Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2451~2461
e-ISSN: 2548-6861

Implementation of HDBSCAN and Bayesian Optimization for Clustering Flood-Affected Regions in Indonesia

Nasywa Azzah Nabila [1]*, Aviolla Terza Damaliana [2], Shindi Shella May Wara [3]
[1,2,3] Data Science, Universitas Pembangunan Negeri "Veteran" Jawa Timur
22083010051@student.upnjatim.ac.id [1], aviolla.terza.sada@upnjatim.ac.id [2], shindi.shella.fasilkom@upnjatim.ac.id [3]

Article Info

Article history:

Received 2026-04-06
Revised 2026-05-09
Accepted 2026-05-25

Keyword:

Bayesian Optimization,
Clustering,
DBCV,
Floods,
HDBSCAN.

ABSTRACT

Floods are among the most frequent natural disasters in Indonesia, with thousands of events causing significant impacts on infrastructure damage and human lives. The substantial increase in the number of victims and flood-related damages in 2024 indicates that flood disaster mitigation efforts in Indonesia remain suboptimal. Consequently, a clustering-based analytical approach is required to understand patterns of flood impact across provinces. This study aims to cluster provinces in Indonesia based on flood impact indicators using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) method with Bayesian Optimization to obtain optimal hyperparameters. This study comprises several stages, including data collection, data standardization, statistical tests, data reduction, hyperparameter optimization, HDBSCAN algorithm, model evaluation, and analysis of clustering results. The results show that HDBSCAN with Bayesian Optimization yields a well-separated cluster structure with a DBCV value of 0.515. The clustering results consist of three primary clusters and one noise cluster. Cluster 0 (High Displacement & Inundation) comprising 5 provinces, Cluster 1 (High Fatality & Structural Damage) comprising 4 provinces, Cluster 2 (Low Impact) comprising 21 provinces, and the noise cluster comprising 8 provinces. These findings are intended to provide a foundation for the government to formulate targeted flood mitigation strategies tailored to the flood impact characteristics of each province.

I. INTRODUCTION

Floods are among the most frequent natural disasters in Indonesia, with thousands of events causing significant impacts on infrastructure damage and human lives. In 2023, the National Disaster Management Agency recorded 1,255 flood events across all provinces in Indonesia, affecting 3,860,136 individuals and causing 1,083,004 instances of damage [1]. These figures increased in 2024, with 1,420 flood events affecting 6,391,056 individuals and causing 2,005,567 instances of damage [2]. Notably, floods ranked as the most dominant natural disaster in Indonesia in 2024, highlighting their significant prevalence compared to other types of disasters. The substantial increase in the number of affected individuals and damages compared to the previous year indicates that flood mitigation efforts in Indonesia remain suboptimal. Therefore, a more comprehensive analytical approach is required to understand the characteristics of flood impacts across regions and support more targeted mitigation strategies.

Clustering is an effective approach for identifying patterns of flood impacts and grouping provinces in Indonesia based on similar impact characteristics. In the context of inter-provincial flood impact analysis, several previous studies have applied common clustering methods, such as K-Means [3], K-Medoids [4], and Fuzzy C-Means [5]. However, these methods have limitations in handling datasets with varying densities and the presence of noise, which may reduce clustering quality [6]. In practice, disaster-related data, particularly flood impact indicators across provinces are generally heterogeneous, characterized by substantial variations in impact severity among regions as well as the presence of several provinces with extreme impact characteristics. These conditions pose

a risk for partition-based and fuzzy clustering methods to disproportionately distort cluster boundaries, leading to less representative outcomes. To overcome these limitations, this study employs the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). A previous study [7] demonstrated that HDBSCAN can overcome these limitations through a hierarchical density-based approach that effectively clusters data with complex density structures while automatically identifying noise. Moreover, HDBSCAN is capable of automatically identifying the optimal number of clusters, in contrast to other clustering methods that require this value to be predefined [8].

HDBSCAN is an extension of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), developed to enhance clustering performance on datasets with varying density structures. A previous study [9] demonstrated that HDBSCAN outperforms DBSCAN in identifying clusters with varying sizes, shapes, and densities without requiring the manual specification of the epsilon parameter as in DBSCAN. In DBSCAN, the epsilon parameter is used as a global threshold to determine the neighborhood density of a data point. This approach becomes less effective when applied to inter-provincial flood impact data, where density distributions vary significantly across regions. Under such conditions, a single epsilon value capable of representing high-density regions may fail to capture lower-density clusters, potentially leading to unrepresentative results. To overcome these limitations, HDBSCAN transforms the DBSCAN concept into a hierarchical density-based clustering approach that evaluates all possible epsilon values simultaneously. Consequently, HDBSCAN can adapt to local density variations without depending on a single global threshold by utilizing the mutual reachability distance metric to construct a hierarchical structure based on data point density and extract the most stable clusters through a cluster condensation process. These capabilities make HDBSCAN a highly suitable method for clustering disaster-related data, which is typically complex and heterogeneous across regions. Therefore, this study's first novelty lies in the application of the HDBSCAN to cluster provinces in Indonesia based on flood impact indicators, which has been relatively underexplored in similar studies.

Furthermore, previous studies [5] have generally utilized housing damage variables as a broad aggregate, failing to delineate the specific severity of such damage. In practice, the degree of housing damage can represent different levels of flood impact across regions. Therefore, the second novelty of this study involves the application of more granular flood impact indicators by disaggregating housing damage into severely damaged houses, moderately damaged houses, and slightly damaged houses, as well as incorporating the number of flooded houses into the analysis. By including the variables of the number of deaths and missing victims, the number of injured and displaced victims, the number of severely damaged houses, the number of moderately damaged houses, the number of slightly damaged houses, and the number of flooded houses, this study aims to produce a more specific and comprehensive mapping of flood impact characteristics across provinces in Indonesia.

However, the effectiveness of the HDBSCAN model strongly depends on the selection of hyperparameters, particularly the minimum samples and minimum cluster size parameters [10]. Inappropriate hyperparameter selection may lead to less representative clustering results. Therefore, an optimization method is required to determine the optimal hyperparameter values. One such method is Bayesian Optimization, a probabilistic and model-based technique designed to efficiently search the parameter space [11]. Unlike conventional search methods that require numerous trials, Bayesian Optimization utilizes Bayes' theorem and Gaussian Processes to model the objective function, enabling a more efficient hyperparameter search process in terms of both accuracy and computational cost [12]. According to a prior study [13], Bayesian Optimization successfully determined the optimal hyperparameters for HDBSCAN and significantly enhanced clustering performance, with the silhouette score improving from 0.15 to 0.65. Consequently, integrating Bayesian Optimization with HDBSCAN provides an effective strategy for obtaining optimal hyperparameters and improving clustering performance.

This study aims to implement HDBSCAN with Bayesian Optimization to cluster provinces in Indonesia based on flood impact indicators. The novelty of this study is reflected in three primary aspects. First, while HDBSCAN and Bayesian Optimization have been applied in various domains, the combined application of these methods to cluster Indonesian provinces based on flood impacts remains relatively limited; thus, this study offers a methodological contribution to disaster risk analysis in Indonesia. Second, this study utilizes more detailed and multidimensional flood impact indicators compared to previous studies, specifically accounting for the severity of housing damage and incorporating the number of submerged houses into the analysis. Third, Bayesian Optimization is employed to systematically determine the optimal HDBSCAN hyperparameters, ensuring that the clustering results do not rely on arbitrary parameter selection and enhancing overall clustering performance. Consequently, the findings of this study are expected to offer insights to policymakers and relevant agencies in developing more targeted flood mitigation strategies based on the specific flood-impact characteristics of each province in Indonesia.

II. METHOD

The systematic workflow is illustrated in Figure 1.

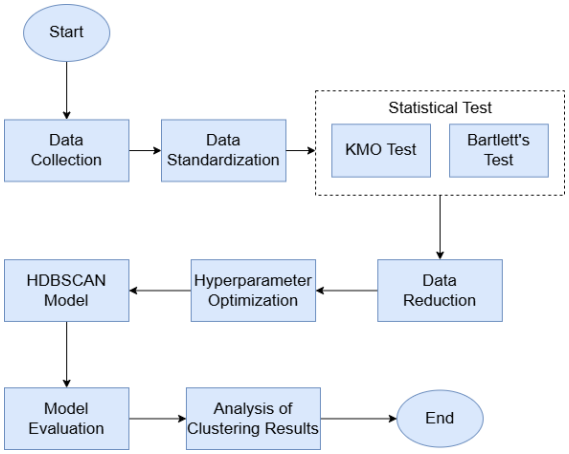

Figure 1. Systematic Workflow

A. Data Collection

In this study, the datasets used are secondary data obtained from the official website of the National Disaster Management Agency. The dataset represents aggregated flood impact statistics for all 38 provinces in Indonesia for the year 2024. Each observation corresponds to a province, with values representing the total number of flood-related impacts recorded within the year. The variables used in this study include the number of deaths and missing victims (X1), the number of injured and displaced victims (X2), the number of severely damaged houses (X3), the number of moderately damaged houses (X4), the number of slightly damaged houses (X5), and the number of flooded houses (X6). All variables are measured in absolute counts and reflect cumulative impacts of flood events within the specified period.

B. Data Standardization

Data standardization aims to transform variable scales so that their value ranges become comparable across variables. This transformation is crucial in the context of flood impact clustering, as the variables used in this study exhibit significantly different measurement scales and numerical ranges across provinces. Without standardization, variables with inherently larger numerical values could potentially dominate the distance calculations in the HDBSCAN algorithm, leading to biased and unrepresentative cluster structures. In this study, data standardization was performed using the z-score method with the *StandardScaler()* function, which ensures scale consistency across numerical variables in the dataset [14]. This method works by subtracting the variable's mean and dividing the result by the standard deviation, ensuring a uniform distribution with a mean of 0 and a standard deviation of 1 [15]. The operation of StandardScaler is formally defined by the following equation:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Where x is the original value, μ is the mean value, and σ is the standard deviation.

C. Statistical Test

In this study, several statistical tests were conducted to assess data adequacy prior to further analysis, including the Kaiser-Meyer-Olkin (KMO) test and Bartlett's Test of Sphericity. The KMO test evaluates sampling adequacy by examining the relationship between observed correlations and partial correlations among variables. The KMO value ranges from 0 to 1, where a KMO value > 0.50 indicates that the dataset is adequate and suitable for further analysis [16]. The KMO statistic is defined as:

$$KMO = \frac{\Sigma_{i=1}^{p} \quad \Sigma_{j=1}^{p} \quad r_{ij}{}^2}{\Sigma_{i=1}^{p} \quad \Sigma_{j=1}^{p} \quad r_{ij}{}^2 + \Sigma_{i=1}^{p} \quad \Sigma_{j=1}^{p} \quad a_{ij}{}^2} \quad (2)$$

Concurrently, Bartlett's Test of Sphericity is employed to evaluate the presence of significant correlations between variables. If variables were entirely independent, the correlation matrix would function as an identity matrix. Therefore, a significant result in this test confirms that the matrix deviates from an identity structure, validating the suitability of PCA for the dataset [17]. The Bartlett's test is computed as follows:

$$X = -\left(n - 1 - \frac{2p+5}{6}\right) In|R| \quad (3)$$

In this formulation, n denotes the total number of observations, p represents the number of variables, and $|R|$ signifies the determinant of the correlation matrix. If the statistics satisfies $X^2_{count} > X^2_{table}$ or p-value < 0.05, the null hypothesis is rejected, confirming that the correlation matrix significantly deviates from the identity matrix and indicating a significant correlation among the variables.

D. Data Reduction

Data reduction aims to minimize the number of variables while preserving the core informational integrity and mitigating multicollinearity concerns among the features. This study utilizes Principal Component Analysis (PCA) for dimensionality reduction, transforming interrelated variables into a smaller set of orthogonal components that preserve a significant proportion of the original dataset's variance [18]. PCA was selected because the flood impact indicators used in this study are potentially intercorrelated. For instance, provinces with a high number of flooded houses often exhibit high numbers of displaced victims or damaged dwellings. Such correlations can impair clustering performance and increase data redundancy. Consequently, PCA was employed to achieve a more compact and representative feature space prior to the application of the HDBSCAN algorithm. The PCA is initiated by calculating the covariance matrix of the standardized data:

$$cov(X) = \frac{1}{n-1} X^T X \quad (4)$$

Where n is the number of observations, X is the standardized data matrix, and X^T is the transpose of X. Next, eigenvalues and eigenvectors are obtained from the covariance matrix using the following equations:

$$cov(X)v = \lambda v \quad (5)$$

$$det(cov(X) - \lambda I = 0 \quad (6)$$

The eigenvalues represent the proportion of variance captured by each principal component, while the eigenvectors specify the linear combinations of the original variables that form these components. The principal components are selected using the Kaiser criterion of eigenvalues > 1 and a minimum cumulative variance of 80% [19]. These criteria ensure that the retained components preserve most of the information contained in the original dataset while reducing dimensional complexity. Finally, the data are projected onto the selected principal components using the following transformation:

$$Z = XV \quad (7)$$

Where V is the eigenvector matrix of the selected principal components. The resulting matrix Z represents the dimensionality-reduced dataset that serves as the input for the subsequent HDBSCAN clustering process.

E. Hyperparameter Optimization

In this study, Bayesian Optimization is employed to efficiently search for the optimal hyperparameters of HDBSCAN. This optimization method utilizes the principle of Bayes' theorem to construct a probabilistic model of the objective function by using a Gaussian Process as the primary approach for modeling the relationship between parameters and the function value [20]. At each iteration, the Gaussian Process is updated based on all previously evaluated hyperparameter configurations and their associated scores, producing a posterior distribution that estimates both the expected performance and the uncertainty of unexplored configurations. A key advantage of this method is its capability to identify the global optimal solution in fewer iterations, as the search process is guided by probabilistic information from previous evaluations rather than by random exploration [21].

The optimization process focuses on two key hyperparameters of the HDBSCAN, namely minimum samples (*min_samples*) and minimum cluster size (*min_cluster_size*). The search space of hyperparameters is defined as *min_samples* ∈ [1, 6] and *min_cluster_size* ∈ [2, 7], with both parameter bounds were determined based on the dataset size of 38 provincial observations, ensuring that the search space remains practically meaningful and computationally feasible. The objective of the optimization is to maximize the clustering quality measured using the Density-Based Clustering Validation (DBCV) index. The DBCV index was selected as the objective function because it is specifically designed to evaluate density-based clustering results by considering both the intra-cluster density and the inter-cluster density separability, making it more appropriate for assessing HDBSCAN outputs. During the optimization process, the objective function evaluates the clustering result generated by a set of hyperparameters and returns the corresponding DBCV score. The optimal hyperparameters are obtained by maximizing the following objective function:

$$x^* = argmax_{x \in X} f(x) \quad (8)$$

Where x represents the set of hyperparameters, X represents the search space, and $f(x)$ represents the objective function that returns the DBCV score. During the optimization process, Bayesian Optimization iteratively evaluates different hyperparameter configurations and updates the surrogate model to guide the search toward parameter combinations that yield higher DBCV scores. The resulting optimal hyperparameters are subsequently used in the HDBSCAN clustering process.

F. HDBSCAN Algorithm

After obtaining the optimal hyperparameter values through Bayesian Optimization, the clustering process is performed using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm. HDBSCAN is a density-based clustering method that extends DBSCAN by using the mutual reachability distance metric to construct a hierarchical structure based on data-point density [22]. The HDBSCAN workflow is illustrated in Figure 2.

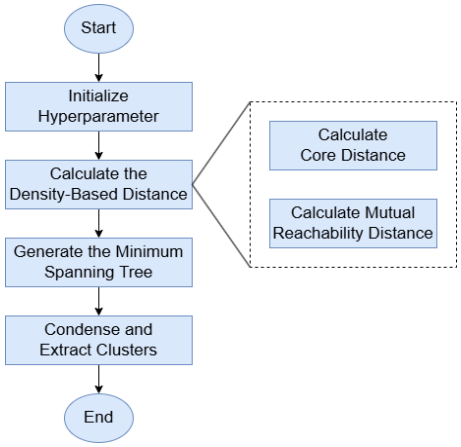

Figure 2. HDBSCAN Workflow

1) Initialize Optimal Hyperparameter

The clustering process begins by initializing the optimal hyperparameter values (*min_samples* and *min_cluster_size*) obtained from the Bayesian Optimization procedure. The *min_samples* parameter dictates the necessary neighbor count for a data point to be designated as a core point, thereby affecting local density estimation. Meanwhile, *min_cluster_size* defines the lower threshold for a group of points to be recognized as a legitimate cluster [23].

2) Calculate the Density-Based Distance

The core distance is derived from the proximity of each point to its *k*-nearest neighbor, with *k* is governed by the *min_samples* parameter. This value facilitates the calculation of the mutual reachability distance between points *a* and *b*, process intended to sharpen the separation of low-density regions, as given by:

$$d_{mreach-k}(a,b) = max\{core_k(a), core_k(b), d(a,b)\} \quad (9)$$

Where $core_k(a)$ and $core_k(b)$ are the core distances of points a and b, respectively, and $d(a,b)$ is the Euclidean distance.

3) Generate the Minimum Spanning Tree (MST)

A weighted graph is constructed in which each data points represents a nodes and edges are weighted by mutual reachability distances [24]. Subsequently, a MST is generated by connecting all nodes with a minimal total edge weight, ensuring the resulting graph remains free of cycles [25]. The MST is subsequently transformed into a hierarchical cluster structure by iteratively removing edges from highest to lowest weight.

4) Condense and Extract Clusters

The final stage involves condensing the hierarchical tree to identify stable clusters, defined as groups of points that remain consistent across certain density levels. Clusters that do not exhibit sufficient stability are not retained in the hierarchy. Data points that fail to align with any stable cluster are subsequently classified as noise.

G. Model Evaluation

Performance evaluation of the model relies on the DBCV and Silhouette Score to ensure the generation of well-defined and well-separated clusters.

1) DBCV (Density-Based Clustering Validation)

In this study, DBCV is employed as the primary evaluation metric because it is specifically designed for density-based clustering algorithms such as HDBSCAN and can effectively account for noise as well as variations in data density. The DBCV score is computed by comparing the internal density connectivity of clusters with the density separation among different clusters [26]. This score is bounded between −1 and 1, with a higher scores signify more well-defined clusters. The DBCV index for clustering result C is computed as:

$$DBCV(C) = \sum_{i=1}^{l} \frac{|C_i|}{|O|} VC_i \quad (10)$$

Where $|C_i|$ is the number of points in cluster C_i, $|O|$ is the total number of points in the dataset (including noise), and VC_i is the validity value of cluster C_i that is defined as:

$$VC_i = \frac{min_{0\le j\le l, j\neq i}(DSPC_{i,j}) - DSPC_i}{max(min_{0\le j\le l, j\neq i}(DSPC_{i,j}) - DSPC_i)} \quad (11)$$

Where $(DSPC_i)$ is the maximum internal edge weight of the MST of cluster C_i, and $(DSPC_{i,j})$ is the minimum mutual reachability distance between points belonging to two different clusters (C_i and C_j).

2) Silhouette Score

To ensure a comprehensive evaluation of the clustering outcomes, the Silhouette Score is employed as a secondary metric to complement DBCV, offering a more nuanced assessment of the compactness and isolation of the clusters produced by HDBSCAN. As an internal evaluation metric, ithis metric quantifies clustering quality by evaluating the balance between intra-cluster cohesion and inter-cluster separation. The Silhouette Score ranges from -1 to 1, where a value close to 1 indicates well-defined clusters [28]. The Silhouette Score is calculated using the following equation:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (12)$$

Where $a(i)$ denotes the mean intra-cluster distance for point *i*, whereas $b(i)$ signifies the minimum mean distance between point *i* and all points in the closest adjacent cluster.

H. Analysis of Clustering Results

The final step of this study is the analysis of clustering results. Each cluster is analyzed to identify its characteristics based on the flood impact indicators. These characteristics are determined by averaging the values of each variable within each cluster, which is then transformed into relative percentages across clusters. The percentage contribution of cluster *k* to variable *j* is calculated as follows:

$$P_{k,j} = \frac{\underline{x}_{k,j}}{\sum_{k=1}^{K} \underline{x}_{k,j}} \times 100\% \quad (12)$$

Where $P_{k,j}$ is the percentage contribution of cluster *k* to variable *j*, K is the total number of clusters, and $\underline{x}_{k,j}$ is the mean value of variable *j* within cluster *k*.

III. RESULTS AND DISCUSSION

A. Data Collection

The variables used in this study include the number of deaths and missing victims (X1), the number of injured and displaced victims (X2), the number of severely damaged houses (X3), the number of moderately damaged houses (X4), the number of slightly damaged houses (X5), and the number of flooded houses (X6). A sample of the dataset is presented in Table 1.

TABLE I
SAMPLE OF THE DATASET

Province	X1	X2	X3	….	X6
Riau	0	248958	28	…	86147
Jambi	4	315814	528	…	103894
West Sumatra	72	303357	1363	…	68550
….	….	….	….	….	….
South Papua	0	3010	0	….	54

B. Data Standardization

The *StandardScaler()* function was employed to perform z-score standardization. The results of the standardization are presented in Table 2.

TABLE II
RESULTS OF DATA STANDARDIZATION

Province	X1	X2	X3	….	X6
Riau	-0.502	0.356	-0.381	….	0.528
Jambi	-0.179	0.643	-0.331	….	1.118
West Sumatra	5.314	0.589	4.779	….	0.749
….	….	….	….	….	….
South Papua	-0.502	-0.700	-0.439	….	-0.689

C. Statistical Test

Prior to performing PCA, the KMO test and Bartlett's test were applied to examine sampling adequacy and the correlation structure among the variables. Based on the calculation, the KMO value is 0.678, exceeding the minimum threshold of 0.50. This result indicates that the dataset is adequate and appropriate for further analysis. Furthermore, the Bartlett's test yielded a p-value of less than 0.05, indicating that the correlation matrix is significantly distinct from an identity matrix. This implies that a significant correlation exists among the variables. Based on these results, it can be concluded that the dataset satisfies the necessary assumptions for PCA. A high KMO value signifies sufficient sampling adequacy, while a significant result from Bartlett's validates the existence of correlations. Therefore, dimensionality reduction can be performed to identify the principal components that capture the major variation in the flood impact indicators.

D. Data Reduction

The number of principal components is determined based on the Kaiser criterion (eigenvalues > 1) and a minimum cumulative explained variance of 80%. The eigenvalues and cumulative explained variance of each principal component are presented in Table 3.

TABLE III
EIGENVALUES AND CUMULATIVE PROPORTION OF TOTAL VARIANCE

Principal Component	Eigenvalue	Cumulative Proportion of Total Variance (%)
PC1	3.628	58.88%
PC2	1.677	86.09%
PC3	0.484	93.93%
PC4	0.263	98.20%
PC5	0.080	99.50%
PC6	0.031	100%

Based on Table III, the results show that the first principal component (PC1) yields an eigenvalue of 3.628, explaining 58.88% of the total variance, whereas the second principal component (PC2) possesses an eigenvalue of 1.677, representing an additional 27.21% of the variance. Both components exhibit eigenvalues > 1 and collectively explain 86.09% of the total data variance. Consequently, based on the Kaiser criteria of an eigenvalue > 1 and a cumulative variance proportion of ≥ 80%, two principal components were selected as they are considered sufficient to represent the majority of the information within the dataset. These two components will serve as the input for the optimization process and clustering modeling. The results of the two-component dimensionality reduction are presented in Table 4.

TABLE IV
RESULTS OF THE TWO-COMPONENT DIMENSIONALITY REDUCTION

Province	PC1	PC2
Riau	-0.310	0.820
Jambi	0.184	1.258
West Sumatra	8.975	-4.028
….	….	….
South Papua	-0.917	-0.524

E. Hyperparameter Optimization

Before the clustering process, Bayesian Optimization was employed to optimize the hyperparameter values of *min_samples* and *min_cluster_size*, with the DBCV score used as the objective function. The search space was defined as *min_samples* ∈ [1, 6] and *min_cluster_size* ∈ [2, 7]. The optimization process was conducted over a total of 38 iterations and consisted of two sequential phases. The first phase involved 8 initial random exploration iterations, during which hyperparameter configurations were sampled randomly across the search space to initialize the Gaussian Process surrogate model. The second phase consisted of 30 guided optimization iterations, in which the surrogate model was iteratively updated and the Expected Improvement acquisition function was employed to guide the search toward hyperparameter configurations with higher predicted DBCV scores. This two-phase strategy provides a balance between broad exploration of the hyperparameter space during the early stages and focused exploitation of promising regions in subsequent iterations, thereby reducing the likelihood of premature convergence to a local optimum. The optimization results are summarized in Table V.

TABLE V
PARAMETER OPTIMIZATION RESULTS

Parameter	Search Space	Optimal Value
min_samples	1-6	2
min_cluster_size	2-7	2
Best DBCV Score		0.515

The optimal hyperparameter configuration was identified as *min_samples* = 2 and *min_cluster_size* = 2, yielding the highest DBCV score of 0.515. A DBCV score of 0.515 indicates a moderate to good clustering quality, suggesting that the resulting clusters possess a sufficiently clear internal density structure and meaningful separation

between clusters. The convergence visualizations of the optimal parameter search process are presented in the Figure 3 and Figure 4.

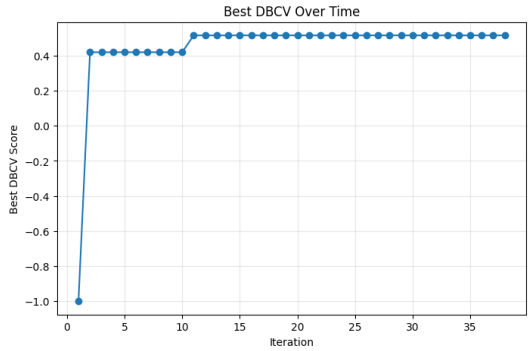

Figure 3. Best DBCV Score Over Time

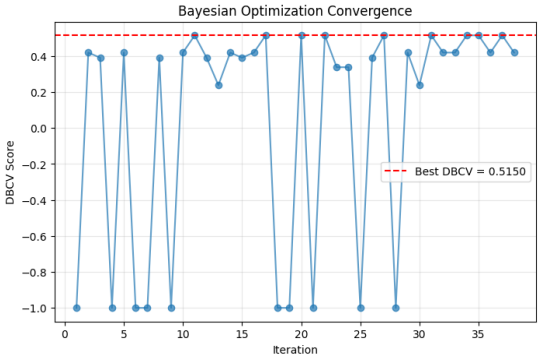

Figure 4. Bayesian Optimization Convergence

As illustrated in Figure 3, the best DBCV score increased rapidly during the early iterations and reached its maximum value at iteration 11. After this point, the best score remained stable at 0.515, indicating that the optimization process had converged to a stable solution. Furthermore, Figure 4 presents the convergence behavior of Bayesian Optimization throughout the iterations. Although the individual DBCV scores fluctuated during the search process due to the exploration and exploitation mechanism of Bayesian Optimization, the best observed value remained consistently at 0.515 after iteration 11.

These results demonstrate that the Bayesian Optimization process successfully identified the optimal hyperparameters, enabling HDBSCAN to effectively detect cluster structures within the heterogeneous distribution of flood impact characteristics across provinces in Indonesia. The optimal hyperparameters obtained from this optimization process (*min_samples* = 2 and *min_cluster_size* = 2) were subsequently used in the HDBSCAN clustering stage.

F. HDBSCAN Algorithm

The HDBSCAN algorithm was applied using the optimal parameters configurations derived from Bayesian Optimization (*min_samples* = 2 and *min_cluster_size* = 2). An optimal *min_samples* value of 2 signifies that a data point qualifies as a core point provided it possesses at least two neighboring points within its specified reachability distance. Meanwhile, an optimal *min_cluster_size* value of 2 simplies that the HDBSCAN algorithm allows for the establishment of clusters containing as few as two data observations. Based on these parameters, the clustering process identified three primary clusters and one noise cluster. The resulting clusters are graphically represented as a scatter plot in Figure 5, showing the distribution of the identified clusters.

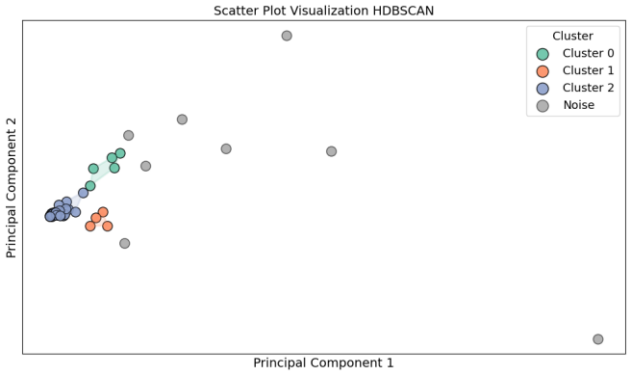

Figure 5. Cluster Distribution

Figure 5 shows the clustering results obtained by projecting the dataset onto a primary PCA defined two-dimensional subspace. Cluster 0 (green) is located in the center-to-upper region of the data distribution and forms a relatively compact group, indicating a high degree of similarity among its members. Cluster 1 (orange) is positioned adjacent to Cluster 0 but tends to be lower on the PC2 axis, suggesting distinct pattern differences despite their proximity in the feature space. Meanwhile, Cluster 2 (blue) occupies the left side of the PC1 axis and forms the densest group with the largest number of members, indicating that the provinces within this cluster possess the most homogeneous characteristics among the three. The gray points scattered outside these three clusters are identified as noise, representing data points that do not meet the density criteria for inclusion in any cluster.

Overall, the three main clusters exhibit relatively clear boundaries and are well-separated from one another, demonstrating that the model effectively groups the data based on density levels. Furthermore, the presence of noise does not interfere with the structure of the primary clusters, but represents data points that lack sufficient density to be categorized. This indicates that the HDBSCAN algorithm with Bayesian Optimization is capable of achieving optimal data clustering.

G. Model Evaluation

To further substantiate the efficacy of the proposed method, the performance of HDBSCAN with Bayesian Optimization was compared with that of DBSCAN optimized using the same optimization framework. The

comparison was conducted using two evaluation metrics, namely the DBCV index as the primary evaluation metric and the Silhouette Score as a complementary metric. The comparison results are presented in Table VI.

TABLE VI
COMPARISON OF CLUSTERING PERFORMANCE

Evaluation	DBSCAN+BO	HDBSCAN+BO
DBCV	0.505	**0.515**
Silhouette Score	0.416	**0.645**

As presented in Table VI, the comparison results demonstrate that HDBSCAN with Bayesian Optimization outperforms DBSCAN with Bayesian Optimization across all evaluation metrics. The DBCV score approaching 1 signifies a well-defined clustering structure with high internal density cohesion and clear inter-cluster separation. In terms of DBCV score, HDBSCAN achieved a higher value of 0.515 compared to 0.505 for DBSCAN, indicating a superior density-based cluster structure. Furthermore, the Silhouette Score that is also approaching 1 signifies that the clusters are well-separated and distinct. The Silhouette Score obtained by HDBSCAN reached 0.645, substantially higher than the 0.416 achieved by DBSCAN, suggesting that HDBSCAN generated more well-separated clusters. Overall, these results confirm that HDBSCAN is more effective than DBSCAN in identifying flood impact patterns.

H. Analysis of Clustering Results

The clustering results reveal substantial variation in flood impact characteristics across Indonesian provinces. The HDBSCAN algorithm identifies three primary clusters and one noise cluster. The provincial composition of each cluster is presented in Table VII.

TABLE VII
DISTRIBUTION OF PROVINCES ACCROSS CLUSTERS

Cluster	Number of Provinces	Provinces
0	5	Aceh, Riau, Jambi, East Java, Central Kalimantan
1	4	Banten, Central Sulawesi, North Maluku, Central Papua
2	21	Bengkulu, Bangka Belitung Islands, Lampung, Riau Islands, Bali, DI Yogyakarta, DKI Jakarta, West Nusa Tenggara, East Nusa Tenggara, South Kalimantan, East Kalimantan, Southeast Sulawesi, North Sulawesi, West Sulawesi, Gorontalo, Maluku, Southwest Papua, West Papua, Papua, South Papua, Highland Papua
-1	8	West Sumatra, South Sumatra, North Sumatra, West Java, Central Java, South Sulawesi West Kalimantan, North Kalimantan

To characterize each cluster, the percentage contribution of each cluster to the total national mean per flood impact indicator is computed. These percentages reflect the relative share of each cluster's average value compared to the sum of all cluster averages, and therefore do not represent absolute flood impact magnitudes but rather indicate which clusters dominate each indicator at the national level. The characteristics of each cluster based on the percentage contribution of flood impact indicators are summarized in Figure 6.

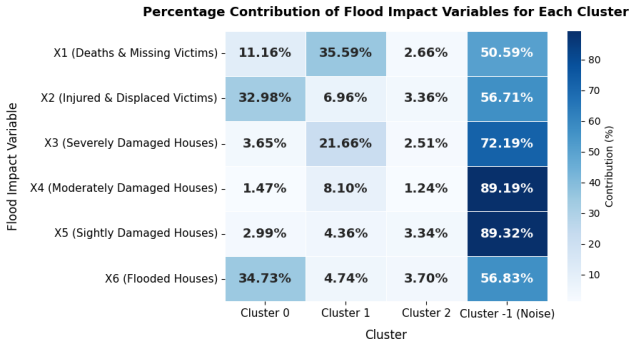

Figure 6. Percentage Contribution of Flood Impact Variables

Based on the cluster characteristics in Figure 6, Cluster 0 (High Displacement & Inundation) comprising 5 provinces, which are Aceh, Riau, Jambi, East Java, and Central Kalimantan. This cluster accounts for the highest proportions of injured and displaced victims (32.98%) and flooded houses (34.73%) compared to other valid clusters. However, its contribution to fatality and housing damage indicators remains relatively low. This pattern suggests that Cluster 0 experience flood events that significantly affect population exposure and displacement, although the level of structural damage is not as extensive as in other clusters. Consequently, this cluster can be characterized as regions with moderate flood impacts with relatively higher risks of displacement and residential inundation.

Cluster 1 (High Fatality & Structural Damage) comprising 4 provinces, which are Banten, Central Sulawesi, North Maluku, and Central Papua. This cluster exhibits the highest proportions of death and missing victims (35.59%) and severely damaged houses (21.66%) compared to the other valid clusters. In contrast, its contribution to injured and displaced victims is relatively lower compared to Cluster 0. These results indicate that flood events in Cluster 1 tend to produce more severe structural damage and higher fatality rates, even though the number of affected or displaced residents may be relatively lower. Therefore, this cluster can be categorized as regions

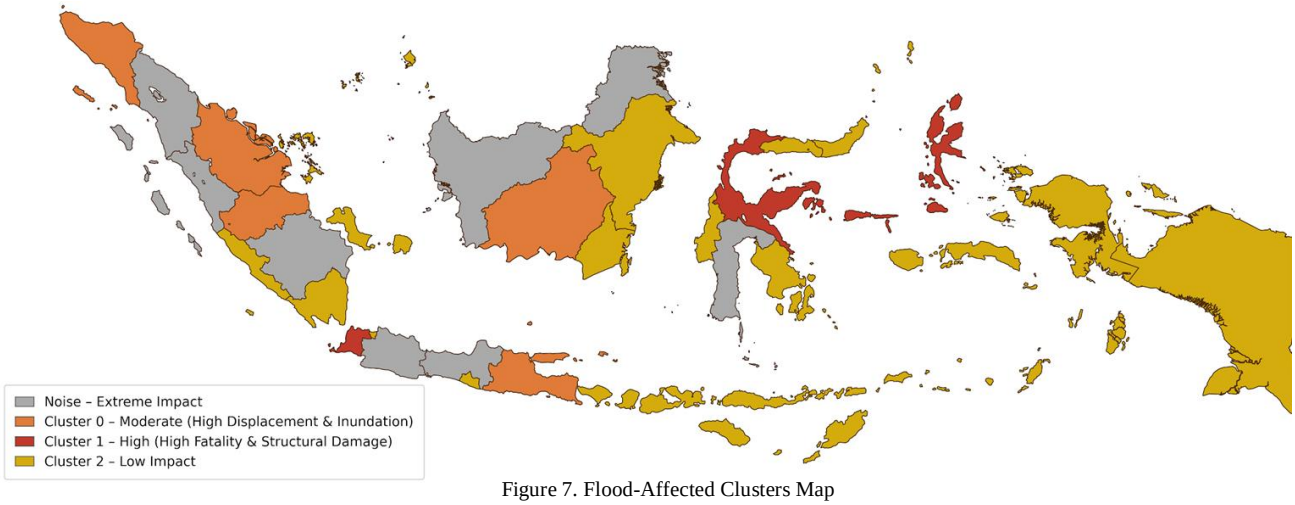

Figure 7. Flood-Affected Clusters Map

with high flood impacts with high structural damage and fatality risks.

Cluster 2 (Low Impact) is the largest clusters, comprising 21 provinces. However, this cluster exhibits relatively low proportions across all flood impact indicators. The highest contributions are observed in lightly damaged houses (3.34%) and flooded houses (3.70%), both of which represent the least severe impact categories. This indicates that the majority of provinces in Cluster 2 experience relatively low and evenly distributed flood impacts, without being dominated by any specific impact indicator. Therefore, this cluster can be categorized as regions with low flood risk characteristics.

Meanwhile, Cluster -1 (Noise) comprising 8 provinces, which are West Sumatra, South Sumatra, North Sumatra, West Java, Central Java, South Sulawesi West Kalimantan, North Kalimantan. Although the HDBSCAN algorithm does not assign these provinces to any specific cluster and classifies them as noise, this group exhibits exceptionally high proportions across all flood impact indicators. Therefore, this cluster can be categorized as regions with extreme flood impacts.

The spatial distribution of flood impact clusters across Indonesian provinces is illustrated in Figure 7.

As shown in Figure 7, the spatial distribution of the clustering results reveals distinct geographical patterns regarding the severity of flood impacts across Indonesia. Cluster 0 (orange) is primarily concentrated in the western part of Sumatra comprising Aceh, Riau, and Jambi, along with Central Kalimantan and East Java, reflecting regions that frequently experience large-scale floods characterized by extensive residential inundation and population displacement. Cluster 1 (red) exhibits a dispersed distribution spanning from western to eastern Indonesia, encompassing Banten in Java, Central Sulawesi, North Maluku, and Central Papua, suggesting that high-fatality and structurally destructive flood events are not geographically confined but may arise under specific local topographical, environmental, or infrastructural conditions across different island groups. Cluster 2 (yellow) is the most spatially dominant cluster, covering the majority of provinces across Indonesia, particularly in central Kalimantan, most of Sulawesi, Nusa Tenggara, Maluku, and Papua, representing regions with relatively low and evenly distributed flood impacts across all indicators. Meanwhile, provinces classified as noise (grey) are distributed across Indonesia's most densely populated and economically significant islands, including Sumatra, Java, Kalimantan, and Sulawesi. The concentration of noise provinces in these major islands is consistent with their extreme flood impact profiles, as higher population density and more extensive human settlements in these regions tend to amplify the scale of flood-related casualties, displacement, and structural damage relative to less densely populated provinces.

Overall, the clustering results provide practical implications for the government in designing more targeted flood mitigation strategies. By grouping provinces based on their flood impact characteristics, policymakers can allocate resources more efficiently and prioritize interventions according to regional needs. Cluster 0, which records the highest proportions of injured and displaced victims as well as flooded houses, indicates that flood events in these provinces primarily threaten population mobility and residential safety rather than causing severe structural damage. Therefore, provinces within this cluster are recommended to prioritize the strengthening of community-based early warning systems, evacuation management protocols, and drainage infrastructure improvements to minimize the extent of inundation. Cluster 1, characterized by the highest proportions of fatalities and missing persons as well as severely damaged houses, reflects flood events with highly destructive physical consequences. Provinces in this cluster therefore require

policy interventions focused on structural resilience and emergency response capacity, including the implementation of flood-resistant construction standards, the strategic pre-positioning of search and rescue (SAR) equipment, and the strengthening of emergency medical supply chains.

Cluster 2, which consistently exhibits low proportions across all impact indicators, suggests that flood events in these provinces can still be managed within existing mitigation frameworks. Accordingly, provinces in this cluster are advised to maintain and further optimize current strategies, including routine drainage system maintenance and systematic environmental monitoring, to prevent significant escalation of flood risks in the future. Meanwhile, provinces classified as noise (Cluster -1) which display very high proportions across all flood impact indicators, require the highest level of policy attention. The extreme and heterogeneous nature of their impact profiles necessitates comprehensive in-depth assessments and the development of region-specific policy frameworks, ranging from spatial planning and land-use governance to the formulation of dedicated cross-sectoral mitigation strategies. Therefore, these clustering results can support data-driven decision-making for flood disaster mitigation planning.

IV. CONCLUSIONS

This study aimed to identify patterns of flood impacts across Indonesian provinces using the HDBSCAN with Bayesian Optimization for automated parameter selection. The Bayesian Optimization process yielded an ideal configuration *min_samples* = 2 and *min_cluster_size* = 2, yielding a DBCV score of 0.515, indicating a relatively well-defined clustering structure. Based on these optimal parameters, the HDBSCAN algorithm identified three primary clusters and one noise cluster. Cluster 0 represents provinces with moderate flood impacts dominated by large-scale displacement and residential inundation. Cluster 1 represents provinces with high flood impacts characterized by elevated fatality rates and substantial structural damage. Cluster 2, comprising the majority of Indonesian provinces, exhibits relatively low proportions across all flood impact indicators, representing regions with comparatively lower flood risks. Provinces classified as noise (Cluster -1) exhibit exceptionally high and heterogeneous flood impact profiles across all indicators, suggesting that their extreme characteristics cannot be adequately captured by standard cluster structures.

The spatial distribution of the identified clusters highlights the geographical diversity of flood impacts across Indonesia, reflecting variations in environmental conditions, population exposure, and infrastructure resilience among provinces. These findings demonstrate the implementation of HDBSCAN with Bayesian Optimization for provincial-level flood impact clustering while providing a more detailed characterization of flood impacts through multidimensional indicators. From a practical perspective, the clustering results may support disaster management authorities in developing region-specific flood mitigation strategies based on the risk characteristics of each province. Nevertheless, the findings are limited to a specific observation period and may not fully represent flood impact patterns under different temporal conditions. Consequently, future study is encouraged to integrate longitudinal datasets, a broader range of environmental variables, and higher-resolution spatial scales, while extending the application of this methodology to other disaster types within the Indonesian context.

REFERENCES

[1] Data Bencana Indonesia 2023. Pusat Data Informasi dan Komunikasi Kebencanaan Badan Nasional Penanggulangan Bencana, 2023.

[2] Data Bencana Indonesia 2024. Pusat Data Informasi dan Komunikasi Kebencanaan Badan Nasional Penanggulangan Bencana, 2025.

[3] W. T. Oktaviany, F. Insani, and A. Nazir, "Pengelompokan Wilayah Bencana Banjir di Indonesia Menggunakan Algoritma K-Means," BULLETIN OF COMPUTER SCIENCE RESEARCH, vol. 5, no. 4, pp. 542–552, Jun. 2025, doi: 10.47065/bulletincsr.v5i4.608.

[4] U. Islamy et al., "Pengelompokkan Provinsi Di Indonesia Berdasarkan Indikator Dampak Bencana Banjir Tahun 2017-2020 Menggunakan K-Medoids," Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya, vol. 11, no. 2, pp. 381–388, 2022.

[5] M. N. Hayati et al., "Pengelompokan Provinsi Di Indonesia Berdasarkan Data Jumlah Kejadian Dan Dampak Bencana Banjir Menggunakan Metode Fuzzy C-Means," Variansi: Journal of Statistics and Its Application on Teaching and Research, vol. 6, no. 01, pp. 21–34, 2024, doi: 10.35580/variansiunm167.

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

[7] D. N. Amalina and A. Fauzan, "A Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) Approach for Identifying Potential Villages in Buleleng Regency," Knowledge Engineering and Data Science, vol. 7, no. 2, Dec. 2024, doi: 10.17977/um018v7i22024p187-199.

[8] M. Mouhiha and A. Mabrouk, "An empirically-driven clustering framework for NoSQL data warehouse conversion: Optimizing column family design from relational big data using HDBSCAN," Inf Softw Technol, vol. 195, p. 108117, Jul. 2026, doi: 10.1016/j.infsof.2026.108117.

[9] M. Aljibawi, H. K. Algabri, and Z. I. Rasool, "Adaptive Clustering Using Enhanced DBSCAN: a Dynamic Approach to Optimizing Density-based Clustering," Statistics, Optimization and Information Computing, vol. 14, no. 4, pp. 1980–1991, Sep. 2025, doi: 10.19139/soic-2310-5070-2484.

[10] D. P. Uddandarao, M. R. Konatham, and R. K. Vadlamani, "Robust and Scalable Statistical Models for High-Dimensional Marketing Data Applications: A Comprehensive Review," in 2025 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), IEEE, Sep. 2025, pp. 1–8. doi: 10.1109/ICERECT65215.2025.11377070.

[11] M. Binois and N. Wycoff, "A Survey on High-dimensional Gaussian Process Modeling with Application to Bayesian Optimization," ACM Transactions on Evolutionary Learning and Optimization, vol. 2, no. 2, pp. 1–26, Jun. 2022, doi: 10.1145/3545611.

[12] Md. S. S. Islam et al., "Optimizing Short-Term Photovoltaic Power Forecasting: A Novel Approach with Gaussian Process Regression

and Bayesian Hyperparameter Tuning," Processes, vol. 12, no. 3, p. 546, Mar. 2024, doi: 10.3390/pr12030546.
[13] C. Thanos, C. Meghini, V. Bartalesi, and G. Coro, "An exploratory approach to data driven knowledge creation," J Big Data, vol. 10, no. 1, p. 29, Mar. 2023, doi: 10.1186/s40537-023-00702-x.
[14] M. A. Mohammed, "Effect of Using Numerical Data Scaling on Supervised Machine Learning Performance," Global Libyan Journal, vol. 67, pp. 1–21, 2024.
[15] F. Aldi, F. Hadi, N. A. Rahmi, and S. Defit, "Standardscaler's Potential In Enhancing Breast Cancer Accuracy Using Machine Learning," Journal of Applied Engineering and Technological Science, vol. 5, no. 1, pp. 401–413, 2023.
[16] Zhang, T. Sangsawang, K. Vipahasna, and M. Pigultong, "A Mixed-Methods Data Approach Integrating Importance-Performance Analysis (IPA) and Kaiser-Meyer-Olkin (KMO) in Applied Talent Cultivation," Journal of Applied Data Sciences, vol. 5, no. 1, pp. 256–267, Jan. 2024, doi: 10.47738/jads.v5i1.170.
[17] S. Nabhan and A. Habók, "The Digital Literacy Academic Writing Scale: Exploratory Factor Analysis," Sage Open, vol. 15, no. 1, Jan. 2025, doi: 10.1177/21582440241311709.
[18] S. H. A. Latif, A. S. Alwan, and A. M. Mohamed, "Principal component analysis as tool for data reduction with an application," EUREKA: Physics and Engineering, no. 5, pp. 184–198, Sep. 2022, doi: 10.21303/2461-4262.2022.002577.
[19] S. Ramasubramanian, S. C.N, A. J. Athreya, A. Devarajan, A. U. Shankar, and R. Kumar P, "Data Dimensionality Reduction Using Principal Component Analysis: A Case Study," in 2024 1st International Conference on Communications and Computer Science (InCCCS), IEEE, May 2024, pp. 1–6. doi: 10.1109/InCCCS60947.2024.10593421.
[20] A. Hebbal, M. Balesdent, L. Brevault, N. Melab, and E.-G. Talbi, "Deep Gaussian process for multi-objective Bayesian optimization," Optimization and Engineering, vol. 24, no. 3, pp. 1809–1848, Sep. 2023, doi: 10.1007/s11081-022-09753-0.
[21] X. Wang, Y. Jin, S. Schmitt, and M. Olhofer, "Recent Advances in Bayesian Optimization," ACM Comput Surv, vol. 55, no. 13s, pp. 1–36, Dec. 2023, doi: 10.1145/3582078.
[22] D. Vijayan and I. Aziz, "Adaptive Hierarchical Density-Based Spatial Clustering Algorithm for Streaming Applications," Telecom, vol. 4, no. 1, pp. 1–14, Dec. 2022, doi: 10.3390/telecom4010001.
[23] A. Sante, A. S. Font, D. Mistry, S. Ortega-Martorell, and I. Olier, "Optimized HDBSCAN clustering for reconstructing the merger history of the Milky Way: applications and limitations," Mon Not R Astron Soc, Mar. 2026, doi: 10.1093/mnras/stag503.
[24] L. Wang, P. Chen, L. Chen, and J. Mou, "Ship AIS Trajectory Clustering: An HDBSCAN-Based Approach," J Mar Sci Eng, vol. 9, no. 6, p. 566, May 2021, doi: 10.3390/jmse9060566.
[25] A. Mashreghi and V. King, "Broadcast and minimum spanning tree with o(m) messages in the asynchronous CONGEST model," Distrib Comput, vol. 34, no. 4, pp. 283–299, 2021.
[26] D. Chicco, G. Sabino, L. Oneto, and G. Jurman, "The DBCV index is more informative than DCSI, CDbw, and VIASCKDE indices for unsupervised clustering internal assessment of concave-shaped and density-based clusters," PeerJ Comput Sci, vol. 11, p. e3095, Aug. 2025, doi: 10.7717/peerj-cs.3095.
[27] Z. Teng, J. Yan, D. Liu, and P. Zhang, "When Does the Silhouette Score Work? A Comprehensive Study in Network Clustering," Dec. 2025, [Online]. Available: http://arxiv.org/abs/2512.24841