Landslide Risk Assessment using Granular Fuzzy Rule-Based Modeling: A Case Study on Earthquake-triggered Landslides

BINGXIN SHI1,3,4, TING ZENG2, CHUAN TANG1, LIFANG ZHANG5, ZHUOJUAN XIE5, GUOJUN LV6 and QIHONG WU7

1State Key Laboratory of Geohazard Prevention and Geoenvironment Protection, Chengdu University of Technology, Chengdu 610059, China
2Science & Technology for Development Research Center of Sichuan Province, Chengdu, 610041, China
3Sichuan Earthquake Administration, Boulder, Chengdu, 610041, China
4Institute of Engineering Mechanics, China Earthquake Administration, Harbin, 150080, China
5Natural Hazard,Ministry of Emergency Management of the People's Republic of China, Beijing, 100085, China
6Hebei Earthquake Administration, Shijiazhuang,050021, China
7School of Architecture and Civil Engineering, Chengdu University, Chengdu, 610106, China

Corresponding author: Ting Zeng (tzeng_scu@163.com).

This work was supported in part by scientific research fund of institute of engineering mechanics (China earthquake administration) under Grant 2020EEEVL0202, in part by the National Key R & D programs of China under Grant 2017YFC1500405-94, in part by the institute of crustal dynamics (China earthquake administration) under Grant ZDJ2019-11, and in part by the regional innovation cooperation programs of Sichuan province, China under Grant 2021YFQ0050.

ABSTRACT Landslides are one type of destructive and recurring natural calamities in the mountainous regions. The landslide occurrences often lead to immense damage to local infrastructure and loss of land, human lives and livestocks. Data-driven risk assessment of landslide risk plays a crucial role in preventing the incoming landslide occurrences and recurrences. In this research, we developed a human-centric framework using information-granules to perform risk assessment of a group of landslides. First, the density-based spatial clustering of applications with noise (DBSCAN) has been selected as the backbone unsupervised learning method to subclusters for landslide risk indication. The clustering outcomes are visualized via t-distributed stochastic neighbor embedding (t-SNE) in the 2-D embedding space. Second, the prototype points within the subclusters produced by DBSCAN are computed for granular construction. Third, interval-based information-granules are constructed and measured via coverage, specificity and area under the coverage-specificity curve (AUC). Last, with the optimal information-granules constructed, two risk measures namely Value-at-Risk (VaR) and Conditional-Value-at-Risk (CVaR) are computed to interpret the rule-based information-granules with respect to the key attributes. Comparative experiments have also been performed against other benchmarking clustering approaches. Computational results indicate that the information-granules constructed from DBSCAN subclusters offered enhanced performance in reveal meaningful information-granules and provide promising results. The proposed approach can capture the main essence of landslide pattern with higher interpretability and help to reduce the computing overhead.

INDEX TERMS DBSCAN, Landslide Risk, Information granule, VaR, CVaR

I. INTRODUCTION
Landslides are a global cascading geo-hazard that can result severe economic and societal losses. The majority of the landslide events are water-induced where the natural and human related water activities reduced the slope stability and induced slope failures [1]. Data-driven approaches to manage the landslides are becoming more popular in recent years as they necessarily characterize the landslides both internally and externally [2]. The external characteristics of landslides are often geomorphology related such as distance to rivers, distance to road, elevation, and slope. These external factors can be easily observed and measured via on-site investigation by geologists with domain knowledge. Meanwhile, measuring the internal characteristics is a more challenging task. The practitioners need to rely on point sensors, physical samples, soil samples, and many other in-depth characteristics. Both
external and internal measures are often merged to construct the meta-dataset for the risk analytics of landslide occurrences. Based on the meta-dataset, many conventional data-driven analysis and recent approaches can be performed to indicate the potential risk of the underlying landslide. However, despite many data-driven techniques aiming to assess the landslide risks using meta-data, the interpretability of such results is often insufficient and vague for readers. As a consequence, the reproducibility of those techniques can be only limited to a small dataset but can hardly be applied universally to the research problem with sufficient level of interpretability.

In the past decade, the data-driven approaches of assessing landslide risk have been widely and successfully applied. Generally, they can be categorized into conventional approaches and more recent approaches. For conventional models, Kirschbaum et al. [3] proposed a meta-heuristic fuzzy overlay model to create a regional susceptibility map and evaluated the performance using receiver operating characteristic (ROC) curves. Van et al. [4] integrated the analytical hierarchy process and weighted linear combination to construct the risk assessment model for landslide occurrences. Nefeslioglu et al. [5] improved the vanilla analytical hierarchy process and evaluated the landslide susceptibility. Ahmed. [6] performed comparative analysis among various state-of-the-art approaches such as Artificial Hierarchy Process (AHP), Weighted Linear Combination (WLC), and Ordered Weighted Average (OWA) to assess landslide risks. Nine different thematic layers or landslide causative factors were considered in the study. Overall, the statistical models provided promising results with limited factors considered and lower-level of nonlinearity between the landslide risks and these factors respectively.

For the time being, the more recent approaches using machine-learning and artificial intelligence are becoming the mainstream of predicting landslide risks. They often describe the highly nonlinear relationships between internal/external factors and landslide risks and can formulate the problem as classification or regression problem quantitatively. Althuwaynee et al. [7] initially performed multivariate analysis of landslide risk assessment using decision trees (DT) and logistic regression (LR) models. Huang et al. [8] modelled the landslide susceptibility using support vector machine (SVM) with a case study analysis in Nantian area of China. Youssef et al. [9] evaluated the landslide risk using random forest (RF) model to classify the landslide risk levels including low, moderate, high and extremely high. Eleven landslide-conditioning factors were prepared in the study which contains both internal and external factors. Gorsevski et al. [10] performed landslide risk assessment with a case study in the Cuyahoga Valley National Park, Ohio using artificial neural network (ANN) integrated with lidar data. The ANN algorithms provided superior performance in construct nonlinear mapping between the landslides and predictor attributes and the prediction accuracy outperforms on average. Huang et al. [11] further improved the prediction performance of landslide risk classification using a fully-connected sparse autoencoder which is a deep learning algorithm that contains higher capacity of handling nonlinearity between the risks and predictors.

Previous research of data-driven models on landslide risk assessment has their inherent advantages and disadvantages. For the advantages, the machine-learning models have provided improved modeling and prediction performance in risk assessment tasks with respect to accuracy and the capacity of handling highly nonlinear mapping between the input features and landslide risk levels. The mostly frequently used models such as ANN, SVM, RF, and ELM all produced high prediction accuracy while the well-trained models can be applied broadly to similar datasets. However, at the same time, all these models are in the format of “black-box” functions which lacks interpretability of the decision-making process. All geological domain knowledge is not adequately displayed within the “black-boxes” which inevitably deteriorate their wide application in the field.

In recent years, a new sub-field in data science namely granular computation has led to a series of breakthroughs in the explanation of the decision-making process. The granular computational approach construct data-driven information granular to indicate the decision boundaries based on the dataset with higher-level interpretability. In addition, the granular computation does not require tuning many hyper-parameters of the machine-learning models and thus can avoid overfitting to the local optimal solution. The granular computation has been successfully applied in many typical data-driven problems such as classification [12-15], regression [16], time-series prediction [17-18] and anomaly detection [19]. Especially inspired by the work of Ouyang et al. [14], the utilization of granular computation in landslide risk assessment has been adopted in this research.

Based on the discussed outlined above, a data-driven granular computation framework is proposed in this study to assess the landslide risks. First, the density-based spatial clustering of applications with noise (DBSCAN) algorithm is adopted to perform the clustering analysis to cluster the landslide dataset into subgroups. Second, a collection of prototypes within each cluster are selected as the representatives of each sub-cluster and the interval-based information granules are constructed accordingly. The neighboring data points are computed with respect to the density to discover the “thinnest” cluster which is the optimality with respect to coverage and specificity. In addition, to measure the classification performance of landslide risks using these constructed interval-based granular, three metrics including coverage, specificity, and area under the coverage-specificity curve (AUC) are computed. Comparative analysis is also performed against the benchmarking clustering algorithms such as fuzzy C-means (FCM), k-mean and k-medoids. The two statistical measures namely Value-at-Risk and Conditional-Value-at-Risk are computed to interpret the
generated rule-based information-granules. The main contribution of this paper is as follows:

- This research firstly introduced the concept of granular computation into the field of landslide risk assessment. None of the related work has been discussed in this field yet.
- Second, the proposed approach provided explicit decision boundaries with respect to attributes which illustrates the decision-making process of the landslide risk assessment. The level of interpretability outperforms the previous state-of-art machine-learning models and the decision rules can be easily adopted by field engineers.

To realize this proposed approach, this paper is organized as follows. Section II discuss the methods utilized in this research. Section III introduces the data collection process and the background of our case study area. Section IV illustrates experimental results including clustering analysis, interval-based granularity construction, sensitivity analysis, as well as the granular visualization and rule interpretation. Section V provides the discussion and Section VI presents the conclusion of this research.

II. METHODOLOGY

A. DBSCAN ALGORITHM

Density-based spatial clustering of applications with noise (DBSCAN) [20] is a popular density-based clustering algorithm which has been widely applied in many data-science domains with promising outcomes. It is designed to discover clusters of arbitrary shape without requiring prior knowledge of the number of clusters. There are two major contributions of DBSCAN: First, it is a formal model to generate clusters based on the density of data points. Second, it is a database-driven algorithm for cluster discovery that adheres to the underlying data model [21]. It has been successfully applied to various domains such as finance, environmental science, transportation, and astronomy.

The DBSCAN algorithm uses a simple minimum density level estimation which rely on two major parameters: the scanning radius $\varepsilon$ within an arbitrary distance and minPts which denotes the density threshold deciding whether a point is a core point or not. Given a dataset $S$ containing $N$ k-dimensional data points, the user-defined distance (e.g., Euclidean distance, etc) between the two data points $s_i$ and $s_j$ can be expressed as $d_{ij}$ and the density $\rho_{ij}$ of $s_j$ in the $\varepsilon$-neighborhood of $s_i$ can be expressed in (1) as follows:

$$\rho_{ij} = \sum_{j=1}^{N} I(d_{ij} \leq \varepsilon)$$

(1)

where $I()$ is the indicator function determining whether $s_j$ is within the $\varepsilon$-neighborhood of $s_i$. The concepts of DBSCAN algorithm has been illustrated in Fig. 1.

As illustrated in Fig. 1, the red data points are core points and the black points are border points. The circles are outliers which denotes the noise within the dataset and the arrows between the core points indicates direct density reachability. All core points and border points are density connected as they are reachable from the starting core points from the density perspective. Instead, the noise points are not reachable according to density and hence are considered outliers. All the outliers failed to contain the sufficient number of points to form a cluster and they do not fit any clusters. The time-complexity for running a DBSCAN algorithm is $O(n^2)$.

![Figure 1. Illustration of the DBSCAN algorithm clustering concept.](Image)

B. Visualization Using T-SNE

To visualize multi-dimensional dataset, linear compression techniques such as Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) are being widely used in previous literature. They often well perform with low-dimensional dataset with fair quality of preserving the original data structure. However, one major drawback of these conventional methods is that they often fail to generate high-quality low-dimensional representation of the original high-dimensional dataset [22]. To address this issue, t-distributed Stochastic Neighbor Embedding (t-SNE) proposed by Van der Maaten & Hinton [23] offered a more reliable solution and it has been widely applied in many fields in engineering, natural language processing, and cancer analysis.

Given an input data matrix in the size of $N \times k$ where $N$ denotes the total number of data points and $k$ denotes the dimensions, the pairwise distances between any data points are transformed into probability distributions that represents the neighborhoods in an intuitive method. First, for any pair of data points $x_i$ and $x_j$ in the high-dimensional dataset, t-SNE computes the similarity $p_{ij}$ between them which is based the conditional probability that $x_i$ picks $x_j$ as the neighbor point given the condition that $x_i$ is selected as the center point. The conditional probability $p_{ij}$ is proportional to the probability density function of a high-dimensional Gaussian distribution $P$ as expressed in (2):

$$p_{ij} = \frac{\exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma^2)}$$

(2)

where $\sigma$ denotes the standard deviation from the Gaussian distribution centered around the point $x_i$. Hence, the similarity $p_{ij}$ between any two points $x_i$ and $x_j$ can be expressed in (3):

$$p_{ij} = \frac{p_{ij} \times p_{ij}}{2N}$$

(3)
Subsequently, the t-SNE algorithm reconstructs another probability distribution \( Q \) that projects \( P \) from high-dimensional to low-dimensional space (e.g., 2-D or 3-D) in order to provide straightforward visualization. Hence, \( Q \) is the projection of \( P \) in the low-dimensional space. Here, unlike \( P \), \( Q \) is modeled using a heavy-tailed Student’s t-distribution and is not parametrized with a variable neighborhoods density (e.g., no \( \sigma_i \)). The similarity between the projected two points in low-dimensional space can be computed in (4):

\[
KL(P\|Q) = \sum_{i,j}^{} p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}
\]

C. PROTOTYPE SELECTION

According to the principles of constructing information granules, the granular descriptor of a certain region should be justifiable based on its position and size [13]. When locating a information granule’s position, the prototype (cluster center) represents a group of data points that are considered the member of the constructed granular. Hence, the prototype locates an information granule and the size factor decides its description range. The computational details can be summarized as follows.

First, a Fuzzy-C-means (FCM) [25] algorithm has been applied to partition the clusters generated by DBSCAN algorithm into several sub-clusters. The FCM is a popular fuzzy clustering algorithm that partition the given dataset into \( c \) fuzzy clusters while outputs a partition matrix \( U \) and \( c \) prototypes (cluster centers). The partition matrix is expressed in (5) and the entries of the elements in \( U \) are computed by (6) as follows:

\[
U = \begin{bmatrix}
u_{i1} & \cdots & u_{iN} \\
\vdots & \ddots & \vdots \\
u_{c1} & \cdots & u_{cN}
\end{bmatrix}
\]

\[
u_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \| x_i - y_{jk} \| \right)^{m}}
\]

where \( U \) is the partition matrix representing each point’s membership to all clusters generated; \( x_i \) denotes a randomly selected data point from the dataset; \( u_{ik} \) is the membership score for \( x_i \) with respect to the \( i \)-th sub-cluster which is in the range of \([0,1]\); \( m \) is a fuzzification coefficient; and the summation of the membership score to all the clusters should be 1 intuitively as \( \sum_{i=1}^{N} u_{ik} = 1 \). Then, the prototypes (cluster center) namely \( v_i \) of the \( i \)-th sub-cluster can be computed in (7):

\[
v_i = \frac{\sum_{k=1}^{N} u_{ik} x_k}{\sum_{k=1}^{N} u_{ik}}
\]

Next, several iterations are performed to minimize the value of the loss function (see Eq.(8)). The iteration are set to stop at a given termination condition and a final partition matrix plus the prototypes are obtained from the computational steps. The loss function has been expressed in (8) respectively.

\[
Loss = \sum_{k=1}^{N} \sum_{j=1}^{c} u_{ik}^{m} \| x_k - v_i \|^2
\]

D. MEASUREMENT MATRICES

With the computed prototypes of the sub-clusters, the information granule construction also requires the determination of the granular boundaries for decision-making tasks. According to Ouyang et al. [12], two temporary templates points from the sub-clusters are selected to indicate the upper and lower boundaries for the underlying information granule.

Here, we initialize with two randomly selected points inside each sub-cluster and name them as \( x_i^u \) for the upper bound and \( x_i^l \) for the lower bound. Then, we can apply two measurement matrices namely coverage and specificity to evaluate the quality of the initial constructed information granules. The coverage is a reflection of an information granule’s capacity to cover data points [26]. It computes the summation of all membership scores for the data covered by the lower or upper bound. The coverage for the initialized upper bound can be computed by (9) and the coverage for the initialized lower bound can be computed by (10) as follows:

\[
cov(x_i^u) = \sum_{i \in | x_i - v_i > 0 \land x_i < x_i^u |} u_{ik}
\]

\[
cov(x_i^l) = \sum_{i \in | x_i - v_i < 0 \land x_i > x_i^l |} u_{ik}
\]

where \( u_{ik} \) is the membership score of the data point \( x_k \) with respect to the \( i \)-th sub-cluster; \( v_i \) is the computed prototype for the \( i \)-th sub-cluster from the FCM algorithm; and \( x_i^u \) and \( x_i^l \) are randomly initialized points indicating the upper and lower boundaries presented above.

Meanwhile, specificity indicates the precision of the constructed information granule. Higher value of specificity reflects smaller size of a granule intuitively. The computation of specificity for the upper and lower bounds are defined in (11) and (12) as follows:

\[
sp(x_i^u) = 1 - \frac{\| x_i^u - v_i \|}{\| x_{max} - v_i \|}
\]

\[
sp(x_i^l) = 1 - \frac{\| v_i - x_{min} \|}{\| x_i^l - x_{min} \|}
\]

where \( \| \cdot \| \) is the distance measure between the data points; \( v_i \) is the computed prototype from the FCM algorithm; \( x_i^u \) and \( x_i^l \) are randomly initialized points indicating the upper and lower bounds; \( x_{max} \) and \( x_{min} \) are the maximum points encountered within the generated data sub-clusters for which the information granule is developed.

A high-quality information granule is expected to have both high coverage and high specificity. However, in practice, the increase of coverage always associates with the decrease of specificity. A trade-off is expected to balancing the two conflicting requirements by sweeping the parameters of bounds localization. In this research, the curve reflecting the product of both coverage and specificity is proposed in this research. The area under the coverage-specificity curve (AUC) is regarded as the global evaluation criteria for the information granules constructed as expressed in (13):

\[
AUC = \int_{0}^{1} cov(x) sp(x) dx
\]

where \( x \) is the data point to define boundaries of the granule.

To obtain the optimal information granules, the cost function for the granule construction process is expressed as follows in Eq. (14):
where \( x \) is the data point to define boundaries of the granule; and \( PI \) is the performance indicator expressed as the product of coverage and specificity. The optimization process aims to increase the value of \( PI \) iteratively and the value of AUC can be maximized.

**E. INFORMATION GRANULES & RISK MEASURES**

By repeating the process as described in Section II-D, the optimal solution for the location and size of the information granule \( C_i \) can be derived. Hence, we can express the fuzzy rule-based information granules as:

\[
R_i: \text{If } x_k \text{ is in the neighborhood of } v_i, \text{ THEN } y_k \text{ should be classified as risk group } C_i \quad (15)
\]

where \( x_k \) is the candidate data point; \( v_i \) is the selected prototype; \( y_k \) is the risk label for the data point \( x_k \). The definition of “neighborhood” refers to the range of \([x^l_i, x^u_i]\).

In order to increase the interpretability of the fuzzy rule, two risk measures namely Value-at-Risk (VaR) and Conditional-Value-at-Risk (CVaR) are introduced in this research for interpreting the landslide risks in different risk attributes. Suppose one attribute of all covered data points within an information granule follows a certain distribution as \( x \), then we can fit the distribution can obtain its probability density function (PDF) via the maximum likelihood approach. The cumulative density function (CDF) can be obtained as follows:

\[
F(x) = \int_{-\infty}^{x} f(x)dx \quad (16)
\]

where \( f(x) \) and \( F(x) \) are the PDF and CDF for the distribution of the attribute of all data points within the underlying information granular. Since the parametric distributions are selected in this study, both the PDF and CDF are differentiable.

The VaR can be conceived as the inverse computation of the CDF with respect to the probability percentile \( \alpha \) as expressed in (17). The CVaR is the conditional expectation of the distribution when the attributes of the data points exceed the VaR threshold. The computation of CVaR is expressed in (18) as follows.

\[
VaR_\alpha(x) = F^{-1}_x(\alpha) \quad (17)
\]

\[
CVaR_\alpha(x) = VaR_\alpha(x) + \frac{\mathbb{E}[X|X>VaR_\alpha(x)]}{1-\alpha} \quad (18)
\]

**III. FIELD INVESTIGATION AND DATA COLLECTION**

For this research, the study area is located within Jiuzhaigou county, China. It is situated in northern Sichuan province with a study area of 545.19 \( \text{km}^2 \) between 33.03–33.35\(^\circ\)N latitude and 103.63–104.05\(^\circ\)E longitude. The whole area is in the mountainous region on the estern margin of the Tibetan Plateau. Overall, the whole region has complex geological conditions with strong tectonic movement which uplifted the entire region of Jiuzhaigou. Early slope failures have been frequently observed in the eastern region of Jiuzhaigou county.

On Aug 8th 2017, a devastating earthquake of magnitude 6.5 struck hit the whole Jiuzhaigou region and triggered a huge amount of co-seismic landslides. The majority of the earthquake-triggered landslides are shallow rock slides and rock falls. They seriously threaten the anthropogenic activities, local traffic, as well as tourist facilities of the region. Hence, it is essential to build landslide risk assessment model to classify the landslide risks for the safety of tourism and natural landform protection [27-28].

The dataset used in this study was collected by geological experts with thousand times of on-site investigations. In total of 4800 landslide occurrences are reported in the dataset and mapped on the satellite remote sensing images. According to field expertise, several landslide risk-related attributes are considered into this study. Several attributes including elevation, slope angle, relief amplitude, fault distance, area, volume, runout distance and vertical drop are measured by onsite visit and remote sensing maps. The name, unit, and range of the attributes are summarized in Table I as follows.

**IV. EXPERIMENTAL RESULTS**

**A. CLUSTERING ANALYSIS**

In this section, the clustering analysis is performed using the DBSCAN algorithm. The landslides dataset utilized in this study is firstly split into the training/testing data according to 70%/30% rule and bootstrap technique is applied over the...
training dataset to form several pairs of training and testing data sets according to [29]. In these pairs, the testing dataset is always the same but the training dataset is not independent across each other as re-sampling is applied to regenerate the training set in each bootstrap experiment. In total, we performed 200 bootstrap resampling experiments over the training set and tested the proposed framework on the testing dataset instead of one single split and test in order to avoid overfitting of model parameters. In total of 8 landslide attributes as described in Table I are utilized in this study. All experiments were conducted on a machine equipped with Intel(R) Core(TM) i5-4300U, 2.90 GHz, 16GB RAM computer. All codes are programmed on the MATLAB 2020a environment.

In the first experiment, we use the default setting of the scanning radius parameter ε=0.3 to perform the clustering analysis with different size of the attributes. Among all 8 risk-related attributes, we perform the clustering in an incremental manner to increase the number of utilized attributes K from 2, 4, 6, and 8 in this study. Here, with different K, the inclusion of attributes are in the same order regarding their geological importance from the field experts. For instance, when K=2, the two most important geological attributes as landslide area and runout distance are selected in the experiments. In the scenario of K=4, four attributes including landslide area, runout distance, total volume and vertical drop. When K=6, only slope angle and elevation are excluded from the selected attributes and K=8 refers to all attributes are included in the experiments. To test the effectiveness of the DBSCAN algorithm, we measured the running time with different size of training samples from 100 to 2800. The computational time measured in seconds for each trail has been illustrated in Fig. 2 as follows.

According to Fig. 2, the trail with 8 attributes are more efficient for clustering with DBSCAN algorithm. In comparison, the trail with only 2 attributes cost the largest amount of time to converge to the optimal solution. This indicates that the 8 attributes together contain important information which strengthens the partition of different risk groups. In comparison, datasets with only 2 attributes can only provide limited information and thus deteriorate the clustering performance intuitively.

In the second experiment, we performed clustering analysis with the training dataset and tried various levels of the scanning radius ε from 0.1 to 0.9 in an incremental way. To test the effectiveness of the DBSCAN algorithm, we measured the running time with different size of training samples from 100 to 2800. All 8 attributes are utilized in this experiment. The measured running time for various scanning radius parameter under different dataset size are displayed in Fig. 3 as follows.

In Fig. 3, due to the limit of space, the run-time curves for only four incremental radius parameters are displayed. The small radius value (ε=0.2) consumes the largest amount of computational time. This can be mainly attribute to the nature of low density among landslides dataset which indicates the sparsity between the different landslide points with respect to the distance measures with all 8 attributes. Hence, it takes more time to converge as the scanning process cost more time to cover all points within a cluster. On the contrary, ε=0.8 is the most efficient parameter for clustering with DBSCAN algorithm.

Furthermore, we also checked the summation of inner cluster distances with various ε values. It is discovered that ε=0.8 provides the smallest summation of inner cluster distances and hence further demonstrates its effectiveness in clustering the underlying landslide dataset.

**B. VISUALIZATION IN THE EMBEDDING SPACE**

In this subsection, the correctness of the DBSCAN algorithm has also been evaluated. In total of 3 subclusters are generated from the DBSCAN algorithm with multiple experiments. We input the testing dataset into the pre-trained clusters and visualize them using the t-SNE algorithm. Here, we use the best performing scanning radius parameter setting as ε=0.8 and we randomly sampled N (N=200, 400, 600, and 800) number of testing data points to input into the t-SNE algorithm for demonstration purposes. Meanwhile, we also visualize the clustering performance by using different number of attributes K. The number of attributes K were increased from 2, 4, 6, and 8. All visualizations of the testing dataset were projected in to the 2-D low-dimensional space by t-SNE and the visualization outcome has been presented in Fig. 4 as follows.

In Fig. 4, we displayed the clustering outcome of the testing dataset with different number of attributes and different sampling size of the testing points. The meaning of different K values denotes various subset of selected attributes according to their importance from the perspective of geological experts and the details are described in Section IV.A above. According to the expertise provided from geologists, we can label them into three risk levels as low risk (light blue), medium risk (dark blue), and high risk (red). First, the testing data points reflect homogenous patterns as they reflect similar distribution over the t-SNE embedding space for all values of N. As K increases, the more
distinctions between the clusters are displayed. This indicates all the 8 features are the most informative scenario for constructing information granules for landslide risk assessment.

FIGURE 3. DBSCAN clustering with different radius parameter setting.

C. CONSTRUCTING INFORMATION-GRANULES

With the clusters generated from the DBSCAN algorithm, the prototypes are computed using the method introduced in Section II-C. Here, each selected prototype point serves as the most “representative” point of a group of landslide data points and the information granules are constructed using the prototype as cluster centers. Then, the decision-boundaries can be computed using these prototypes as centers with appropriate distance measures.

To obtain the fuzzy rules of for the constructed information granules, various distance measures are tested in this research. Here, we utilized the 2-D low-dimensional embeddings from t-SNE to display the prototypes of the granules and the geometry of the granular descriptors using various definitions of coverage by three distances including Euclidean, Manhattan, and Tchebychev [30]. To generate Fig. 5, the $\varepsilon$ was set as 0.8; $N$ was set as 800; $K$ was set as 8; and $c$ was set as 3 as it presents the total of 3 clusters generated.

As displayed in Fig. 5, (a) is the original t-SNE visualization of a landslide dataset on the 2-D embedding space with 800 sampled data and all 8 features; (b) is the granules constructed using Euclidean distance with various coverages; (c) is the granules constructed using Manhattan distance with various coverages; and (d) is the granules constructed using Tchebychev distance with various coverages. The center black dots denote the computed prototypes for each cluster according to Section II-C. It is obvious that high-specificity (dash-lines) would lead to low coverage and the risk measures are inadequate to describe the characteristic of the risks in the underlying landslide group. On the contrary, higher coverage with less compactness is a more suitable solution for constructing the decision-boundaries of the information granules in this study. In this study, the Euclidean distance is selected as the distance metric in the landslide dataset.

FIGURE 5. Generating rule-based information granules using three distance measures.

For global evaluation of the granular quality, two metrics including coverage and specificity are computed following the description in Section II-D. In this research, we compute the metrics using both upper and lower bound points simultaneously and average their values. Then, the coordinates of various coverage-specificity curves are obtained and can be plotted for the experiments tested. We have examined various levels scanning radius parameters in combination with different number of attributes in the testing dataset. All experimental results for this test have been presented in Table II and illustrated in Fig. 6 below.

As illustrated in Fig. 6, the coordinates of coverage-specificity computation are displayed and the area under the coverage-specificity curves (AUCs) can be measured as the global performance measurement. Each row displays the curve under $K$ number of attributes and each column indicates the clusters generated by DBSCAN with different scanning radius parameters. The last row in Fig. 6 has the largest area converged by the curves and thus indicate improved performance with more attributes included in the dataset.

In Table II, the performance measures including coverage, specificity and AUC are computed for all possible combinations of the number of input attributes $K$ and the radius factor $\varepsilon$. For most of the pairs, the values of specificity
are smaller than coverage which indicates that the constructed information granules input more emphasis on covering homogenous points for each sub-cluster. With respect to the number of attributes $K$, it is obvious that all performance measures are increased. Intuitively, this can be attributed to the richer information provided for the granules from the added attributes which is valuable for the improvement of both coverage and specificity. Meanwhile, there’s no obvious pattern between the performance measures and the radius parameter $\varepsilon$.

In Table II, the top performance with the highest AUC value was achieved with $K=8$ and $\varepsilon=0.8$. It further demonstrates the number of attributes are key impact factors for the quality of constructed information granules. In sum, the number of attributes included in the dataset is the key determinant factor impacting the quality of constructed information granules. Meanwhile, there’s no obvious pattern between the performance measures and the radius parameter $\varepsilon$.

In Table II, the top performance with the highest AUC value was achieved with $K=8$ and $\varepsilon=0.8$. It further demonstrates the number of attributes are key impact factors for the quality of constructed information granules. In sum, the number of attributes included in the dataset is the key determinant factor impacting the quality of constructed information granules. In sum, the number of attributes included in the dataset is the key determinant factor impacting the quality of constructed information granules. With all attributes included (when $K=8$), the AUC values do not significantly increased with respect to the values $\varepsilon$ in DBSCAN. This indicates there’s no significant difference among the clusters generated by DBSCAN algorithm with different scanning radius parameter values.

FIGURE 6. Coverage and specificity curves for all the scenarios.

D. COMPARATIVE ANALYSIS

To assess the performance of the proposed approach using DBSCAN generated clusters for information granule construction, we also performed comparative analysis on the same dataset using traditional clustering algorithms such as k-mean, k-medoids, as well as FCM. Using the same landslide dataset, we measure the quality of information granules constructed with coverage, specificity, and AUC. Also, the computational time with respect to training the clustering algorithms has been considered. The comparative analysis for all algorithms has been summarized in Table III in detail as follows.

Table III shows the comparison of granular quality between the DBSCAN-based granules and granules constructed based on other clustering algorithms. Here, the k-Mean, k-Medoids, and FCM algorithm are selected as benchmarking algorithms for comparison. Similar to Table
II, the specificity of all granules constructed have higher coverage than specificity which indicates the granules focus more on covering homogenous data points into sub-clusters than differentia them into others. The DBSCAN-based granules outperforms in specificity, coverage, and AUC which indicates that the proposed DBSCAN-based information granules dominates all the compared clustering algorithms. The computational time measured in seconds has also been provided in the last row of Table III. Since DBSCAN algorithm is a density based clustering algorithm instead of measuring inner sub-cluster distances, it requires much more computational time intuitively. Although DBSCAN requires more computational time compared with k-mean and k-medoids algorithms, there exists a significant improvement with respect to the specificity, coverage and AUC values. Thus, the DBSCAN is more suitable for providing clusters to construct information granules in this research.

E. RISK MEASURES & HAZARD MAPPING

The granular descriptors are constructed and optimized according to the work presented in Section IV.A-C. In order to increase the interpretability of the information granules, the two risk measures [31] including VaR (see Eq.(17)) and CVaR (see Eq.(18)) are computed in this study for all attributes. To indicate both the lower and upper bounds of the granules, the 5% and 95% percentiles of VaR and CVaR are computed for the three most informative attributes including area, runout distance, and volume. These three attributes are selected by geologists considered as top factors evaluating landslide risks in most of the field investigations. The VaRs and CVaRs are computed for the three landslide risk groups from low to high and are summarized in Table IV in detail. The risk measures can provide direct interpretation for the rules generated from the information granules to guide the field engineers to classify risk levels of incoming landslide in the future. The box-plots indicating the constructed interval-based granules are also illustrated in Fig. 7 respectively.

| Algorithm | k-Mean | k-Medoids | FCM | DBSCAN |
|-----------|--------|-----------|-----|--------|
| Mean      | 0.71   | 0.61      | 0.75 | 0.87   |
| 95% CI    | (0.62,0.80) | (0.63,0.79) | (0.67,0.83) | (0.80,0.94) |

TABLE II
Summary of the global performance measures of information granules

| Metric   | ε = 0.2 | ε = 0.4 | ε = 0.6 | ε = 0.8 |
|----------|---------|---------|---------|---------|
|          | Mean    | 95% CI  | Mean    | 95% CI  | Mean    | 95% CI  | Mean    | 95% CI  |
| Specificity | 0.63  | (0.59,0.66) | 0.56  | (0.51,0.61) | 0.50  | (0.44,0.56) | 0.69  | (0.60,0.78) |
| Coverage  | 0.90    | (0.89,0.91) | 0.84  | (0.82,0.86) | 0.84  | (0.78,0.90) | 0.76  | (0.70,0.82) |
| AUC       | 0.66    | (0.62,0.70) | 0.69  | (0.64,0.74) | 0.78  | (0.68,0.88) | 0.68  | (0.59,0.77) |
| Specificity | 0.60  | (0.58,0.62) | 0.68  | (0.66,0.70) | 0.59  | (0.58,0.60) | 0.63  | (0.61,0.65) |
| Coverage  | 0.85    | (0.83,0.87) | 0.92  | (0.91,0.93) | 0.85  | (0.84,0.86) | 0.91  | (0.85,0.97) |
| AUC       | 0.79    | (0.70,0.88) | 0.74  | (0.65,0.83) | 0.68  | (0.65,0.71) | 0.73  | (0.72,0.74) |
| Specificity | 0.71  | (0.68,0.74) | 0.78  | (0.76,0.80) | 0.74  | (0.68,0.80) | 0.79  | (0.72,0.86) |
| Coverage  | 0.84    | (0.79,0.89) | 0.87  | (0.80,0.94) | 0.89  | (0.88,0.90) | 0.81  | (0.71,0.91) |
| AUC       | 0.79    | (0.77,0.81) | 0.84  | (0.82,0.86) | 0.86  | (0.85,0.87) | 0.82  | (0.73,0.91) |
| Specificity | 0.72  | (0.68,0.76) | 0.87  | (0.85,0.89) | 0.59  | (0.54,0.64) | 0.81  | (0.74,0.88) |
| Coverage  | 0.88    | (0.82,0.94) | 0.70  | (0.61,0.79) | 0.89  | (0.81,0.97) | 0.81  | (0.75,0.87) |
| AUC       | 0.83    | (0.75,0.91) | 0.84  | (0.81,0.87) | 0.88  | (0.84,0.92) | 0.89  | (0.85,0.93) |

TABLE III
Comparative analysis of granular quality between DBSCAN clusters and others

| Metric   | k-Mean | k-Medoids | FCM | DBSCAN |
|----------|--------|-----------|-----|--------|
| Mean     | 0.71   | 0.61      | 0.75 | 0.87   |
| 95% CI   | (0.62,0.80) | (0.63,0.79) | (0.67,0.83) | (0.80,0.94) |

E. RISK MEASURES & HAZARD MAPPING

The granular descriptors are constructed and optimized according to the work presented in Section IV.A-C. In order to increase the interpretability of the information granules, the two risk measures [31] including VaR (see Eq.(17)) and CVaR (see Eq.(18)) are computed in this study for all attributes. To indicate both the lower and upper bounds of the granules, the 5% and 95% percentiles of VaR and CVaR are computed for the three most informative attributes including area, runout distance, and volume. These three attributes are selected by geologists considered as top factors evaluating landslide risks in most of the field investigations. The VaRs and CVaRs are computed for the three landslide risk groups from low to high and are summarized in Table IV in detail. The risk measures can provide direct interpretation for the rules generated from the information granules to guide the field engineers to classify risk levels of incoming landslide in the future. The box-plots indicating the constructed interval-based granules are also illustrated in Fig. 7 respectively.

| Algorithm | k-Mean | k-Medoids | FCM | DBSCAN |
|-----------|--------|-----------|-----|--------|
| Mean     | 0.71   | 0.61      | 0.75 | 0.87   |
| 95% CI   | (0.62,0.80) | (0.63,0.79) | (0.67,0.83) | (0.80,0.94) |
In Table IV, it presents the risk measures including the upper/lower bounds for the top three attributes including area, runout and volume. Here, the upper bounds ($VaR_{.95}$) and lower bounds ($VaR_{.05}$) for the risk measure $VaR$s are computed indicating the thresholds for the information granules. Meanwhile, the upper bounds ($CVaR_{.95}$) and lower bounds ($CVaR_{.05}$) for the risk measure $CVaR$s are also computed. This provides a further conservative thresholds for classifying risk levels for each group.

In Figure 7 above, the box-plots for the constructed information granules in the aspect of area, runout and volume are illustrated. For the risk levels, we name them from high to low as C1, C2, and C3 as instructed by the experts. Then, the information granular characteristics such as mean, upper bounds ($VaR_{.95}$), and lower bounds ($VaR_{.05}$) are all plotted in Figure 7. This provides a directly visual interpretation of the information granules constructed which is valuable for the risk classification tasks for field engineers.

Meanwhile, mapping the risk labels (low, medium, high) on the landslide inventory map is crucial for the geologists to identify landslide hazard and perform risk assessment. As the risk labels are computed from the pre-constructed information-granules, all risk indices can be imported into ArcGIS environment to produce the landslide risk map. For better visualization, the risk labels including low risk (green color), medium risk (yellow color), and high risk (red color) are projected on to the inventory map according to their historic occurrence and location in Jiuzhaigou county. Fig. 8 below shows the landslide risk map of all the landslide data collected from our case study area. It can be observed that most of the high-risk landslides have wide area, large runout distance, and large slide volume which confirms the knowledge provided by the geology experts. Hence, the risk labels provided from information-granules on the inventory map can visualize high-risk landslide-prone zones effectively and assist geologists to prevent and reduce potential damages.

### V. DISCUSSION

In this research, 8 factors were selected as the attributes for the development of information granules to classify the risk levels of earthquake-triggered landslides. The selection of these factors were mainly based on the experts opinions. In practice, the landslides conditioning factors may vary with respect to the study area and its geographical locations. Hence, the set of attributes selected in this study can be applied particular to the type of earthquake-triggered landslides used in this research. Here, we also compared with some previous related research in the aspect of attributes selected for the risk analytics of landslides as summarized in Table V as follows:

In Table V, the selected attributes of landslide characteristics in this study are compared with the ones selected in previous literature. It is obvious that the geomorphological factors including elevation, slope angle, runout, area, and others are always considered important factors for landslides risk assessment.

In addition, the ranking of importance of each attribute within the attribute set were generally determined by
geologists using their field expertise. However, the human determination of the ranking order can be different from the computational results provided from the algorithms.

**TABLE V**

| Literature                     | Attributes                                      |
|-------------------------------|-------------------------------------------------|
| Fan et al. (2014) [32]        | Area, volume, runout relief, distance to river, and H/L ratio |
| Lee & Pradhan. (2007) [33]    | Slope aspect, curvature, distance from drainage, soil type, area, precipitation |
| Suzen and Kaya (2011) [34]    | Elevation, surface roughness, slope length, slope aspect, slope angle, slope curvature |
| This study                    | Elevation, slope angle, relief amplitude, fault distance, area, volume, runout, vertical drop |

The Relief algorithm an instance-based learning algorithm for ranking the importance of candidate attributes in a supervised manner. Given a training dataset, a sample size, and a threshold, it identifies the attributes that are statistically relevant to the target output. In Table IV below, we compared the ranking of all 8 attributes provided by experts and Relief algorithm which is a supervised attribute ranking algorithm.

**TABLE VI**

| Attribute Name | Experts | Relief Algorithm |
|----------------|---------|------------------|
| Area           | 1       | 2                |
| Runout         | 2       | 1                |
| Volume         | 3       | 3                |
| Vertical drop  | 4       | 6                |
| Relief amplitude | 5   | 5                |
| Fault distance | 6       | 8                |
| Slope angle    | 7       | 4                |
| Elevation      | 8       | 7                |

According to Table VI, the order of ranking between the experts and Relief algorithm is slightly different. Especially when considering the top three attributes including area, runout, and volume, there’s very slight difference in the importance ranking outcome provided by either experts or Relief algorithm. However, there also exits difference in the ranking order in other attributes except the top three ones. This can be attributed to the fact that the experts opinion are more based on their field experience. However, the ranking from Relief algorithm are mainly based on the statistical correlation.

**VI. CONCLUSION**

This paper proposes a novel granular computation approach to discover the fundamental structure within the landslide risk dataset. The DBSCAN firstly clustered the multiple landslide data points into sub-clusters for generating rule-based granules. Graphs are also generated using t-SNE to project the landslide dataset in subclusters into 2-dimensional embedding space for data visualization. Then, we construct information-granules by computing prototypes for each cluster and then set up decision boundaries by intervals. The rule-based information-granules are constructed to interpret the decision-making process for risk classification. We also optimize the granular structures by considering the coverage, specificity, and AUC and obtained the optimal granules for classifying landslide risks into three levels. Two quantitative risk measures including the VaR and CVar are both computed to indicate the decision boundaries for the risky landslides. Comparative analysis against benchmarking clustering algorithms has been performed. Computational results over the testing dataset demonstrates that the proposed approach outperforms the benchmark clustering algorithms with respect to the quality of constructed information granules.

Case study on landslides from Jiuzhaigou County, China demonstrates that the proposed approach to construct information granules can effectively generate interpretable decision-rules to group the landslides with higher risks compared with others. The DBSCAN algorithm can effectively generate subclusters and the rule-based information granules are constructed using these subclusters. The selected prototypes in the center of each cluster represents the geometric center of constructed information granule with respect to landslide risk. Risk maps of landslide in the case study area has been produced based on the obtained risk labels and the original landslide inventory map. In the future, in-detail analysis will be performed considering the scenarios that data points containing missing values in certain attributes. Thus, it will provide us more in-depth insights regarding the risk patterns within the landslide dataset.
FIGURE 8. Risk mapping of landslides classified by rule-based information granules.

REFERENCES

[1] J. Gasmo, H. Rahardjo, and E. Leong, "Infiltration effects on stability of a residual soil slope," Computers & Geotechnics, vol. 26, no. 2, pp. 145-165, 2000.

[2] J. Boyd, J. Chambers, P. Wilkinson, et al., “linked geomorphological and geophysical modelling methodology applied to an active landslide,” Landslides, to be published. DOI: 10.1007/s10346-021-01666-w

[3] D. Kirschbaum, T. Stanley, and S. Yatheendradas, "Modeling landslide susceptibility over large regions with fuzzy overlay," Landslides, vol. 13, no. 3, pp. 485-496, 2016.

[4] N. T. H. Van, P. Van Son, N. H. Khanh, & L. T. Binh, “Landslide susceptibility mapping by combining the analytical hierarchy process and weighted linear combination methods: a case study in the upper Lo River catchment (Vietnam),” Landslides, vol. 13, no. 5, pp. 1285-1301, 2016.

[5] H. A. Nefeslioglu, E. A. Sezer, C. Gokceoglu, & Z. Ayas, “A modified analytical hierarchy process (M-AHP) approach for decision support systems in natural hazard assessments,” Computers & Geosciences, vol. 59, pp. 1-8, 2013.

[6] B. Ahmed, “Landslide susceptibility mapping using multicriteria evaluation techniques in Chittagong Metropolitan Area, Bangladesh,” Landslides, vol. 12, no. 6, pp. 1077-1095, 2015.

[7] O. F. Althuwaynee, B. Pradhan, H. J. Park, & J. H. Lee, “A novel ensemble decision tree-based CHi-squared Automatic Interaction Detection (CHAID) and multivariate logistic regression models in landslide susceptibility mapping,” Landslides, vol. 11, no. 6, pp. 1063-1078, 2014.

[8] F. Huang, C. Yao, W. Liu, Y. Li, & X. Liu, “Landslide susceptibility assessment in the Nantian area of China: a comparison of frequency ratio model and support vector machine. Geomatics,” GEOMAT NAT HAZ, vol. 9, no. 1, pp. 919-938, 2018.

[9] A. M. Youssif, H. R. Pourghasemi, Z. S. Pourtaghi, & M. M. Al-Katheeri, “Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia,” Landslides, vol. 13, no. 5, pp. 839-856, 2016.

[10] P. V. Gorsevski, M. K. Brown, K. Panter, C. M. Onasch, A. Simic, & J. Snyder, “Landslide detection and susceptibility mapping using LiDAR and an artificial neural network
approach: a case study in the Cuyahoga Valley National Park, Ohio,” Landslides, vol. 13, no. 3, pp. 467-484, 2016.

[11] F. Huang, J. Zhang, C. Zhou, Y. Wang, J. Huang, & L. Zhu, “A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction,” Landslides, vol. 17, no. 1, pp. 217-229, 2020.

[12] T. Ouyang, W. Pedrycz, & N. J. Pizzi, “Record linkage based on a three-way decision with the use of granular descriptors,” Expert Systems with Applications, vol. 122, pp. 16-26, 2019.

[13] T. Ouyang, W. Pedrycz, O. F. Reyes-Galaviz, & N. J. Pizzi, “Granular description of data structures: A two-phase design,” IEEE Trans Cybern, vol. 51, no. 4, pp. 1902-1912, 2019.

[14] T. Ouyang, W. Pedrycz, & N. J. Pizzi, “Rule-based modeling with DBSCAN-based information granules,” IEEE Trans Cybern, to be published. DOI: 10.1109/TCYB.2019.2902603.

[15] X. Zhang, H. Gou, Z. Lv, & D. Miao, “Double-quantitative distance measurement and classification learning based on the tri-level granular structure of neighborhood system,” KNOWL-BASED SYST, vol. 217, pp. 106799, 6 April 2021.

[16] Y. Chen, &D. Miao, “Granular regression with a gradient descent method,” NFORM SCIENCES, vol. 537, pp. 246-260, 2020.

[17] W. Froelich, & W. Pedrycz, “Fuzzy cognitive maps in the modeling of granular time series,” KNOWL-BASED SYST, vol. 115, pp. 110-122, 2017.

[18] H. Guo, W. Pedrycz, & X. Liu, “Hidden Markov models based approaches to long-term prediction for granular time series,” IEEE T FUZZY SYST, vol. 26, no. 5, pp. 2807-2817, 2018.

[19] L. Ester, D. Leite, F. Viola, & D. Bonacorsi, “Comparison of evolving granular classifiers applied to anomaly detection for predictive maintenance in computing centers,” in 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS) (pp. 1-8). IEEE.

[20] M. Ester, H. Kriegel, J. Sander and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” International Conference on Knowledge Discovery in Databases and Data Mining (KDD-96), Portland, August 1996, pp. 226-231.

[21] E. Schubert, Sander, M. Ester, H. P. Kriegel, & X. Xu, “DBSCAN revisited, revisited: why and how you should (still) use D J.BSCAN,” ACM Transactions on Database Systems (TODS), vol. 42, no. 3, pp. 1-21, 2017.

[22] A. Platzer, “Visualization of SNPs with t-SNE,” PloS one, vol. 8, no. 2, pp. e56883, 2013.

[23] L. Van der Maaten, & G. Hinton, “Visualizing data using t-SNE,” Journal of machine learning research, vol. 9, no. 11, pp. 2579–2605, 2008.

[24] A. Chatzimparmpas, R. M. Martins, & A. Kerren, “t-viSNE: Interactive Assessment and Interpretation of t-SNE Projections,” IEEE T VIS COMPUT GR, vol. 26, no. 8, pp. 2696-2714, 2020.

[25] S. B. Roh, S. K Oh, & W. Pedrycz, “Identification of black plastics based on fuzzy RBF neural networks: Focused on data preprocessing techniques through Fourier transform infrared radiation,” IEEE T IND INFORM, vol. 15, no. 5, pp. 1802-1813, 2017.

[26] S. Wang, & W. Pedrycz, “Data-driven adaptive probabilistic robust optimization using information granulation,” IEEE T CYBERNETICS, vol. 48, no. 2, pp. 450-462, 2016.

[27] H. Li, Q. Xu, Y. He, & J. Deng, “Prediction of landslide displacement with an ensemble-based extreme learning machine and copula models,” Landslides, vol. 15, no. 10, pp. 2047-2059, 2018.

[28] H. Li, Q. Xu, Y. He, X. Fan, S. Li, “Modeling and predicting reservoir landslide displacement with deep belief network and EWMA control charts: a case study in Three Gorges Reservoir,” Landslides, vol. 17, no. 3, pp. 693-707, 2020.

[29] Harrell Jr, F. E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. springer. 2015.

[30] T. Ouyang, Y. He, & H. Huang, “Monitoring Wind Turbines' Unhealthy Status: A Data-Driven Approach,” IEEE TETCI, vol. 3, no. 2, pp. 163-172, 2018.

[31] Q. Xu, H. Li, Y. He, F. Liu, & D. Peng, “Comparison of data-driven models of loess landslide runout distance estimation,” B ENG GEOL ENVIRON, vol. 78, no. 2, pp. 1281-1294, 2019.

[32] Fan, X., Rossiter, D. G., van Westen, C. J., Xu, Q., & Görum, T. (2014). Empirical prediction of coseismic landslide dam formation. Earth surface processes and landforms, 39(14), 1913-1926.

[33] Lee, S., & Pradhan, B. (2007). Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides, 4(1), 33-41.

[34] Sützen, M. L., & Kaya, B. Ş. (2012). Evaluation of environmental parameters in logistic regression models for landslide susceptibility mapping. International Journal of Digital Earth, 5(4), 338-355.