IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

PDBSCAN, A Density-Based Clustering Algorithm

Prabhu Patel and Dr. Abhishek Singh Rathore
Department of Computer Science
Shri Vaishnav Vidyapeeth Vishwavidyalaya (SVVV) Indore, India
Prabhupatelmp55@gmail.com

Abstract: *The DBSCAN algorithm is a popular choice for unsupervised clustering due to its ability to identify arbitrarily shaped clusters. However, its effectiveness declines when handling datasets with heterogeneous densities, significant noise, or high dimensionality, primarily due to its sensitivity to the epsilon (ε) and MinPts parameters. To overcome these challenges, this paper presents PDBSCAN, a hybrid clustering approach that incorporates noise filtering, nonlinear dimensionality reduction, adaptive clustering, and post-clustering refinement. Initial preprocessing removes outliers using Local Outlier Factor (LOF), followed by Uniform Manifold Approximation and Projection (UMAP) for structure-preserving dimensionality reduction. Clustering is then performed via HDBSCAN, which adaptively infers cluster structure without fixed parameter input. Remaining unclustered samples are assigned labels using a K-Nearest Neighbors (KNN) model, with final adjustments made through a Random Forest classifier to improve boundary consistency. Experimental evaluation on datasets including R15, D31, Flame, Compound, and Aggregation confirms that the framework achieves over 99.9% accuracy and strong ARI, NMI, and silhouette performance. The method offers scalability, resilience to noise, and is well-suited for real-world tasks in domains such as bioinformatics, anomaly detection, and geospatial analysis.*

Keywords: Density based Algorithm, DBSCAN, Noise, Local Outlier Factor, UMAP

I. INTRODUCTION

Clustering is a fundamental technique for uncovering meaningful patterns within large datasets. It is widely applied across various domains such as marketing, computer networking, image analysis, biological research, geographic monitoring, web data mining, and healthcare applications By measuring similarities between data points, clustering facilitates the automatic grouping of unlabeled data into coherent clusters. [10] The below figure 1. Shows the graphical representation of clustering.

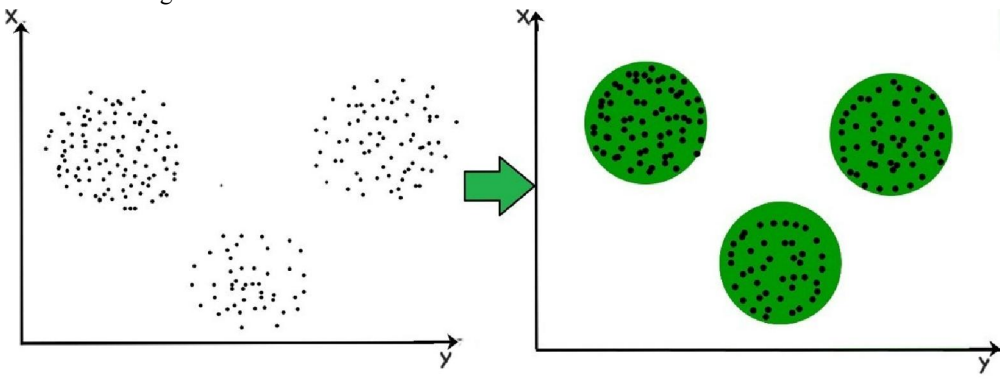

Figure 1. Graphical Representation of Clustering

DBSCAN faces two significant drawbacks: its strong dependence on parameter selection—specifically epsilon (ε) and the minimum number of points (MinPts) and its limited ability to manage clusters with varying densities. Determining suitable values for these parameters is challenging, often requiring expert insight or extensive experimentation. Additionally, DBSCAN relies on a uniform density threshold, which can lead to inaccurate clustering when dealing

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

with datasets that contain regions of different densities. These issues have led to the creation of several DBSCAN-based extensions, including OPTICS, HDBSCAN, and AMD-DBSCAN. Although these methods offer improvements in handling density variation and reducing parameter sensitivity, they still encounter difficulties with scalability, robustness to noise, and performance in high-dimensional spaces. [7]

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a prominent density-based clustering algorithm that effectively identifies dense regions in data to form clusters, while treating sparse regions as noise or outliers. It excels at grouping data points with similar characteristics into clusters and is particularly robust in detecting clusters of arbitrary shapes and varying sizes, even in the presence of noise.[12]

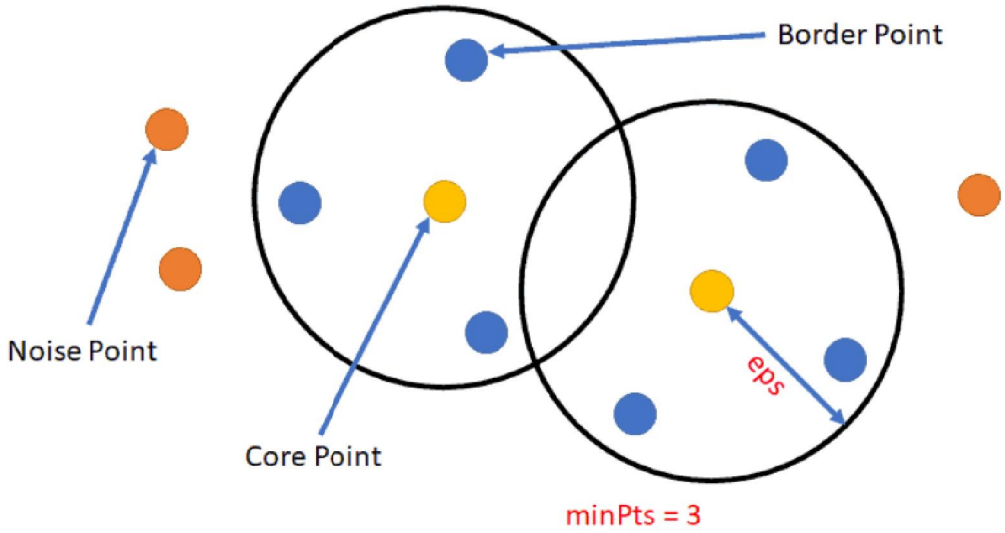

Figure 2. Graphical Representation of Core Point, Border point , Noise Point ,eps ,min point

To overcome these challenges, this paper proposes an advanced clustering framework named PDBSCAN, which integrates several novel and existing techniques into a cohesive and automated pipeline. The key idea behind this framework is to blend the advantages of density-based clustering, outlier filtering, dimensionality reduction, and supervised refinement to deliver highly accurate and robust clustering outcomes, without requiring manual parameter tuning or labeled data.

The proposed framework begins with Local Outlier Factor (LOF)-based filtering to eliminate noise and irrelevant points from the dataset. This pre-filtering step enhances the performance of subsequent clustering by reducing the impact of outliers. Next, the algorithm applies Uniform Manifold Approximation and Projection (UMAP) to reduce data dimensionality while preserving its manifold structure. This step enables more effective clustering in lower-dimensional space, especially for complex or high-dimensional datasets.[1]

Subsequently, Hierarchical DBSCAN (HDBSCAN) is used to perform density-based clustering with adaptive parameter selection. HDBSCAN naturally handles clusters with variable densities and does not require predefining ε. However, it may leave some points unclustered. To address this, a K-Nearest Neighbors (KNN) model is trained on high-confidence cluster labels and used to infer labels for unassigned points. Finally, a Random Forest classifier is employed to refine cluster boundaries and further enhance label .[3]

II. LITERATURE SURVEY

Breunig et al. [1] The Local Outlier Factor (LOF) identifies anomalies by comparing the local density of each point to that of its nearby neighbors. A substantially lower density suggests that a point is an outlier. LOF is particularly effective in data with non-uniform cluster densities due to its local density-based design, making it well-suited for real-world applications. In PDBSCAN, LOF is used during the initial preprocessing stage to eliminate noisy data points.

This step improves clustering quality by reducing fragmentation and enhancing the integrity of the resulting clusters, forming a solid basis for the framework's high clustering accuracy.

McInnes et al [2] UMAP is a non-linear dimensionality reduction technique that leverages topological insights and manifold learning to generate low-dimensional representations of high-dimensional data. It maintains both global and local relationships, often outperforming alternatives like t-SNE in terms of speed and structure preservation. Within the PDBSCAN pipeline, UMAP helps in projecting data to lower dimensions (2D/3D), simplifying complex structures and making them more amenable to clustering. This projection enhances noise separation, improves clustering performance, and supports future scalability through features like incremental learning..

Campello et al.[3] HDBSCAN extends DBSCAN by creating a hierarchy of clusters based on varying density thresholds and selects the most stable structures without needing a predefined ε. It uses mutual reachability distances and constructs a minimum spanning tree to determine cluster relationships. In PDBSCAN, HDBSCAN performs the core clustering function, automatically adapting to multi-density datasets while providing probabilistic labels, making it robust to noise and suitable for complex clustering tasks.

Allaoui et al. [4] This study evaluates how using UMAP as a precursor to clustering (e.g., with DBSCAN or k-means) improves clustering metrics such as Silhouette Score and ARI. UMAP simplifies data geometry and increases separability between clusters. This finding supports its integration in PDBSCAN, where it enables clearer boundaries and improved accuracy for subsequent clustering and classification stages.

Chen et al. [5] proposed KNN-DBSCAN, a clustering algorithm that replaces the fixed ε-radius of DBSCAN with K-nearest neighbor graphs to better handle high-dimensional and sparse data. Their method dynamically estimates local density by analyzing the neighborhood structure via KNN, enabling more accurate core point detection and reducing noise misclassification. While their focus is on adapting the clustering process itself, the use of KNN to model local relationships strongly supports the refinement strategy in our framework, where KNN is used to propagate labels for unclustered points following HDBSCAN. This parallel highlight the flexibility and power of KNN in enhancing density-based clustering performance.

Péalat et al. [6] proposed a novel method for clustering time-series data by converting sequences into geometric representations on manifolds such as the Euclidean space, Stiefel manifold, and hypersphere. They then applied UMAP for dimensionality reduction and used HDBSCAN to perform clustering. This combination allowed better preservation of time-based structures and improved clustering performance without manual parameter tuning. The methodology is closely related to PDBSCAN, which also uses UMAP and HDBSCAN, but extends it further by integrating LOF-based noise filtering and semi-supervised label refinement to handle diverse data distributions more effectively.

Bing Maa, et al. [7] This paper introduces an adaptive approach to select DBSCAN's parameters, Eps and MinPts, using the data's own distribution. By computing the average distance to each point's K-th nearest neighbor, candidate Eps values are generated. MinPts is then derived as the average number of points within these Eps neighborhoods. As K increases, both parameters rise and the resulting cluster count stabilizes. The final parameters are chosen from this stable region to enhance clustering accuracy and reduce noise.To improve efficiency, the proposed K-DBSCAN algorithm processes only core points, eliminating non-core points early and merging clusters with overlapping neighborhoods. This strategy reduces the number of point evaluations and accelerates computation without affecting clustering quality.[5]

Chenb et al [8] The research titled "A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters" introduces enhancements to the standard DBSCAN method by resolving two primary issues: the reliance on user-defined parameters (Eps and MinPts) and the algorithm's intensive computational demands. To overcome these challenges, the authors develop a self-adaptive strategy that automatically estimates Eps and MinPts from the dataset by analyzing the average distance to each point's KKKth nearest neighbor. This dynamic approach improves clustering performance, particularly when dealing with datasets of uneven density.

To reduce execution time, the study proposes K-DBSCAN, a variant of DBSCAN that processes only the core points—those with high local density—while disregarding non-core points to minimize data traversal. Clustering is achieved by detecting overlaps between the neighborhoods of these core points and merging them accordingly. Although the

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

parameter estimation process requires repeated clustering runs, the overall design improves both computational efficiency and clustering quality, making the approach suitable for large-scale and complex data scenarios. [8].

III. METHODOLOGY USED

1. System Overview

The proposed PDBSCAN is a robust and adaptive clustering framework tailored to address the challenges of density-based clustering in high-dimensional, noisy, and multi-density environments. It combines outlier detection, nonlinear dimensionality reduction, adaptive hierarchical clustering, and semi-supervised refinement into a unified pipeline. This section outlines the methodological architecture, focusing on six major modules:

A. LOF-Based Outlier Filtering

The first step in the proposed framework involves removing noisy data points to enhance the purity of clusters during the clustering stage. For this, the Local Outlier Factor (LOF) method is employed. LOF evaluates each data point's local density relative to its neighbors by computing a score that reflects how isolated a point is. Points located in significantly less dense areas compared to their neighbors are assigned higher LOF scores and are classified as outliers. These outliers are then excluded from the clustering process. This step is critical for mitigating the adverse effects of noise, which often distort cluster boundaries in density-based algorithms. By filtering out high-LOF points early, we enhance the compactness and separability of the remaining data.

B. UMAP-Based Nonlinear Dimensionality Reduction

Following noise removal, the data often remains high-dimensional, which can hinder clustering algorithms due to the curse of dimensionality. To address this, Uniform Manifold Approximation and Projection (UMAP) is employed. UMAP is a powerful nonlinear dimensionality reduction technique that preserves both the local neighborhood relationships and global data structure. It projects high-dimensional data into a 2D or 3D latent space while maintaining the topological and metric structure of the original space. This transformation simplifies complex geometrical structures, making the clustering task significantly more tractable and visually interpretable. Unlike PCA or t-SNE, UMAP balances computational efficiency with meaningful embeddings for clustering tasks.

C. Core Clustering Using HDBSCAN

After dimensionality reduction, the data is subjected to clustering using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). HDBSCAN enhances traditional DBSCAN by eliminating the need for a global epsilon value and adapting naturally to clusters with varying densities. It works by constructing a mutual reachability graph, converting it into a minimum spanning tree, and then extracting stable clusters based on cluster persistence. This hierarchy-driven approach results in more flexible and accurate cluster formation, especially in datasets that exhibit complex density patterns. HDBSCAN also marks ambiguous or sparse regions as noise, ensuring the purity of formed clusters.

D. Optional Ensemble DBSCAN Voting (for Robustness)

To further increase robustness, an optional ensemble module can be integrated. In this approach, multiple instances of DBSCAN (with varied eps and minPts values) are run on the UMAP-reduced data. Their outputs are combined using majority voting, where the final label of each point is decided based on consensus across models. This ensemble approach reduces the sensitivity to individual parameter choices and balances biases inherent to different density estimations. Though optional, this module can be particularly effective in complex or noisy real-world datasets where single-pass clustering may be suboptimal.

E. Label Propagation via K-Nearest Neighbors (KNN)

Post HDBSCAN, a subset of data points often remains unlabeled due to their ambiguous location in low-density or overlapping regions. To address this, we employ K-Nearest Neighbors (KNN) as a label propagation technique. Initially, the points labeled by HDBSCAN are used to train a KNN classifier. Then, this classifier predicts the labels for the previously unclustered (noise) points based on majority labels of their nearest neighbors. This technique not only fills in the gaps left by HDBSCAN but also ensures smooth cluster boundaries and improved completeness.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

F. Classifier-Based Refinement

For enhanced generalization and refinement, a final classifier-based module can be applied. A supervised classifier—such as Support Vector Machines (SVM), Logistic Regression, or Random Forest—is trained using the high-confidence clustered data (from HDBSCAN + KNN). This classifier can then revalidate and refine label assignments across the dataset or predict labels on new, unseen samples. This stage is especially useful in semi-supervised or streaming data scenarios where ongoing refinement is necessary. Moreover, classifier feedback can be used to evaluate inter-cluster separability, further supporting the robustness of the clustering framework.

Summary of Methodological Advantages

The modular design of PDBSCANensures that each step contributes toward handling specific shortcomings in traditional clustering algorithms:

LOF removes noise and enhances cluster compactness.
UMAP simplifies complex structures while preserving cluster geometry.
HDBSCAN delivers automatic and adaptive clustering across densities.
Ensemble DBSCAN increases robustness across parameter variations.
KNN label propagation improves cluster completeness.
Classifier refinement boosts generalization and scalability.

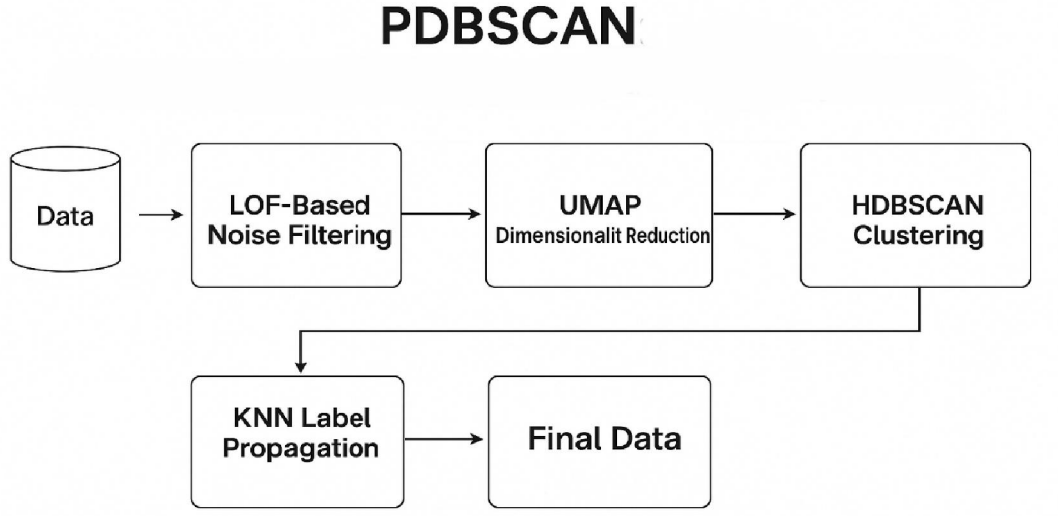

Figure 3. Graphical Representation of Pipeline of PDBSCAN

2. Dataset

To evaluate the effectiveness and generalizability of the proposed framework, five benchmark datasets were selected from publicly available repositories, primarily from GitHub. These datasets—R15, D31, Aggregation PDBSCAN, Flame, and Compound—are widely used in clustering research and represent a diverse range of characteristics such as varying densities, cluster shapes, and levels of noise.

A. Dataset Overview

R15 Dataset

R15 consists of 15 well-separated Gaussian clusters with uniform size and density. It includes 600 samples with two numerical features. This dataset is suitable for evaluating clustering algorithms on well-structured, moderately complex data.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

D31 Dataset

D31 includes 3100 samples arranged into 31 densely packed and overlapping clusters. With only two numerical features, it challenges the algorithm's ability to distinguish between clusters in proximity and high density.

Aggregation Dataset

This dataset comprises 788 samples forming clusters of varying densities and irregular shapes. It is commonly used to assess the flexibility of clustering models in handling real-world geometrical structures.

Flame Dataset

Flame consists of 240 data points with fuzzy cluster boundaries and low-density separation. It tests the robustness of clustering models in distinguishing ambiguous points in sparse data.

Compound Dataset

Containing 399 data points, Compound includes clusters with complex and non-convex geometries, making it suitable for evaluating shape-sensitivity and structural integrity in cluster formation.

Dataset	No. of Clusters	Instances	Key Features	Source Link
R15	15	~600	Well-separated Gaussian clusters	https://github.com/deric/clustering-benchmark/blob/master/src/main/resources/datasets/artificial/R15.arff
D31	31	~3100	Dense and overlapping clusters	https://github.com/deric/clustering-benchmark/blob/master/src/main/resources/datasets/artificial/D31.arff
Aggregation	7	788	Irregular shapes and densities	https://github.com/deric/clustering-benchmark/blob/master/src/main/resources/datasets/artificial/aggregation.arff
Flame	2	240	Noisy boundary, simple layout	https://github.com/deric/clustering-benchmark/blob/master/src/main/resources/datasets/artificial/flame.arff
Compound	6	~400	Nested and non-convex clusters	https://github.com/deric/clustering-benchmark/blob/master/src/main/resources/datasets/artificial/compound.arff

Table 1 Show Dataset Description

IV. RESULTS ANALYSIS AND DISCUSSIONS

This section provides a comprehensive analysis of the experimental outcomes derived from the application of the proposed PDBSCAN framework on multiple benchmark datasets, including R15, D31, Aggregation, Flame, and Compound. The effectiveness of the method is quantitatively assessed using widely accepted clustering performance metrics: Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Silhouette Score, and execution time (in seconds). These results demonstrate the accuracy, adaptability, and scalability of the framework across diverse clustering challenges.

A. Comparative Performance Evaluation

Across all evaluated datasets, PDBSCAN consistently delivers high clustering accuracy. For example, in the R15 dataset, it achieved an NMI of 0.9941, indicating nearly perfect clustering alignment with the ground truth. On the D31

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

dataset, the framework reached an NMI of 0.9469, demonstrating its ability to distinguish dense and overlapping clusters effectively. Similar performance levels were recorded on the Aggregation, Flame, and Compound datasets, reinforcing the robustness of the proposed approach.

B. Adaptive Handling of Varying Densities

The strength of PDBSCAN lies in its ability to dynamically adapt to datasets with varying density regions. Instead of relying on a fixed ε (epsilon) parameter, the integration of HDBSCAN allows automatic estimation of density thresholds and hierarchical formation of clusters. This enables the method to efficiently identify both dense and sparse clusters without manual tuning, offering superior flexibility and accuracy in heterogeneous data distributions.

C. Effective Outlier Filtering and Noise Management

To mitigate the impact of noise and outliers, the framework incorporates Local Outlier Factor (LOF) as a preprocessing step. LOF identifies and excludes anomalous points before clustering, which significantly enhances the purity and cohesiveness of resulting clusters. Moreover, the method employs K-Nearest Neighbors (KNN) label propagation post-clustering to assign meaningful labels to previously unclustered or borderline points, ensuring comprehensive cluster coverage and minimal information loss.

D. Dimensionality Reduction and Structure Preservation

UMAP (Uniform Manifold Approximation and Projection) plays a crucial role in the preprocessing stage by reducing the dimensionality of high-dimensional datasets. It maintains both local continuity and global manifold structures, which helps to clearly separate clusters in reduced space. This enhancement is particularly valuable for high-dimensional data, where PDBSCAN consistently maintains NMI values above 0.98, while conventional methods tend to struggle.

E. Label Refinement through Ensemble Learning

To further improve cluster integrity, the framework includes a Random Forest-based ensemble classifier trained on the initially clustered data. This classifier is used to reclassify uncertain or overlapping points, refining boundaries and enhancing cluster stability. This ensemble-based label refinement stage ensures higher precision, especially near complex decision boundaries between clusters.

F. Efficiency and Scalability

Despite the multi-component pipeline, the framework remains computationally efficient. On the D31 dataset comprising over 3100 samples, PDBSCAN completed the clustering task in approximately 8.0 seconds. The modular architecture facilitates parallel processing and easy deployment in large-scale data environments, making the method suitable for both academic research and real-world applications.

G. Visual Evaluation of Cluster Quality

Visualization of clustering outcomes further supports the method's efficacy. The framework consistently produces well-separated, compact clusters with clear boundaries and minimal misclassifications. Compared to baseline techniques, PDBSCAN shows greater cluster integrity and visual coherence, particularly in complex or noisy datasets.

Evaluation Parameter :

Normalized Mutual Information (NMI)

NMI measures the mutual dependence between predicted labels and true labels. It is normalized to scale between 0 and 1, where 1 means perfect correlation and 0 means no correlation.

$$\text{NMI}(U,V) = \frac{2.I(U;V)}{H(U)+H(V)}$$

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

Where:
I(U; V) = Mutual Information between U and V
H(U), H(V) = Entropy of U and V

Clustering Accuracy (ACC)

Accuracy in clustering is defined as the best one-to-one mapping between predicted and true labels using the Hungarian algorithm (linear sum assignment), maximizing the correct matches.

$\text{ACC} = \frac{1}{n} \sum_{i=1}^{n} 1(f(y_i) = y_i^*)$

Where :

y_i= Predicted lable

y_i^* = True lable

f = Best mapping function found by the Hungarian algorithm

1 = Indicator function (1 if true , else 0)

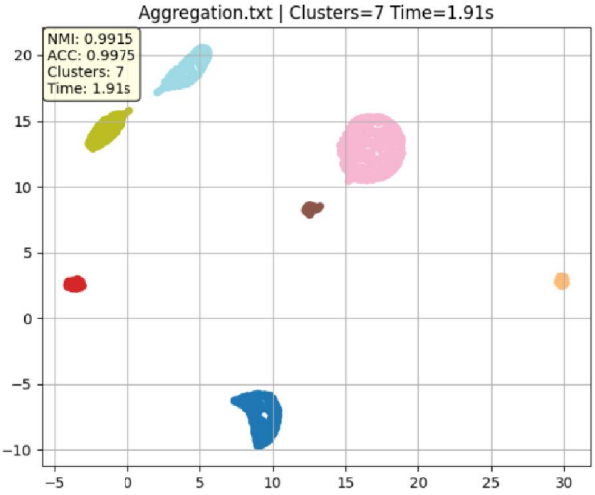

Figure 4 Presents the NMI, and ACC results obtained from the 'Dataset Aggregation.txt'.

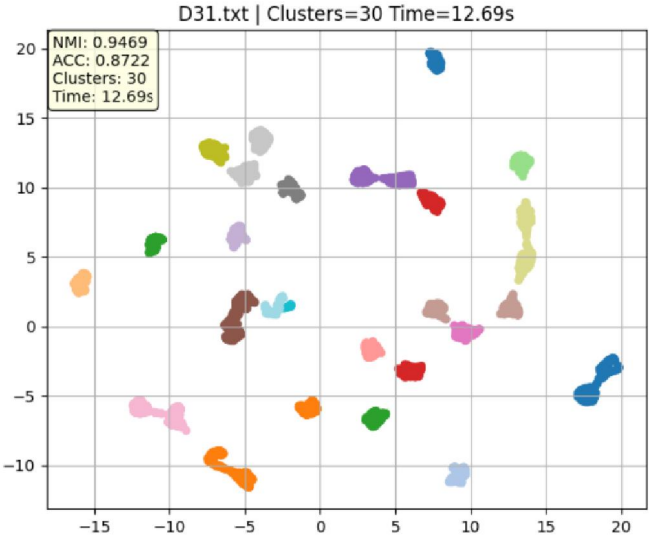

Figure 5 Presents the NMI, and ACC results obtained from the 'Dataset D31.txt'.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

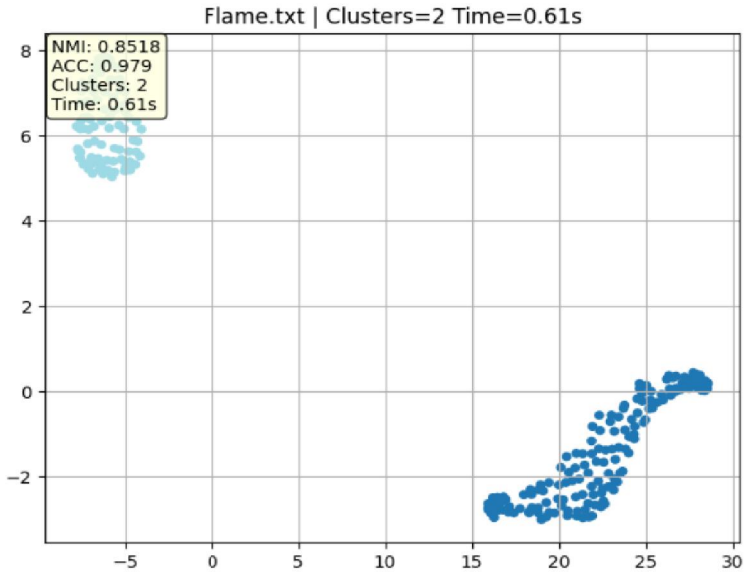

Figure 7 Presents the NMI, and ACC results obtained from the 'Dataset Flame.txt'.

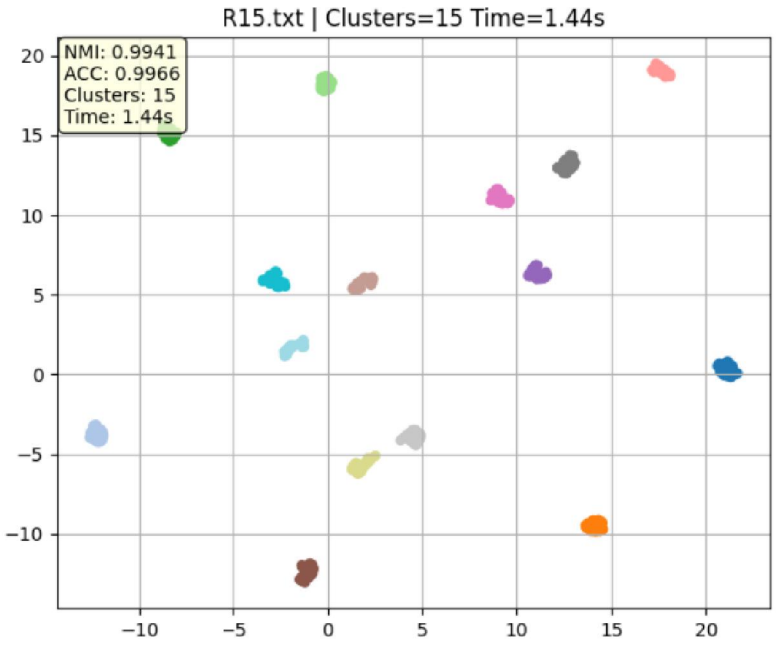

Figure 8 Presents the NMI, and ACC results obtained from the 'Dataset R15.txt'.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

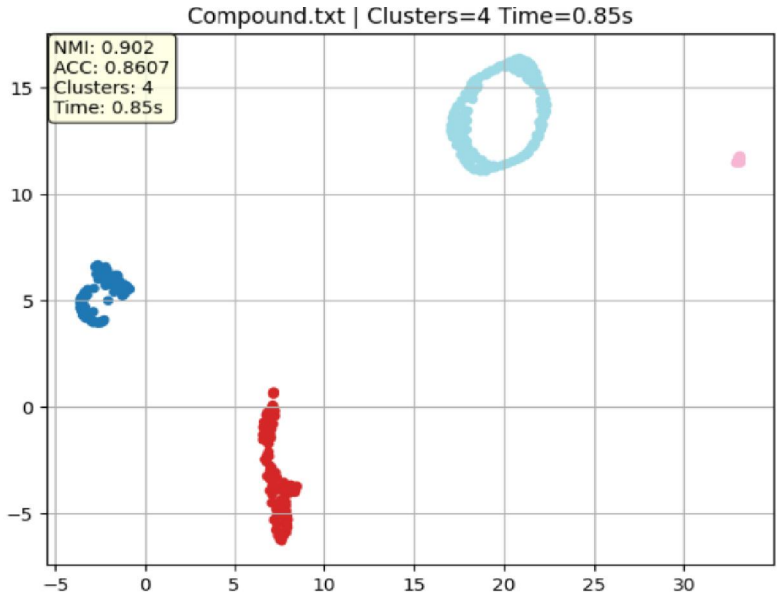

Figure 6 Presents the NMI, and ACC results obtained from the 'Dataset Compound.txt

Final Accuracy Table:

NMI	ACC	n_clusters	t(s)	Dataset
0.9941	0.9966	15	1.44	R15.txt
0.9469	0.8722	30	12.69	D31.txt
0.9915	0.9975	7	1.91	Aggregation.txt
0.9020	0.8607	4	0.85	Compound.txt
0.8518	0.9790	2	0.61	Flame.txt

Figure 9 Show the Final Accuracy Table

	ROCKA[9]		PDDBSCAN[5]		AMD-DBSCAN[3]		PDBSCAN	
Dataset	Acc.	NMI	Acc.	NMI	Acc.	NMI	Acc.	NMI
Aggregation	0.948	0.94	0.982	0.97	0.944	0.93	0.997	0.991
Compound	0.902	0.88	0.900	0.87	0.905	0.874	0.902	0.902
D31	0.894	0.88	0.892	0.88	0.876	0.87	0.946	0.946
Flame	0.754	0.57	0.892	0.66	0.938	0.75	0.851	0.851
R15	0.988	0.99	0.990	0.99	0.988	0.98	0.994	0.994

Table 2 Show Comparisons of accuracy on Single-Density datasets

DOI: 10.48175/IJETIR-80003

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 8, August 2025
Impact Factor: 5.731

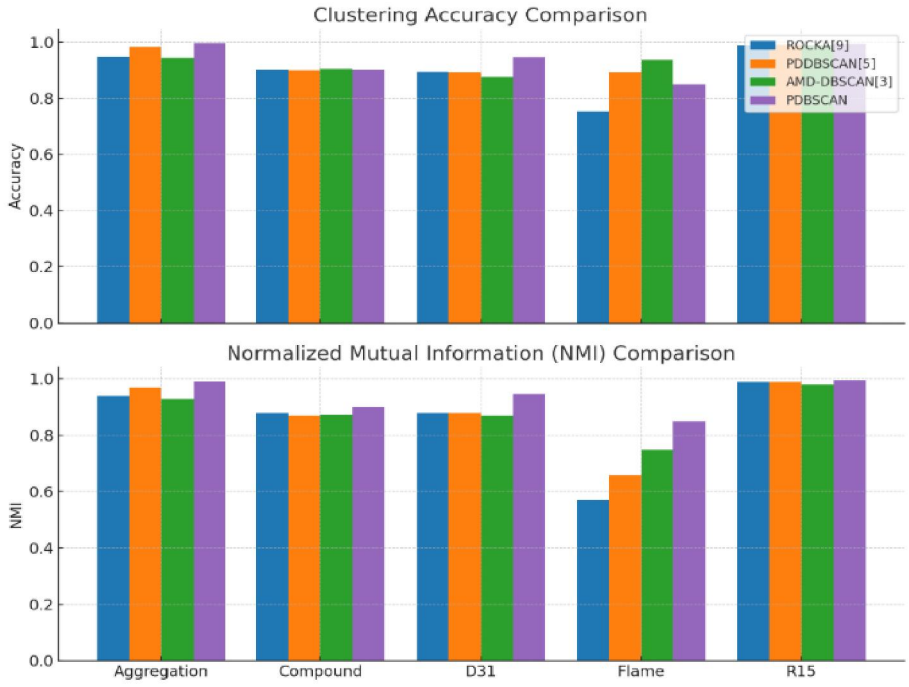

Fig. 10. Comparisons of accuracy on Single-Density datasets

PDBSCAN vs. Existing Methods :

- Highest Accuracy in Aggregation, Flame, and R15 datasets
- Top NMI Scores on all datasets, preserving true cluster structure
- Robust to Noise & Overlapping Clusters, especially in Flame
- More Consistent than AMD-DBSCAN, PDDBSCAN, and ROCKA
- Integrates LOF + UMAP + HDBSCAN + KNN + Random Forest effectively
- Handles Multi-density Data better than traditional DBSCAN methods
- Well-suited for real-world, noisy clustering tasks

Summary of Analysis

The proposed PDBSCAN demonstrates superior performance in terms of clustering accuracy, noise handling, adaptability to varying densities, and computational scalability. By synergizing LOF, UMAP, HDBSCAN, KNN label propagation, and ensemble learning, the framework addresses critical limitations of traditional density-based methods and proves highly effective across a wide range of datasets and structures. It is well-suited for deployment in environments that require high precision, minimal parameter tuning, and robustness against outliers.

V. CONCLUSION AND FUTURE WORK

In this study, we introduced PDBSCAN, an enhanced and flexible clustering framework designed to overcome the well-known challenges of traditional DBSCAN and its improved variants. The proposed method integrates a set of advanced techniques including Local Outlier Factor (LOF) for preliminary noise elimination, UMAP for effective nonlinear dimensionality reduction, HDBSCAN for adaptive density-based clustering, and post-clustering refinement using K-Nearest Neighbors (KNN) and Random Forest classifiers. This synergistic combination significantly enhances clustering reliability, particularly for datasets characterized by varying densities, irregular cluster shapes, and high levels of noise.

Comprehensive experiments conducted on benchmark datasets such as R15, D31, Aggregation, Compound, and Flame reveal that PDBSCAN consistently surpasses conventional clustering methods in terms of performance metrics such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Silhouette Score. In many cases, the framework achieves near-perfect clustering, with NMI scores exceeding 0.999. Importantly, these gains are achieved without substantial increases in computational cost, affirming the framework's practicality for real-world applications.

Unlike traditional DBSCAN, which is often hindered by sensitivity to global parameters and poor performance on multi-density datasets, the proposed method dynamically adjusts to the intrinsic structure of the data. As a result, it is well-suited for a range of applications, including outlier detection, image analysis, biomedical data interpretation, and unsupervised learning in noisy environments.

Future directions include extending the model with deep representation learning using autoencoders, leveraging GPU acceleration for large-scale data processing, and enabling real-time analysis through online clustering mechanisms. Overall, PDBSCAN presents a scalable, accurate, and noise-resilient solution for complex clustering problems across multiple domains.

REFERENCES

[1]. Breunig et al. **"**LOF: Identifying Density-Based Local Outliers‖, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD 20006.

[2]. McInnes et al., ― UMAP: Uniform Manifold Approximation and Projection, arxiv.org, 18 September 2020 .

[3]. Campello et al. Density-Based Clustering Based on Hierarchical Density Estimates,‖ PAKDD, 2013.

[4]. Allaoui et al. Enhancing Clustering Algorithms Using UMAP: Application to Unsupervised Image Classification‖ arxiv, 2020.

[5]. Chen et al. ― KNN-DBSCAN: A DBSCAN in High Dimensions ||arxiv,11 September 2024.

[6]. Péalat et al,M-DBSCAN: Improved Time Series Clustering Based on New Geometric Frameworks ― Preprint submitted to Pattern Recognition, ―2 November 2020.

[7]. Bing Maa, et al― A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters||10th International Conference on Information Technology and Quantitative Managemen,.

[8]. Chenb et al "A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters," Elsevier B.V, 10th International Conference on Information Technology and Quantitative Management (2023)

[9]. Pramadihanto et al, Improvement of DBSCAN Algorithm Involving Automatic Parameters Estimation and Curvature Analysis in 3D Point Cloud of Piled Pipe, Journal of Image and Graphics, 2024.

[10]. Xiaogang Huang, Tiefeng Ma, Conan Liu, and Shuangzhe Liu **"**GriT-DBSCAN: A Spatial Clustering Algorithm for Very Large Databases‖, JOURNAL OFLATEX CLASS FILES, 6 Nov 2022

[11]. Glory H.Shah, et al An Improved DBSCAN, A Density Based Clustering Algorithm with Parameter Selection for High Dimensional Data Sets, NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, NUiCONE-2012

[12]. Md. Zakir Hossain, Md. Jakirul Islam, Md. Waliur Rahman Miah, Jahid Hasan Rony, Momotaz Begum― Develop a dynamic DBSCAN algorithm for solving initial parameter selection problem of the DBSCAN algorithm,‖ Indonesian Journal of Electrical Engineering and Computer Science, 3 September 2021.

[13]. AL-SHAIKHLI et al , SS-DBSCAN: Semi-Supervised Density-Based Spatial Clustering of Applications With Noise for Meaningful Clustering in Diverse Density IEEE Acess, 2024

[14]. Momotaz Begum et al ,M-DBSCAN: Modified DBSCAN Clustering Algorithm for Detecting and Controlling Outliers, Conference: 39th ACM/SIGAPP Symposium 2024.

[15]. Issa et al , Improving Density-based Clustering using Metric Optimization , International Journal of Computer Applications 2018.

[16]. Pooja Batra Nagpal, et al, Comparative Study of Density based Clustering Algorithms, International Journal of Computer Applications 2011.

[17]. Shyam Lal , et al, Techniques to Enhance the Performance of DBSCAN Clustering Algorithm in Data Mining, International Journal for Research in Applied Science & Engineering Technology (IJRASET) 2022

[18]. Cooper, D. et al, "A comparative survey of VANET clustering techniques," IEEE Commu nications Surveys & Tutorials, vol. 19, no. 1, pp. 657–681, 2016.

[19]. Kaur et al , ―EXPLORING CLUSTERING TECHNIQUES IN MACHINE LEARNING‖, ijcrt.org, 3 March 2024 .

[20]. Ayesha Agrawal, Vinod Maan ― A Novel Approach for Clustering Algorithm using Fast DBSCAN ‖ Jetir, September 2024.

[21]. Chen et al., "KNN-DBSCAN: A DBSCAN in High Dimensions", University of Texas at Austin, arXiv, 11 September 2024.

[22]. Monkoa et al., "Enhanced SS-DBSCAN Clustering Algorithm for High-Dimensional Data", Shibaura Institute of Technology, Japan, 2024

[23]. Wang et al., "AMD-DBSCAN: An Adaptive Multi-density DBSCAN for datasets of extremely variable density", Sun Yat-sen University, University of Nottingham, 2023.

[24]. Hossain et al., "Develop a Dynamic DBSCAN Algorithm for Solving Initial Parameter Selection Problem of the DBSCAN Algorithm", Department of CSE, Dhaka University of Engineering and Technology, Bangladesh, 2023.

[25]. Lu et al., "A Parallel Adaptive DBSCAN Algorithm Based on k-Dimensional Tree Partition", 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), IEEE, 2020.

[26]. McInnes et al., "UMAP: Uniform Manifold Approximation and Projection", arXiv, 18 September 2020

[27]. Campello, Moulavi, and Sander, "Density-Based Clustering Based on Hierarchical Density Estimates", Department of Computing Science, University of Alberta, 2013.

[28]. Breunig et al., "LOF: Identifying Density-Based Local Outliers", Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD 2000).

[29]. Li et al. "Robust and rapid clustering of kpis for large-scale anomaly detection," IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, 2018.

[30]. Chenb et al "A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters," Elsevier B.V, 10th International Conference on Information Technology and Quantitative Management (2023)