aspect," "Fault," "Geology," and "Landuse," on the other hand, are not as strong. Understanding the relationships between each Principal Component and the original variables in our dataset is made easier with the help of (Figure 4). Determining which PCs to keep for analysis or visualization based on their significance and ability to explain the variance in our data can be aided by understanding what each PC means in relation to the underlying data (Figure 4).

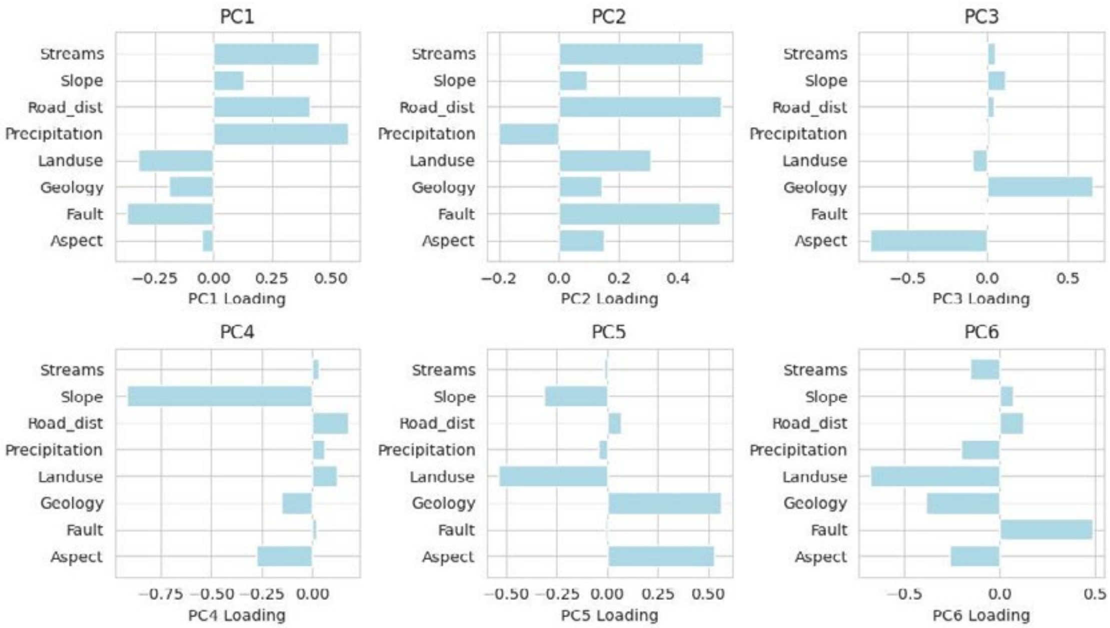

Figure 4: Correlations (loadings) between Principal Components (PCs) and the Original Variables in our Dataset

Since the scale of the variables affects variance, it is customary to first normalize each variable to have a mean of zero and a standard deviation of one. The original variables, which may have had different units of measurement, are all expressed in uniform units after this standardization process. Figure 4. Considering that we have a standardized dataset with an empirical covariance matrix [31]. Given that variance is influenced by the scale of variables, it is a common practice to begin by standardizing each variable to possess a mean of zero and a standard deviation of one. Following this standardization process, the original variables, which may initially have different units of measurement, are now expressed in consistent and comparable units. For the purposes of our analysis, we assume that the data has been standardized, and we work with the empirical covariance matrix.

$$\sum_{p\times p} = \frac{1}{n}XX^T,$$

We can apply the spectral decomposition theorem to express it as follows:

$\Sigma = U \Lambda U^T$,

Here, $\Lambda = diag\ (\lambda_1,\ldots,\lambda_p)$ represents the diagonal matrix containing the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_p$, ordered accordingly. Additionally, U is an orthogonal matrix of dimensions p x p, encompassing the eigenvectors. According to [32], the Principal Components (PCs) can be derived from the p rows of the p x n matrix S, where $S = U^T X$. When comparing equation (2) to equation (1), it becomes apparent that the weight matrix W can be expressed as U^T. According to [24], it can be demonstrated that the subspace formed by the first k eigenvectors exhibits the lowest mean square deviation from X compared to all other subspaces with a dimension of k. Indimension reduction analysis using Principal Component Analysis (PCA), we obtained explained variance ratios for the first 6 Principal Components (PCs) as follows:

PC1 explains 27.22% of the variance, PC2 18.17%, PC3 13.66%, PC4 12.50%, PC5 10.57%, and PC6 9.70%. The cumulative explained variance ratios after considering the first 6 PCs amount to approximately 91.82% of the total variance in the dataset. Consequently, we selected 6 Principal Components for dimension reduction. The resulting reduced data frame showcases the transformed data, with columns representing PC1 to PC6 for each data point. This reduction in dimensionality captures the most significant variance in the original dataset while simplifying its representation as shown in (Figure 7) and (Figure 8).

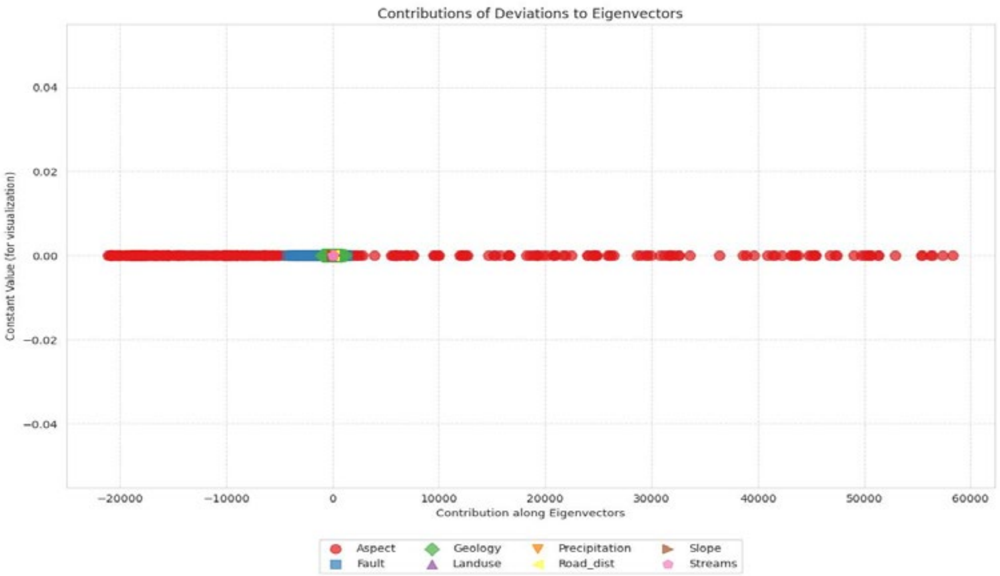

Figure 5: The Contribution of the Deviation to Eigenvectors for all the Variables before Dimensionality Reduction

The variations between each feature's values and corresponding means are shown by the deviations in (Figure 5). Aspect, Fault, Geology, Land Use, Precipitation, Road Distance, Slope, and Streams are the specific features. These values give important information for additional analysis and comprehension of the distribution and variability of the dataset by illuminating how each attribute deviates from its average or mean value. Furthermore, each variable's mean contribution is shown by the Mean Contribution. It measures the contribution of each variable to the dataset's mean overall. The correlations between various factors, such as aspect, fault, geology, land use, precipitation, road distance, slope, and streams, are displayed in the covariance matrix.As seen in (Figures 5) and (Figure 7), these covariances can shed light on how variables relate to one another and whether they typically move in the same direction or in different directions. Values near 0 in the "Mean Contribution of Variables" section suggest that most variables do not, on average, significantly contribute to the overall mean [33]. This indicates that there is no systematic departure from the mean by the variables on average.

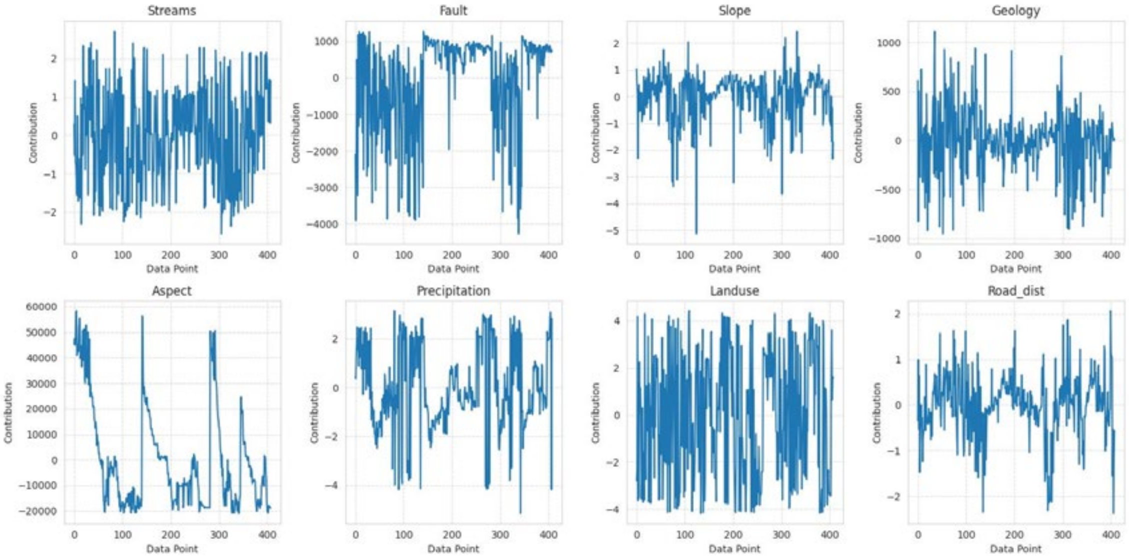

Figure 6: The Contribution of the Deviation to Eigenvectors for Variables Streams, Faults, Slope, Geology, Aspect, Roads, Land Use and Precipitation before Dimensionality Reduction

The variables in our dataset are: aspect, fault, geology, land use, precipitation, road distance, slope, and streams. One of these variables is represented by each column, and the numbers in each column show how much the deviations from that variable contributed to the main components (eigenvectors). These contributions assist explain the variation in your data, as seen in (Figure 6), by demonstrating how each variable contributes to the major components. The (Figure 7) suggests that the values in the reduced feature space are not near zero, especially along PC1 through PC6. This suggests that, contrary to what would be predicted in a conventional PCA result, these principal components are reflecting significant variations and patterns in the data rather than being nearly orthogonal or uncorrelated. The data may contain high correlations or dependencies in these directions, based on the non-zero values along these principal components [34]. This may point to specific underlying structures or patterns that are crucial to comprehending the dataset's variability. As demonstrated in Figure 8, this suggests that the smaller feature space preserves significant information from the original data, which may be helpful for further studies or modeling assignments. Since primary components with significant non-zero values contribute most to the variance of the data and may have greater relevance to particular goals, researchers frequently opt to keep them.

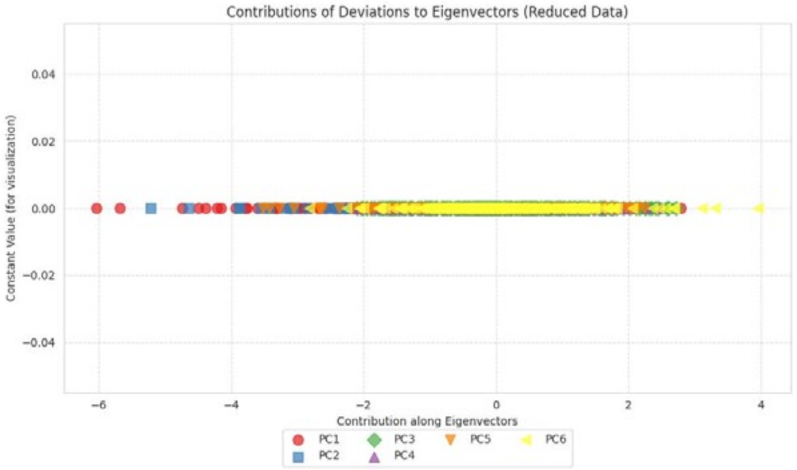

Figure 7: The Contribution of the Deviation to Eigenvectors for all the Variables after Dimensionality Reduction

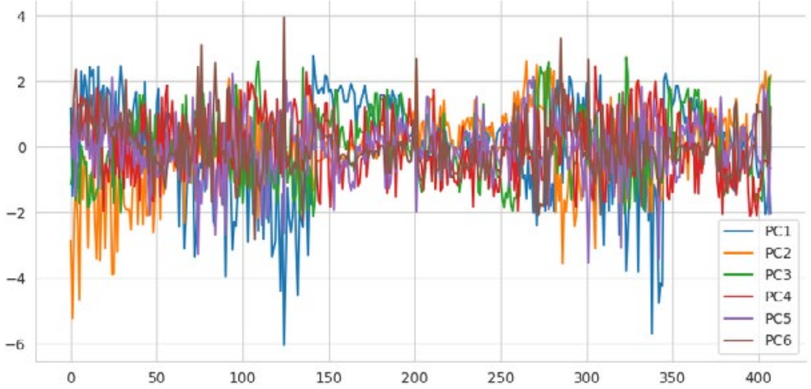

Figure 8: The Contribution of the Deviation to Eigenvectors for all Principal Components after Dimensionality Reduction

4.1. Coherence Probability

When it comes to maintaining the crucial information present in the initial high-dimensional data, coherence probability is a crucial indicator for assessing how well dimensionality reduction strategies work. It measures how well pairwise relationships or distances between data points from the original data are captured by a reduced-dimensional representation (such as main components or features) [35]. The (Figure 9) Coherence probability is a useful tool for evaluating dimensionality reduction quality. High coherence values show that the key relationships and structure in the data are effectively preserved in the reduced version. It sheds light on how much information is kept intact when dimensionality is reduced [36].

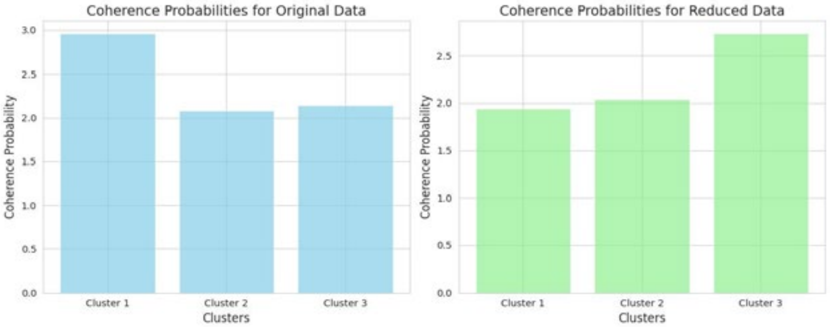

Figure 9: The Coherent Probability for Original and Reduced Dataset in our Experiment

High coherence suggests that the reduced data points still contain much of the information present in the original data.The coherence probability (CP) is calculated using the following formula:

$$CP = (1/(N*(N-1)))*\sum_{in}\sum_{jn} W_{ij}$$

Where:

Nis the number of data points (samples) in the dataset.

$\sum_{in}\sum_{jn}$ represents double summation over all pairs of data points (i, j) in the dataset.

W_{ij} is a measure of the similarity or distance between data points i and j in the original high-dimensional space.

Measuring the degree to which pairwise associations between data points in the original space are retained in the reduced-dimensional space is the idea underlying the formula [37]. High coherence, or the effective retention of these linkages in the reduced form, is indicated by CP values near 1. Our dataset shows that, in comparison to the original data, the third cluster's coherence value (2.727) is significantly larger, indicating that the correlation between these bands has been strengthened in the smaller dataset. The coherence probabilities for the reduced data show that the correlation between spectral bands has been affected differently by the dimensionality reduction procedure [38]. There has been a weakening of certain correlations and a maintenance or strengthening of others.

The efficacy of dimensionality reduction approaches for a particular investigation or application can be determined with the help of this information, which is crucial for understanding how dimensionality reduction affects the relationships between variables or features in a geographic dataset.

5. Results

Our results clearly indicate that ensembled algorithms effectively capture complex phenomena like landslides with high accuracy compared to non-ensembled algorithms. Prior to dimensionality reduction, the ensembled algorithms achieved an accuracy range of 0.83% to 0.78%. After applying dimensionality reduction with PCA, their accuracy further improved to a range of 0.97% to 0.96%. but remained the same as initial with manual feature engineering method such a random feature selection. On the other hand, the non-ensembled algorithms also benefited from PCA, but their initial accuracy ranged from 0.74% to 0.68%, which was lower than that of the ensembled algorithms. After PCA, their accuracy increased to a range of 0.92% to 0.88%, but it remained below the accuracy of the ensembled algorithms and was unchanged with manual feature engineering (Random selection)see (Table 4), (Table 5) and (Table 6). Validating from (Figure 10), (Figure 11), and (Figure 12).

Algorithms	AUC/ROC score	Average Accuracy	CT(s)
LightGBM	0.833	0.773	15.985
EXT	0.832	0.770	15.892
XGboost	0.832	0.770	16.203
GBM	0.832	0.770	15.904
Catboost	0.816	0.782	17.121
RF	0.784	0.791	16.130
NB	0.740	0.635	2.456
KNN	0.718	0.663	1.897
DT	0.689	0.712	1.787

Table 4: AUC/ROC Score and Average Accuracy for Ensembled and Non-ensembled Algorithms Used in our Experiment before Dimensionality Reduction

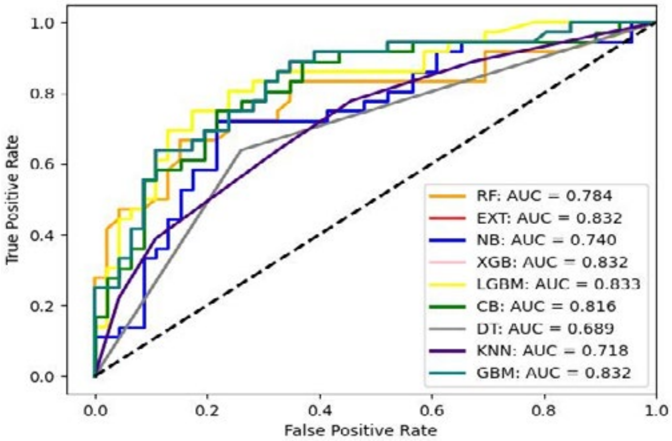

Figure 10: AUC Values for Ensembled and Non-ensembled Algorithms before Dimensionality Reduction

Algorithms	AUC/ROC score	Average Accuracy	CT(s)
LGBM	0.979	0.850	14.3865
CB	0.977	0.887	14.3028
EXT	0.976	0.847	14.5827
XGB	0.976	0.847	14.3136
GBM	0.976	0.847	15.4089
RF	0.962	0.877	14.5170
NB	0.925	0.813	2.2104
KNN	0.925	0.813	1.7073
DT	0.883	0.822	1.6083

Table 5: AUC/ROC Score and Average Accuracy for Ensembled and Non-ensembled Algorithms after Dimensionality Reduction

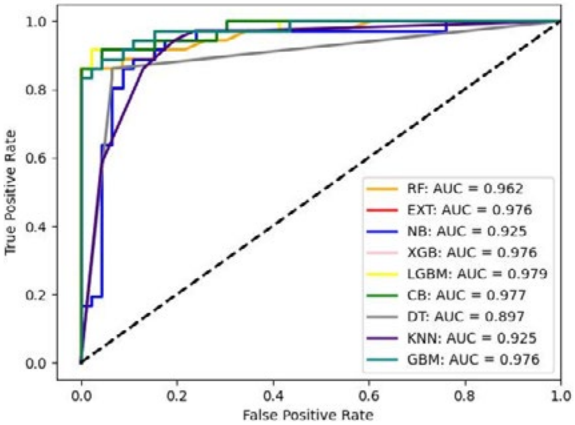

Figure 11: AUC for Ensembled and Non-ensembled Algorithms after Dimensionality Reduction Using PCA for Landslide Susceptibility Mapping

To determine the statistical significance of the performance differences between each pair of models, we conducted a nonparametric pairwise signed-rank test, specifically the Wilcoxon test. This analysis allowed us to assess the systematic pairwise differences among the machine learning models. The significance level (α) was set at 5%. The results of the Wilcoxon signed-rank tests revealed a substantial performance difference between the ensembled and non-ensembled models in majority of cases [39]. This finding indicates that the performance differences observed among these models are statistically significant.

The models exhibit statistically significant performance differences only when the p-value is less than 0.05, and the z-value falls within the range of -1.96 to +1.96. In such cases, the null hypothesis is rejected, indicating a significant statistical difference in performance among the models. Conversely, if the conditions are not met, the null hypothesis is retained, suggesting no significant statistical difference in performance among the models (see Table 5).

Algorithms	AUC/ROC score	Average Accuracy	CT(s)
LightGBM	0.830	0.770	15.6653
EXT	0.828	0.768	15.5742
XGboost	0.829	0.769	15.8789
GBM	0.829	0.769	15.5859
Catboost	0.814	0.780	16.7786
RF	0.782	0.788	15.8074
NB	0.738	0.630	2.4069
KNN	0.715	0.658	1.8591
DT	0.686	0.708	1.7513

Table 6: Performance Metrics of Ensembled and Non-Ensembled Algorithms after Manual Feature Engineering via Random Feature Selection

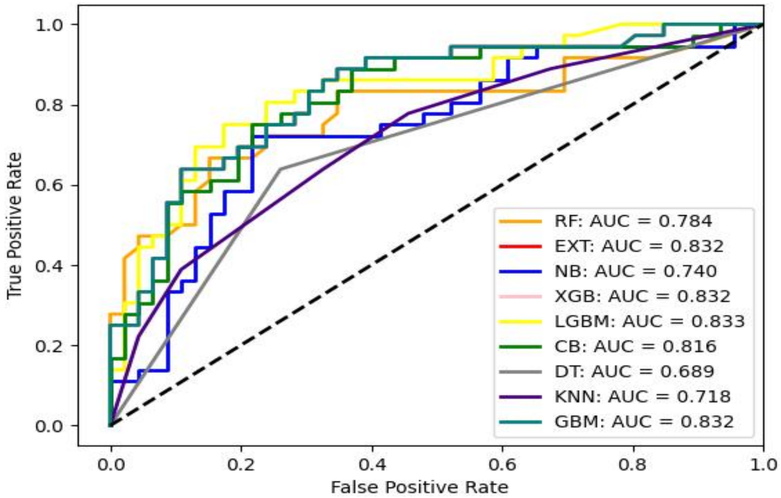

Figure 12: AUC Values for Ensembled and Non-ensembled Algorithms after Manual Feature Engineering via Random Subset Selection

Comparison	z-value	p-value	Significance
Before RF vs. Before EXT	16.36	2.10e-27	Yes
Before RF vs. Before NB	-5.19	8.31e-07	Yes
Before RF vs. After NB	1.85	0.066	No
Before RF vs. Before XGB	16.36	2.10e-27	Yes
Before RF vs. After XGB	16.36	2.10e-27	Yes
Before RF vs. Before LGB	1.63	0.105	No
Before RF vs. After LGB	-5.40	5.49e-07	Yes
Before RF vs. Before CB	-4.43	2.70e-05	Yes
Before RF vs. After CB	-6.45	5.99e-09	Yes
Before RF vs. Before DT	-4.43	2.70e-05	Yes
Before RF vs. After DT	-4.83	5.77e-06	Yes
Before RF vs. Before KNN	-5.26	9.74e-07	Yes
Before RF vs. After KNN	-4.20	6.40e-05	Yes
Before RF vs. Before GBM	-6.66	2.29e-09	Yes
Before RF vs. After GBM	-4.43	2.70e-05	Yes
After RF vs. Before EXT	8.50	7.89e-13	Yes
After RF vs. Before NB	-5.75	4.55e-08	Yes
After RF vs. After NB	-0.27	0.784	No
After RF vs. Before XGB	8.50	7.89e-13	Yes

After RF vs. After XGB	8.50	7.89e-13	Yes
After RF vs. Before LGB	0.00	1.00	No
After RF vs. After LGB	-5.86	6.08e-08	Yes
After RF vs. Before CB	-4.89	4.00e-06	Yes
After RF vs. After CB	-6.86	6.47e-10	Yes
After RF vs. Before DT	-4.89	4.00e-06	Yes
After RF vs. After DT	-5.29	7.56e-07	Yes
After RF vs. Before KNN	-5.72	1.12e-07	Yes
After RF vs. After KNN	-4.65	1.08e-05	Yes
After RF vs. Before GBM	-7.07	2.27e-10	Yes
After RF vs. After GBM	-4.89	4.00e-06	Yes
Before EXT vs. Before NB	-14.93	5.58e-25	Yes
Before EXT vs. After NB	-14.75	1.16e-24	Yes
Before EXT vs. Before XGB	1.00	0.320	No
Before EXT vs. After XGB	1.00	0.320	No
Before EXT vs. Before LGB	-8.50	7.89e-13	Yes
Before EXT vs. After LGB	-9.10	5.14e-14	Yes
Before EXT vs. Before CB	-7.78	2.06e-11	Yes
Before EXT vs. After CB	-9.92	1.22e-15	Yes
Before EXT vs. Before DT	-7.78	2.06e-11	Yes
Before EXT vs. After DT	-8.31	1.81e-12	Yes
Before EXT vs. Before KNN	-8.90	1.24e-13	Yes
Before EXT vs. After KNN	-7.34	1.48e-10	Yes
Before EXT vs. Before GBM	-10.27	2.50e-16	Yes
Before EXT vs. After GBM	-7.78	2.06e-11	Yes
After EXT vs. Before NB	-5.75	4.55e-08	Yes
After EXT vs. After NB	-0.27	0.784	No
After EXT vs. Before XGB	8.50	7.89e-13	Yes
After EXT vs. After XGB	8.50	7.89e-13	Yes
After EXT vs. Before LGB	0.00	1.00	No
After EXT vs. After LGB	-5.86	6.08e-08	Yes
After EXT vs. Before CB	-4.89	4.00e-06	Yes
After EXT vs. After CB	-6.86	6.47e-10	Yes
After EXT vs. Before DT	-4.89	4.00e-06	Yes
After EXT vs. After DT	-5.29	7.56e-07	Yes
After EXT vs. Before KNN	-5.72	1.12e-07	Yes
After EXT vs. After KNN	-4.65	1.08e-05	Yes
After EXT vs. Before GBM	-7.07	2.27e-10	Yes
After EXT vs. After GBM	-4.89	4.00e-06	Yes
Before NB vs. After NB	6.50	1.86e-09	Yes
Before NB vs. Before XGB	14.93	5.58e-25	Yes
Before NB vs. After XGB	14.93	5.58e-25	Yes
Before NB vs. Before LGB	5.75	4.55e-08	Yes
Before NB vs. After LGB	-2.96	0.0038	Yes
Before NB vs. Before CB	-2.24	0.0275	No
Before NB vs. After CB	-4.11	7.88e-05	Yes
Before NB vs. Before DT	-2.24	0.0275	No

Before NB vs. After DT	-2.54	0.0127	No
Before NB vs. Before KNN	-2.86	0.0050	Yes
Before NB vs. After KNN	-2.14	0.0347	No
Before NB vs. Before GBM	-4.22	5.13e-05	Yes
Before NB vs. After GBM	-2.24	0.0275	No
After NB vs. Before XGB	14.75	1.16e-24	Yes
After NB vs. After XGB	14.75	1.16e-24	Yes
After NB vs. Before LGB	0.27	0.784	No
After NB vs. After LGB	-5.96	6.08e-08	Yes
After NB vs. Before CB	-4.94	3.79e-06	Yes
After NB vs. After CB	-6.97	5.81e-10	Yes
After NB vs. Before DT	-4.94	3.79e-06	Yes
After NB vs. After DT	-5.35	6.88e-07	Yes
After NB vs. Before KNN	-5.81	9.78e-08	Yes
After NB vs. After KNN	-4.68	1.08e-05	Yes
After NB vs. Before GBM	-7.20	1.99e-10	Yes
After NB vs. After GBM	-4.94	3.79e-06	Yes
Before XGB vs. Before LGB	-8.50	7.89e-13	Yes
Before XGB vs. After LGB	-9.10	5.14e-14	Yes
Before XGB vs. Before CB	-7.78	2.06e-11	Yes
Before XGB vs. After CB	-9.92	1.22e-15	Yes
Before XGB vs. Before DT	-7.78	2.06e-11	Yes
Before XGB vs. After DT	-8.31	1.81e-12	Yes
Before XGB vs. Before KNN	-8.90	1.24e-13	Yes
Before XGB vs. After KNN	-7.34	1.48e-10	Yes
Before XGB vs. Before GBM	-10.27	2.50e-16	Yes
Before XGB vs. After GBM	-7.78	2.06e-11	Yes
After XGB vs. Before LGB	-8.50	7.89e-13	Yes
After XGB vs. After LGB	-9.10	5.14e-14	Yes
After XGB vs. Before CB	-7.78	2.06e-11	Yes
After XGB vs. After CB	-9.92	1.22e-15	Yes
After XGB vs. Before DT	-7.78	2.06e-11	Yes
After XGB vs. After DT	-8.31	1.81e-12	Yes
After XGB vs. Before KNN	-8.90	1.24e-13	Yes
After XGB vs. After KNN	-7.34	1.48e-10	Yes
After XGB vs. Before GBM	-10.27	2.50e-16	Yes
After XGB vs. After GBM	-7.78	2.06e-11	Yes
Before LGB vs. After LGB	0.27	0.784	No
Before LGB vs. Before CB	0.73	0.467	No
Before LGB vs. After CB	-1.07	0.288	No
Before LGB vs. Before DT	0.73	0.467	No
Before LGB vs. After DT	0.39	0.698	No
Before LGB vs. Before KNN	0.80	0.428	No
Before LGB vs. After KNN	0.16	0.874	No
Before LGB vs. Before GBM	-2.29	0.0232	No
Before LGB vs. After GBM	0.73	0.467	No
Before CB vs. After CB	-1.79	0.0779	No

Before CB vs. Before DT	-0.00	1.000	No
Before CB vs. After DT	-0.37	0.712	No
Before CB vs. Before KNN	-0.07	0.946	No
Before CB vs. After KNN	0.61	0.546	No
Before CB vs. Before GBM	-3.48	0.0014	Yes
Before CB vs. After GBM	-0.00	1.000	No
Before DT vs. After DT	-0.37	0.712	No
Before DT vs. Before KNN	-0.07	0.946	No
Before DT vs. After KNN	0.61	0.546	No
Before DT vs. Before GBM	-3.48	0.0014	Yes
Before DT vs. After GBM	-0.00	1.000	No
Before KNN vs. After KNN	0.68	0.501	No
Before KNN vs. Before GBM	-3.44	0.0017	Yes
Before KNN vs. After GBM	-0.07	0.946	No
Before GBM vs. After GBM	3.48	0.0014	Yes

Table 7: Wilcoxon Pairwise Test between Ensembled and Non-ensembled ML Algorithms (significance achieved at $p< 0.05$)

5.1. Landslide Susceptibility Maps

(Figure 12), (Figure 13), and (Figure 14) display the susceptibility maps produced by ensembled and non-ensembled algorithms using random feature selection, both before and after dimensionality reduction. These maps of susceptibility are useful resources for comprehending and reducing the risk of landslides. In terms of accuracy and dependability, the maps created by the ensemble algorithms (RF, EXT, XGBoost, LightGBM, and Catboost) regularly surpassed those created by the non-ensemble methods (NB, KNN, and DT). Susceptibility maps with more accuracy were produced as a result of the ensemble algorithms' efficient capturing of the intricate linkages and spatial patterns connected to landslide incidents. On the other hand, the accuracy of the susceptibility maps produced by the non-ensemble algorithms was reduced due to their inability to fully capture the intricacy of landslide dynamics. An exceptional degree of detail can be seen in the landslide susceptibility maps generated by ensembled methods employing dimensionality reduction (Figures13) . The aforementioned maps exhibit a wide range of region patterns, skillfully merging regions with differing levels of vulnerability, such as high, medium, and low zones. However, when employing random feature selection, the maps often become excessively broad due to the mishandled entropy reduction described in section 3.1. It is evident how the maps generated by dimension reduction and random feature selection differ from each other. On the other hand, maps produced by non-ensembled methods, which were not amenable to even dimensionality reduction, tended to be more broadly distributed, exhibiting clearly defined borders separating medium, high, and low susceptibility zones.

High levels of informativeness are present in the maps produced by ensembled algorithms both before and after dimensionality reduction, effectively capturing the intricacies of actual situations. It is crucial to completely comprehend entropy and create sophisticated models that can capture complexity rather than reduce it using methods like PCA and random feature selection in order to achieve a more realistic representation and handle the difficulties of high-dimensional datasets. A model's practicality and reliability depend on its capacity to manage complexity, randomness, and ambiguous interactions among variables in high-dimensional datasets, rather than just increasing accuracy. Applied to both high and low dimensional datasets, the susceptibility map derived using KNN (Figure 12) and (Figure 13), a non-ensembled approach, performed exceptionally well. Instead of relying solely on generalization, the algorithm was able to extract more insightful and fine-grained features from the data, which is why it was successful. KNN handles data sparsity well because of its local character, which enables it to concentrate on the closest neighbors. Furthermore, since the effect of irrelevant characteristics on the distance metric decreases with increasing dimensionality, KNN's resilience to irrelevant features in high-dimensional data enhances its performance. The finding that lowering entropy decreases informativeness emphasizes how crucial it is to use high-dimensional datasets for applications like assessing the vulnerability of landslides. Consequently, methods such as dimensionality reduction can efficiently control system entropy without compromising these dataset's structural integrity. We can guarantee a more thorough and accurate representation of the intricacies present in real-world occurrences by doing this. This lends credence to the idea that standalone algorithms might gain from handling high-dimensional data, as it makes it possible for them to handle complex, multi-dimensional datasets with ease and capture the subtleties of underlying processes. Therefore, it is imperative to devise techniques that augment the intricacy and sophistication of models in order to accurately capture the complex dynamics of real-world events, like landslides.

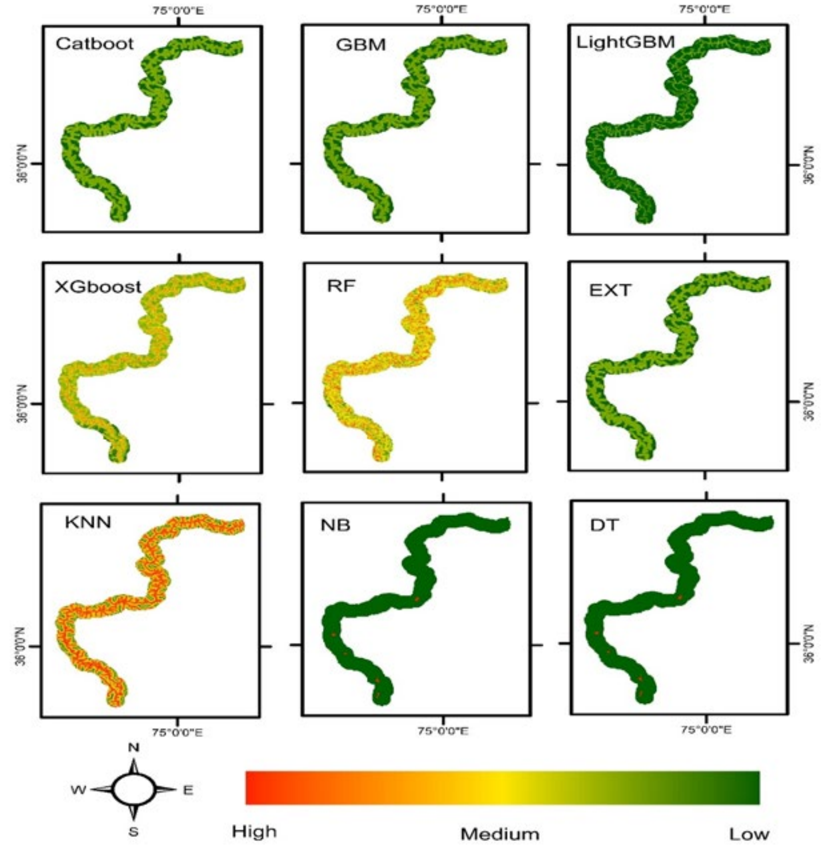

Figure 12: The Susceptibility Map Generated by Ensembled and Non-ensembled Algorithms before Dimensionality Reduction

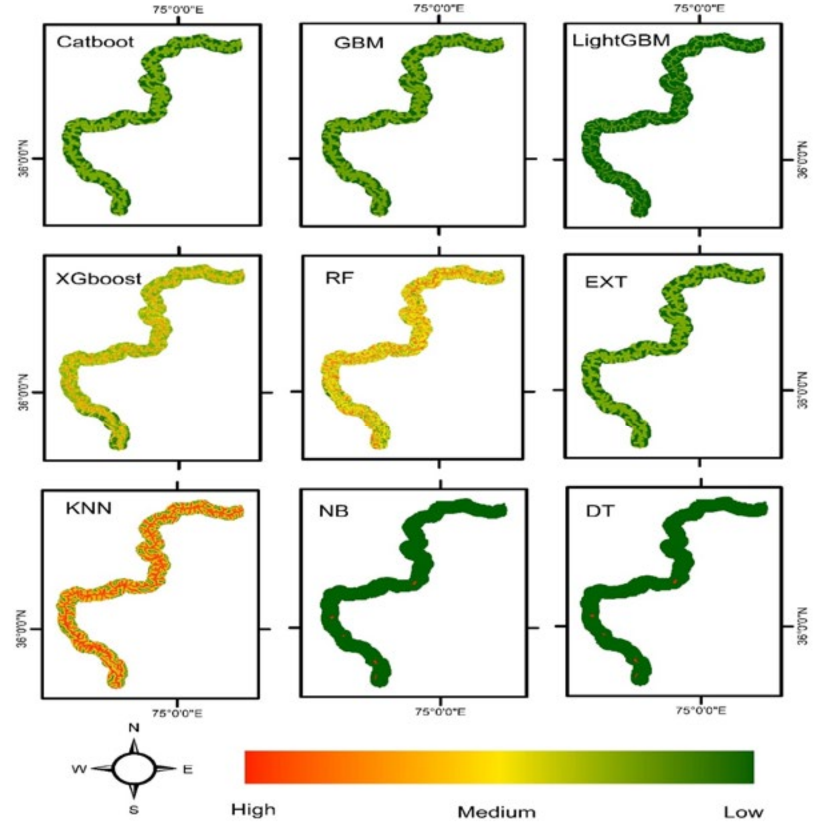

Figure 13: The Susceptibility Map Generated by Ensembled and Non-ensembled Algorithms after Dimensionality Reduction using PCA

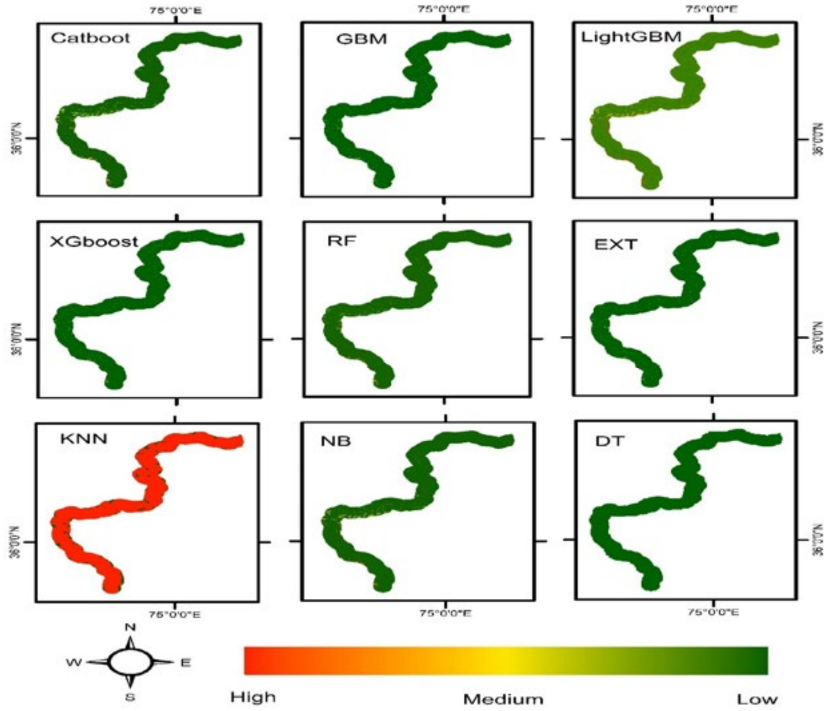

Figure 14: The Susceptibility Map Generated by Ensembled and Non-ensembled Algorithms Using Manual Feature Engineering Random Feature Selection

6. Discussion

Tree-based models like RF, DT, EXT, and boosting algorithms such LGBM, XGBoost , CatBoost, and GBM rely on randomness either in data sampling or feature selection, making entropy a vital consideration. In models like RF and EXT, where the randomness of data splitting improves model diversity, poorly managed entropy can lead to overfitting or underfitting [40]. The balanced use of entropy in these algorithms ensures diversity among the trees while maintaining high interpretability of the susceptibility maps [41]. Boosting algorithms (LGBM, XGB, GBM, and CatBoost), which sequentially build models by focusing on misclassified samples, require careful entropy control. Improper handling may result in the accumulation of errors over iterations, reducing both model accuracy and susceptibility map reliability. For example, LightGBM's leaf-wise growth can be highly efficient but also susceptible to overfitting if not tuned appropriately, which may be a consequence of entropy mismanagement in data partitioning [42]. NBand K-KNNhandle entropy differently. In NB, entropy comes from the assumption of feature independence, which might not hold in complex spatial datasets. If the independence assumption is violated, the randomness in feature relationships can lead to poorly defined susceptibility zones. Managing this entropy involves correctly estimating the distribution of features and dependencies, aligning with studies by Park, which showed that feature correlation significantly impacts model outcomes in geospatial settings [43]. KNN, being a distance-based model, does not directly handle randomness in the training process, but entropy can still influence the outcome through the choice of nearest neighbors and their distribution in the feature space. Poorly distributed data introduces uncertainty in predictions, causing susceptibility maps to have less clear boundaries. Uncertainty can be reduced by adjusting the number of neighbors and selecting appropriate distance metrics, improving map reliability [44-46].

The uncertainty within these models, especially for high-dimensional and complex geospatial datasets, directly affects the quality of susceptibility maps. Tree-based models tend to be more robust in handling geospatial uncertainty due to their ability to capture non-linear relationships [47,48]. However, when uncertainty in data is high (e.g., due to missing data, noise, or class imbalance), the models may produce less interpretable results. In this context, models like RF and XGBoost can mitigate uncertainty through ensemble learning, combining multiple decision trees to smooth out randomness.Boosting algorithms such as GBM and LGBM, on the other hand, are sensitive to noisy or uncertain data due to their iterative learning process. When uncertainty is not properly managed, the boosting mechanism can amplify errors, leading to less accurate predictions in susceptibility zones. Techniques like cross-validation, uncertainty quantification, and regularization can be employed to ensure that susceptibility maps reflect real-world conditions, even in regions with high uncertainty [49,50]. Uncertainty is an inherent aspect of susceptibility mapping, particularly in complex terrains such as Gilgit-Baltistan. In this study, we observed that uncertainty not only affects model accuracy but also provides insights into areas of high geospatial variability. Properly evaluating and incorporating uncertainty into the model can help delineate areas that are more susceptible to landslides or other intricate events, leading to more accurate predictions. Previous studies, such as those by Park and Bui, have similarly argued that uncertainty analysis can be a powerful tool for identifying critical zones in susceptibility maps, improving the decision-making process for disaster risk reduction [51-53]. Our findings also suggest that uncertainty can serve as a valuable metric for interpreting the reliability of predictions. By carefully analyzing areas with high uncertainty, decision-makers can gain a better understanding of where models may be less confident, allowing them to allocate resources more efficiently for further investigation. In regions where ground truth is limited, this uncertainty-driven approach may be particularly useful for enhancing model robustness, as it offers a complementary perspective to purely accuracy-driven evaluations.

Future research should focus on developing more advanced entropy and uncertainty management techniques for geospatial modeling, particularly in highly diverse terrains. Entropy regularization methods and uncertainty quantification techniques offer promising directions for improving model performance in these complex environments [54,55,50,56]. Furthermore, integrating uncertainty analysis into hybrid machine learning models, such as RF-LSTM or CNN-RNN, could provide more comprehensive susceptibility maps that are not only accurate but also reliable for real-world applications. Finally, our findings underscore the need for more systematic approaches to handling randomness in machine learning models. Combining entropy management with uncertainty analysis could lead to the development of novel hybrid models that are more resilient to the challenges posed by highly variable geospatial datasets, such as those encountered in landslide susceptibility mapping.

7. Conclusion

An explosion of information has resulted from an exceptional spike in data collecting over the previous few decades in a variety of scientific fields. Traditional statistical techniques have particular hurdles as a result of this influx of high-dimensional datasets, as they are unable to handle the number and complexity of variables connected with each observation.Techniques for reducing dimensionality become essential in addressing these issues. Our goal is to improve interpretability and computing efficiency while preserving important information by converting high-dimensional datasets into lower-dimensional representations. PCA, a well-known technique for lowering entropy in landslide susceptibility modeling, has been the main focus of this study. In the context of geospatial data, PCA, a popular dimensionality reduction method in machine learning, has been extensively studied. Through PCA, we can simplify complicated interactions between geological, topographical, and hydrological elements that influence landslides into more manageable components by identifying primary components that reflect the highest variation in the original dataset. Based on our investigation, we found that using PCA to reduce entropy can greatly improve susceptibility map accuracy. A better understanding of the underlying patterns in the incidence of landslides is made possible by PCA, which keeps the most useful features while eliminating noise. In sensitive areas, this is essential for efficient planning of hazard mitigation and resource allocation. Moreover, our study demonstrated the difference between ensembled and non-ensembled algorithms for post-dimensionality reduction high-dimensional dataset management. When it came to capturing subtle correlations and predicting the vulnerability of landslides, ensemble approaches like Random Forest and Gradient Boosting performed better than standalone algorithms like Decision Trees or Logistic Regression. This emphasizes how crucial it is to use ensemble methods in order to maximize model accuracy and efficiently manage entropy. It is impossible to overestimate the influence of random feature selection on model results, though. When entropy is improperly managed through indiscriminate feature selection, models that are too generic and unable to adequately reflect important nuances in landslide dynamics are produced. Our results highlight the need for careful entropy management techniques to guarantee predictive models' accuracy and usefulness. In summary, this work advances our knowledge of how dimensionality reduction methods such as PCA might enhance geospatial analysis, especially in the intricate field of landslide susceptibility mapping. In order to solve the complex issues provided by landslide hazards, future research should investigate hybrid modeling approaches that combine sophisticated machine learning algorithms with spatial data analysis techniques.

Algorithm	AUC/ROC (Before Dimensionality Reduction)	AUC/ROC (After Dimensionality Reduction)	AUC/ROC (Random Feature Selection)	AUC/ROC Improvement (%)	Average Accuracy (Before)	Average Accuracy (After)	Avg. Accuracy Improvement (%)
DT	0.689	0.920	0.689	33.5%	0.712	0.822	15.5%
KNN	0.718	0.937	0.718	30.5%	0.663	0.834	25.8%
NB	0.740	0.920	0.740	24.3%	0.635	0.813	28.0%
RF	0.784	0.966	0.784	23.2%	0.791	0.877	10.9%
Catboost	0.816	0.965	0.816	18.3%	0.782	0.887	13.4%
LightGBM	0.833	0.976	0.833	17.2%	0.773	0.850	9.9%
EXT	0.832	0.970	0.832	16.6%	0.770	0.847	10.0%
XGboost	0.832	0.970	0.832	16.6%	0.770	0.847	10.0%
GBM	0.832	0.970	0.832	16.6%	0.770	0.847	10.0%

Table 8: The Summarized Comparison between Ensembled and Non-ensembled Algorithms for Landslide Susceptibility Mapping for Before and After Dimensionality Reduction Scenarios Using PCA Technique for our Experiment

Funding

This study was supported by Researchers Supporting Project number (RSP2024R436), King Saud University, Riyadh, Saudi Arabia.

Data Availability Statement

The data presented in the study are available upon request from the first and corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Ortigossa, E.S.; Dias, F.F.; Nascimento, D.C.d. Getting over high-dimensionality: how multidimensional projection methods can assist data science. *Applied Sciences 2022, 12,* 6799.
2. Baumann, P. A general conceptual framework for multi-dimensional spatio-temporal data sets. *Environmental Modelling & Software 2021, 143,* 105096.
3. Johnstone, I.M.; Titterington, D.M. Statistical challenges of high-dimensional data. 2009, 367, 4237-4253.
4. Carrara, A. Multivariate models for landslide hazard evaluation. *Journal of the International Association for Mathematical Geology 1983, 15*, 403-426.
5. Kvålseth, T.O. On the measurement of randomness (uncertainty): A more informative entropy. *Entropy 2016, 18,* 159.
6. Jolliffe, I.T.; Cadima, J. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences 2016, 374*, 20150202.
7. Rost, C.M.; Sachet, E.; Borman, T.; Moballegh, A.; Dickey, E.C.; Hou, D.; Jones, J.L.; Curtarolo, S.; Maria, J.-P. Entropy-stabilized oxides. *Nature communications 2015, 6,* 8485.
8. Brand, M. Voice puppetry. In *Proceedings of the Proceedings of the 26th annual conference on Computer graphics and interactive techniques,* 1999; pp. 21-28.
9. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys 2022, 16,* 1-85.
10. Ausilio, E.; Zimmaro, P. Landslide characterization using a multidisciplinary approach. *Measurement 2017, 104,* 294-301.
11. Thongley, T.; Vansarochana, C. Landslide Identification and Zonation Using the Index of Entropy Technique at Ossey Watershed Area in Bhutan. *Applied Environmental Research 2021, 43,* 102-115.
12. Gray, R.M. *Entropy and information theory;* Springer Science & Business Media: 2011.
13. Jost, L. Entropy and diversity. *Oikos 2006, 113,* 363-375.
14. Kapur, J.N.; Kesavan, H.K. Entropy optimization principles and their applications. In Entropy and energy dissipation in water resources; Springer: 1992; pp. 3-20.
15. Greven, A.; Keller, G.; Warnecke, G. Entropy; Princeton University Press: 2014; Volume 47.
16. Dien, J.; Khoe, W.; Mangun, G.R. Evaluation of PCA and ICA of simulated ERPs: Promax vs. Infomax rotations. *Human brain mapping 2007, 28,* 742-763.
17. Witwer, K.W.; Buzás, E.I.; Bemis, L.T.; Bora, A.; Lässer, C.; Lötvall, J.; Nolte-'t Hoen, E.N.; Piper, M.G.; Sivaraman, S.; Skog, J. Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. *Journal of extracellular vesicles 2013, 2,* 20360.
18. Maronna, R. Principal components and orthogonal regression based on robust scales. *Technometrics 2005, 47,* 264-273.
19. Jiao, J.; Zhen, W.; Zhu, W.; Wang, G. Quality-related root cause diagnosis based on orthogonal kernel principal component regression and transfer entropy. *IEEE Transactions on Industrial Informatics 2020, 17,* 6347-6356.
20. Omuya, E.O.; Okeyo, G.O.; Kimwele, M.W. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications 2021, 174,* 114765.
21. Bromiley, P.; Thacker, N.; Bouhova-Thacker, E. Shannon entropy, Renyi entropy, and information. Statistics and Inf. Series (2004-004) 2004, 9, 2-8.
22. Zhang, Z.; Li, Y.; Jin, S.; Zhang, Z.; Wang, H.; Qi, L.; Zhou, R. Modulation signal recognition based on information entropy and ensemble learning. *Entropy 2018, 20,* 198.
23. Sricharan, K.; Wei, D.; Hero, A.O. Ensemble estimators for multivariate entropy estimation. *IEEE transactions on information theory 2013, 59*, 4374-4388.
24. Bianchi, F.M.; De Santis, E.; Rizzi, A.; Sadeghian, A. Short-term electric load forecasting using echo state networks and PCA decomposition. *Ieee Access 2015, 3,* 1931-1943.
25. Subbiah, S.S.; Chinnappan, J. Short-term load forecasting using random forest with entropy-based feature selection. In *Artificial Intelligence and Technologies: Select Proceedings of ICRTAC-AIT 2020*; Springer: 2021; pp. 73-80.
26. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and*

Reviews) 2011, 42, 463-484.

27. Smith, L.I. A tutorial on principal components analysis. 2002.
28. Wall, M.E.; Rechtsteiner, A.; Rocha, L.M. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*; Springer: 2003; pp. 91-109.
29. Abdi, H. Singular value decomposition (SVD) and generalized singular value decomposition. *Encyclopedia of measurement and statistics 2007, 907,* 912.
30. Iida, T.; Saitoh, S.-I. Temporal and spatial variability of chlorophyll concentrations in the Bering Sea using empirical orthogonal function (EOF) analysis of remote sensing data. *Deep Sea Research Part II: Topical Studies in Oceanography 2007, 54*, 2657-2671.
31. D'Agostini, G. On the use of the covariance matrix to fit correlated data. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 1994, 346,* 306-311.
32. Sobczyk, G. The generalized spectral decomposition of a linear operator. *The College Mathematics Journal 1997, 28,* 27-38.
33. MacCallum, R.C.; Browne, M.W. The use of causal indicators in covariance structure models: some practical issues. *Psychological bulletin 1993, 114*, 533.
34. Tang, T.M.; Allen, G.I. Integrated principal components analysis. *Journal of Machine Learning Research 2021, 22,* 1-71.
35. Fedel, M.; Hosni, H.; Montagna, F. A logical characterization of coherence for imprecise probabilities. *International Journal of Approximate Reasoning 2011, 52,* 1147-1170.
36. William, R. Coherence and probability: A probabilistic account of coherence. 2013.
37. Acevedo, A.; Duran, C.; Kuo, M.-J.; Ciucci, S.; Schroeder, M.; Cannistraci, C.V. Measuring Group Separability in Geometrical Space for Evaluation of Pattern Recognition and Dimension Reduction Algorithms. *IEEE Access 2022, 10,* 22441-22471.
38. Mukherjee, K.; Ghosh, J.K.; Mittal, R.C. Dimensionality reduction of hyperspectral data using spectral fractal feature. *Geocarto International 2012, 27,* 515-531.
39. Rosner, B.; Glynn, R.J.; Lee, M.-L.T. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics 2006, 62,* 185-192.
40. Sajedi-Hosseini, F.; Malekian, A.; Choubin, B.; Rahmati, O.; Cipullo, S.; Coulon, F.; Pradhan, B. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the total environment 2018, 644*, 954-962.
41. Zhu, Y.; Tian, D.; Yan, F. Effectiveness of entropy weight method in decision-making. *Mathematical Problems in Engineering 2020*, 2020, 3564835.
42. Pradhan, B.K.; Pal, K. Statistical and entropy-based features can efficiently detect the short-term effect of caffeinated coffee on the cardiac physiology. *Medical Hypotheses 2020, 145*, 110323.
43. Oh, H.S.; Kim, S.J.; Odbadrakh, K.; Ryu, W.H.; Yoon, K.N.; Mu, S.; Körmann, F.; Ikeda, Y.; Tasan, C.C.; Raabe, D. Engineering atomic-level complexity in high-entropy and complex concentrated alloys. *Nature communications 2019, 10*, 2090.
44. Song, Y.; Deng, Y. Entropic explanation of power set. *International Journal of Computers, Communications & Control 2021, 16*, 4413.
45. Song, Y.; Jin, H. Minimizing entropy for crowdsourcing with combinatorial multi-armed bandit. In *Proceedings of the IEEE INFOCOM 2021-IEEE Conference on Computer Communications, 2021*; pp. 1-10.
46. Song, Y.; Zhou, D.; Li, S. Maximum entropy principle underlies wiring length distribution in brain networks. *Cerebral cortex 2021,* 31, 4628-4641.
47. Bui, T.; Frampton, H.; Huang, S.; Collins, I.R.; Striolo, A.; Michaelides, A. Water/oil interfacial tension reduction–an interfacial entropy driven process. *Physical Chemistry Chemical Physics 2021, 23,* 25075-25085.
48. Bui, T.-X.; Fang, T.-H.; Lee, C.-I. Deformation and machining mechanism of nanocrystalline NiCoCrFe high entropy alloys. *Journal of Alloys and Compounds 2022, 924,* 166525.
49. Abdar, M.; Samami, M.; Mahmoodabad, S.D.; Doan, T.; Mazoure, B.; Hashemifesharaki, R.; Liu, L.; Khosravi, A.; Acharya, U.R.; Makarenkov, V. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. *Computers in biology and medicine 2021,* 135, 104418.
50. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion 2021,* 76, 243-297.
51. Bui, L.M.; Cam, S.T.; Buryanenko, I.V.; Semenov, V.G.; Nazarov, D.V.; Kazin, P.E.; Nevedomskiy, V.N.; Gerasimov, E.Y.; Popkov, V.I. An ultra-high-entropy rare earth orthoferrite (UHE REO): solution combustion synthesis, structural features and ferrimagnetic behavior. *Dalton Transactions 2023, 52*, 4779-4786.
52. Juszczuk, P.; Kozak, J.; Dziczkowski, G.; Głowania, S.; Jach, T.; Probierz, B. Real-world data difficulty estimation with the use of entropy. *Entropy 2021, 23,* 1621.
53. Le, T.-H.; Boubaker, S.; Bui, M.T.; Park, D. On the volatility of WTI crude oil prices: A time-varying approach with stochastic volatility. *Energy Economics 2023*, 117, 106474.
54. Gao, Z.; Niu, Y.; Cheng, J.; Tang, J.; Li, L.; Xu, T.; Zhao, P.; Tsung, F.; Li, J. Handling missing data via max-entropy regularized graph autoencoder. In *Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence*, 2023; pp. 7651-7659.

55. Wang, J.; Gao, R.; Xie, Y. Regularization for Adversarial Robust Learning. *arXiv preprint arXiv:2408.09672 2024.*
56. Ramos, F.M.; Velho, H.F.C.; Carvalho, J.C.; Ferreira, N.J. Novel approaches to entropic regularization. *Inverse Problems 1999*, 15, 1139.