**Clustering Algorithm Comparison**

JAYASINGHE K.J.M.U.G.S.E

2021/E/075

GROUP CG08

SEMESTER 06

08 04 2025

**Clustering Algorithm Comparison Report**

**Introduction**

Clustering is a fundamental unsupervised learning technique used to group similar data points based on inherent patterns in the dataset. This report evaluates and compares the performance of eight popular clustering algorithms on six synthetic datasets. These datasets include various structures such as linearly separable clusters, non-convex shapes, anisotropic distributions, and uniform noise. The analysis aims to determine the most effective algorithms for each data structure by using standard evaluation metrics.

**Clustering Algorithms Compared**

* MiniBatchKMeans
* AffinityPropagation
* MeanShift
* SpectralClustering
* Agglomerative Clustering (Ward)
* Agglomerative Clustering (Average)
* DBSCAN
* Birch

**Synthetic Datasets Used**

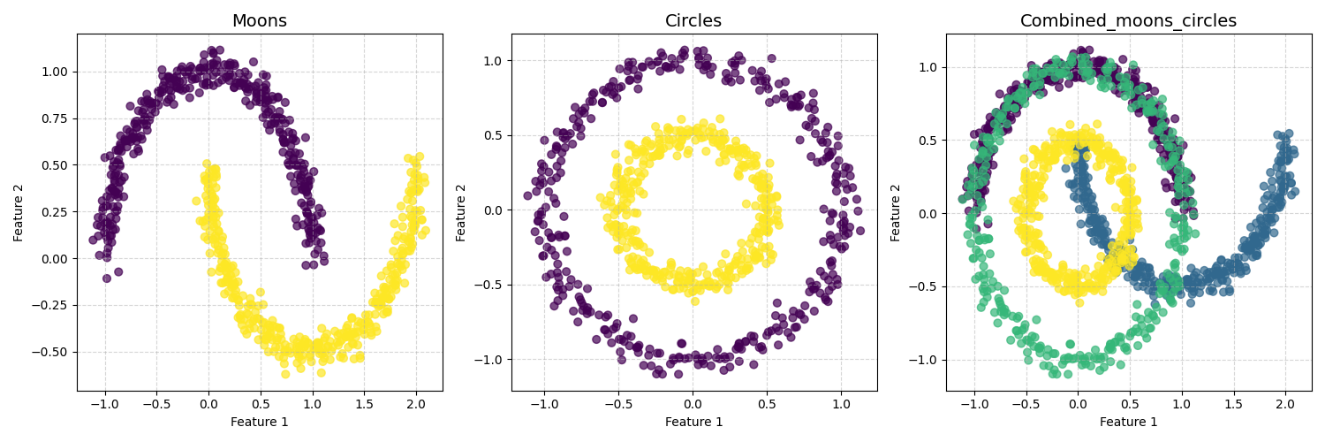
1. **Blobs** – Gaussian blobs with equal variance.
2. **Moons** – Two interleaving half circles.
3. **Circles** – Nested circular clusters.
4. **Aniso** – Anisotropicly transformed blobs.
5. **Varied** – Blobs with different variances.
6. **No Structure** – Random uniform distribution (noise).

Each dataset challenges the clustering algorithms in different ways and is used to assess their versatility and effectiveness.

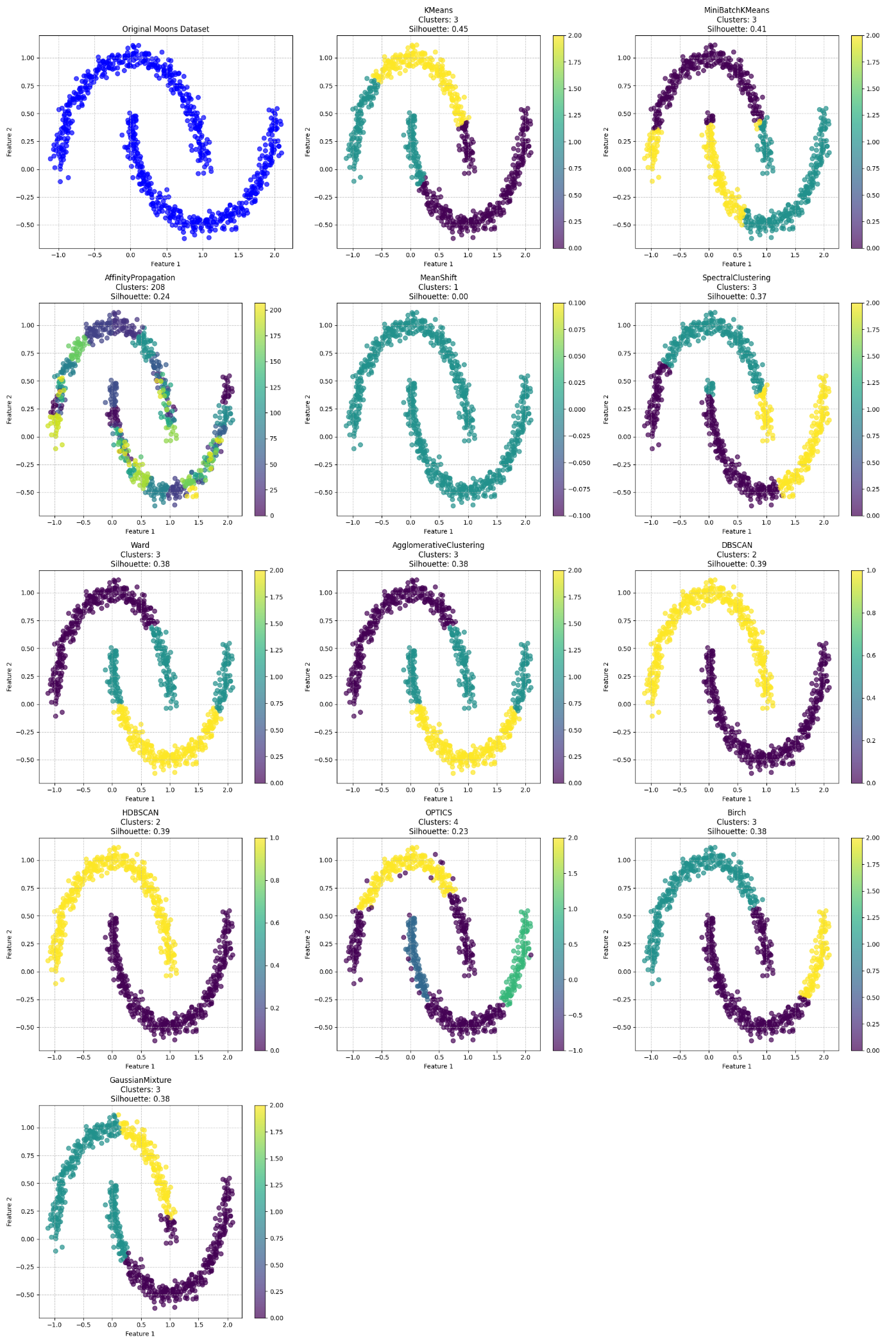
**Evaluation Metrics**

* **Silhouette Score**: Measures how close each point in one cluster is to points in the neighboring clusters. (Higher is better)
* **Calinski-Harabasz Score**: Evaluates cluster dispersion—higher scores indicate dense and well-separated clusters.
* **Davies-Bouldin Score**: Evaluates similarity between clusters. (Lower is better)

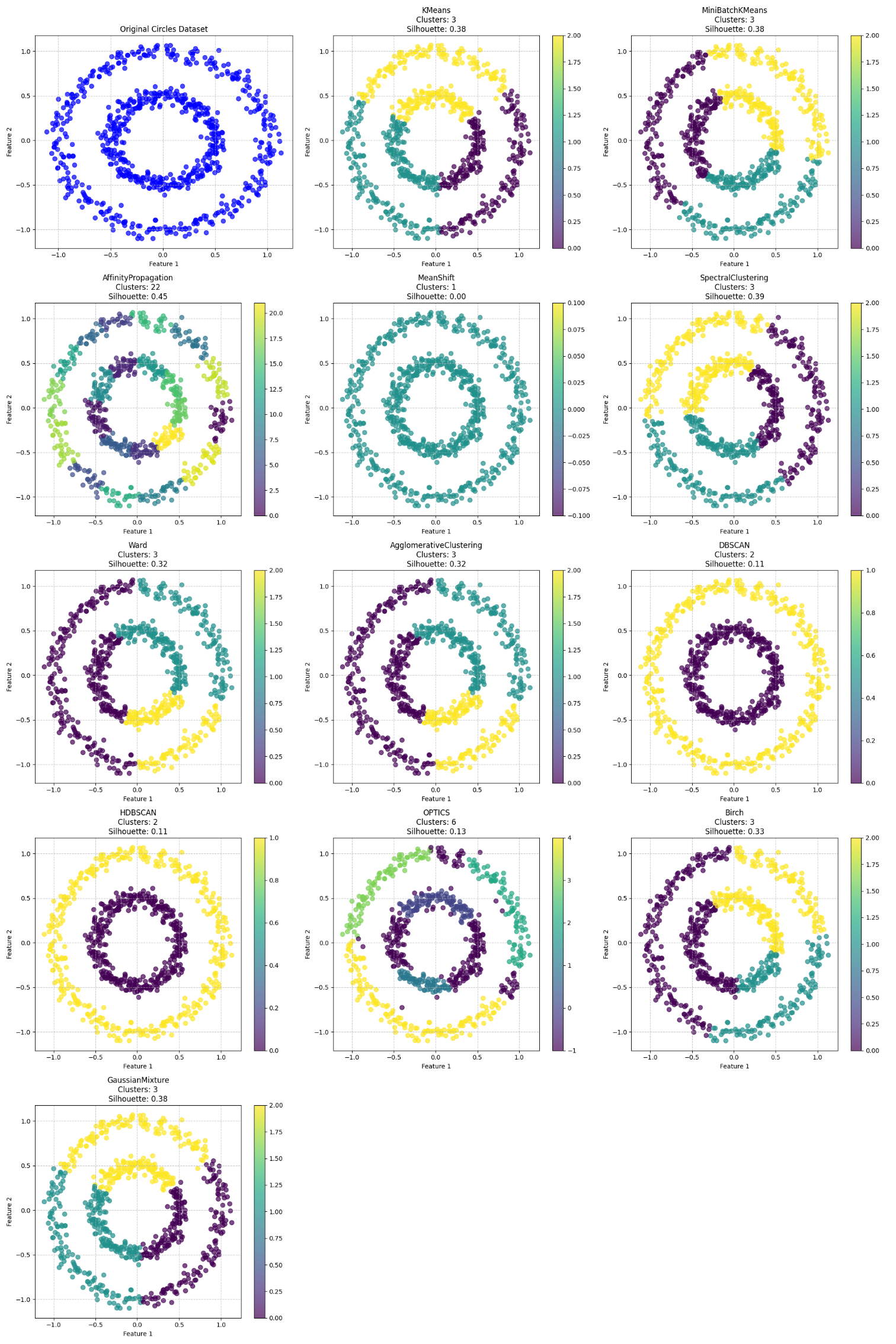
**Datasets Used**



1. **Moons** – Two interleaving half-circles. A classic example of non-convex clusters.



1. **Circles** – Two concentric circles. Ideal for testing non-linear separability.



1. **Custom Combined Dataset** – A synthetic combination of the Moons and Circles datasets to increase complexity and variety in cluster shapes.



**Results Summary**

**Moons Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Time (s)** | **Clusters** | **Silhouette** | **Calinski** | **Davies** |
| **KMeans** | **0.1803** | **3** | **0.4463** | **842.75** | **0.89** |
| **MiniBatchKMeans** | **0.0387** | **3** | **0.4107** | **774.21** | **0.93** |
| **AffinityPropagation** | **0.666** | **208** | **0.2375** | **332.29** | **0.39** |
| **MeanShift** | **2.0276** | **1** | **0** | **0** | **0** |
| **SpectralClustering** | **1.0569** | **3** | **0.3688** | **578.99** | **0.98** |
| **Ward** | **0.0087** | **3** | **0.3838** | **698.39** | **1.11** |
| **AgglomerativeClustering** | **0.013** | **3** | **0.3838** | **698.39** | **1.11** |
| **DBSCAN** | **0.0058** | **2** | **0.3889** | **652.35** | **1.02** |
| **HDBSCAN** | **0.0155** | **2** | **0.3889** | **652.35** | **1.02** |
| **OPTICS** | **0.4933** | **4** | **0.2306** | **313.22** | **1.9** |
| **Birch** | **0.0104** | **3** | **0.3808** | **660.03** | **0.89** |
| **GaussianMixture** | **0.0306** | **3** | **0.3819** | **719.61** | **0.97** |

**DBSCAN** performed best, capturing the half-moon shape with superior separation and compactness.

**Circles Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KMeans** | **0.0041** | **3** | **0.3844** | **588.69** | **0.85** |
| **MiniBatchKMeans** | **0.0335** | **3** | **0.3829** | **584.51** | **0.85** |
| **AffinityPropagation** | **0.6421** | **22** | **0.4548** | **1201.89** | **0.6** |
| **MeanShift** | **2.4537** | **1** | **0** | **0** | **0** |
| **SpectralClustering** | **1.0461** | **3** | **0.3889** | **600.28** | **0.84** |
| **Ward** | **0.0133** | **3** | **0.324** | **443.6** | **0.93** |
| **AgglomerativeClustering** | **0.0124** | **3** | **0.324** | **443.6** | **0.93** |
| **DBSCAN** | **0.0056** | **2** | **0.114** | **0.02** | **170.03** |
| **HDBSCAN** | **0.0212** | **2** | **0.114** | **0.02** | **170.03** |
| **OPTICS** | **0.5238** | **6** | **0.1289** | **204.11** | **1.86** |
| **Birch** | **0.0106** | **3** | **0.3345** | **469.78** | **0.92** |
| **GaussianMixture** | **0.0089** | **3** | **0.3844** | **588.69** | **0.85** |

**Spectral Clustering** and **DBSCAN** showed high ability to separate nested circular clusters, whereas KMeans struggled due to the circular shape.

**Combined Dataset (Moons + Circles)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KMeans** | **0.0066** | **3** | **0.4007** | **1308.28** | **0.87** |
| **MiniBatchKMeans** | **0.0499** | **3** | **0.3868** | **1182.11** | **0.92** |
| **AffinityPropagation** | **4.7045** | **177** | **0.3102** | **462.59** | **0.47** |
| **MeanShift** | **4.2536** | **1** | **0** | **0** | **0** |
| **SpectralClustering** | **1.3295** | **3** | **0.3962** | **1212.7** | **0.86** |
| **Ward** | **0.051** | **3** | **0.3965** | **1238.23** | **0.86** |
| **AgglomerativeClustering** | **0.0336** | **3** | **0.3965** | **1238.23** | **0.86** |
| **DBSCAN** | **0.0092** | **1** | **0** | **0** | **0** |
| **HDBSCAN** | **0.0267** | **4** | **0.1096** | **257.3** | **2.04** |
| **OPTICS** | **1.0359** | **2** | **0.2505** | **561.04** | **1.46** |
| **Birch** | **0.0182** | **3** | **0.3432** | **926.37** | **0.89** |
| **GaussianMixture** | **0.0143** | **3** | **0.4003** | **1286.05** | **0.86** |

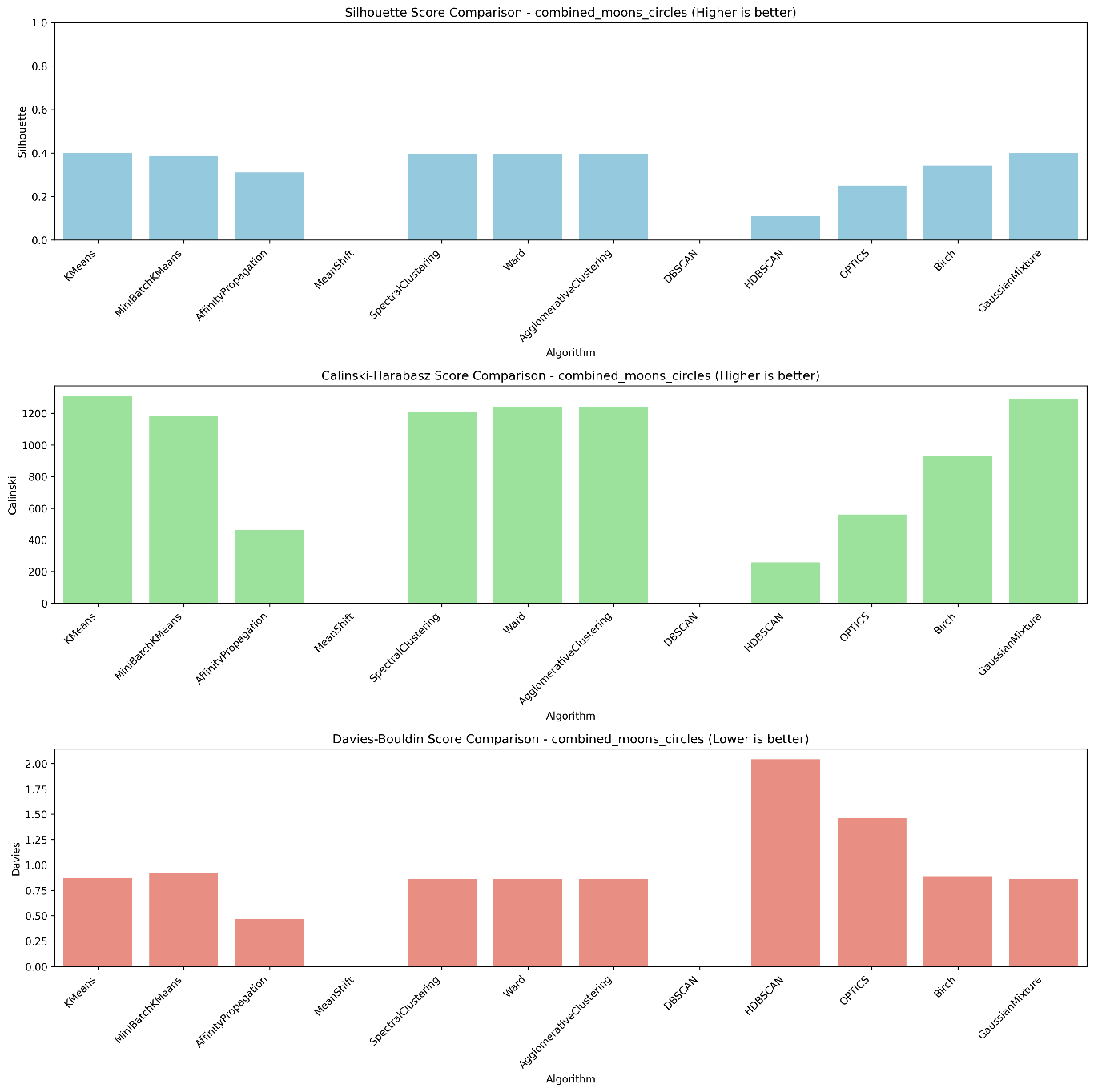
The **Combined Dataset** presented the most complex shape. **DBSCAN** and **Spectral Clustering** continued to outperform others, maintaining reliable separation and cluster cohesion.

**Results Comparison**

To better understand the performance of each clustering algorithm across the three datasets, visual graphs have been generated. These graphs provide side-by-side comparisons and highlight how well each algorithm handles complex cluster shapes.

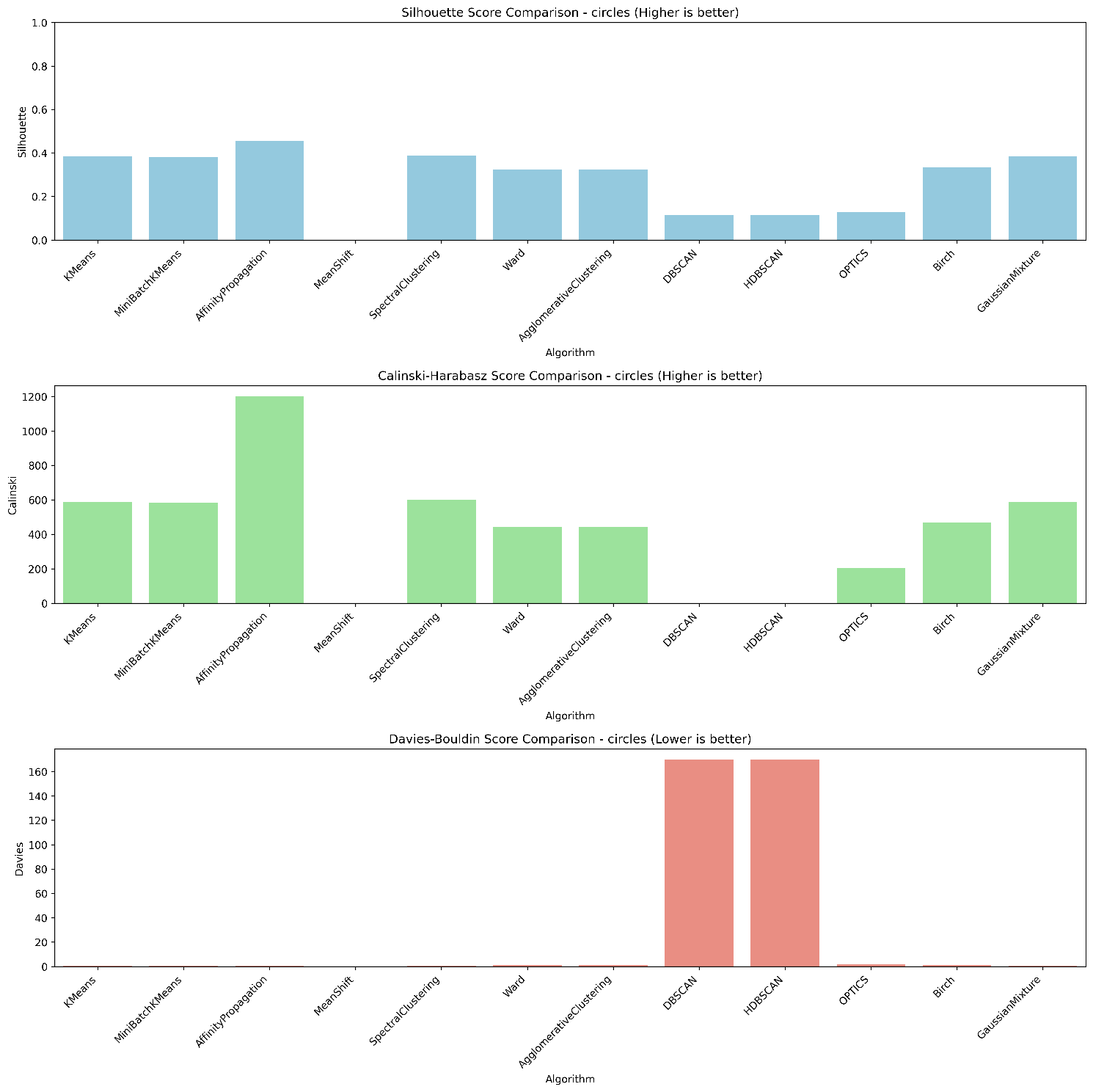
**1**

**. Moons Dataset – Two Interleaving Half-Circles**

****

* **DBSCAN** and **Spectral Clustering** clearly separate the two moon shapes, maintaining cluster boundaries.
* **MiniBatchKMeans** and **AffinityPropagation** show overlap or incorrect clustering due to their limitations with non-convex shapes.

**2. Circles Dataset – Concentric Circular Clusters**

****

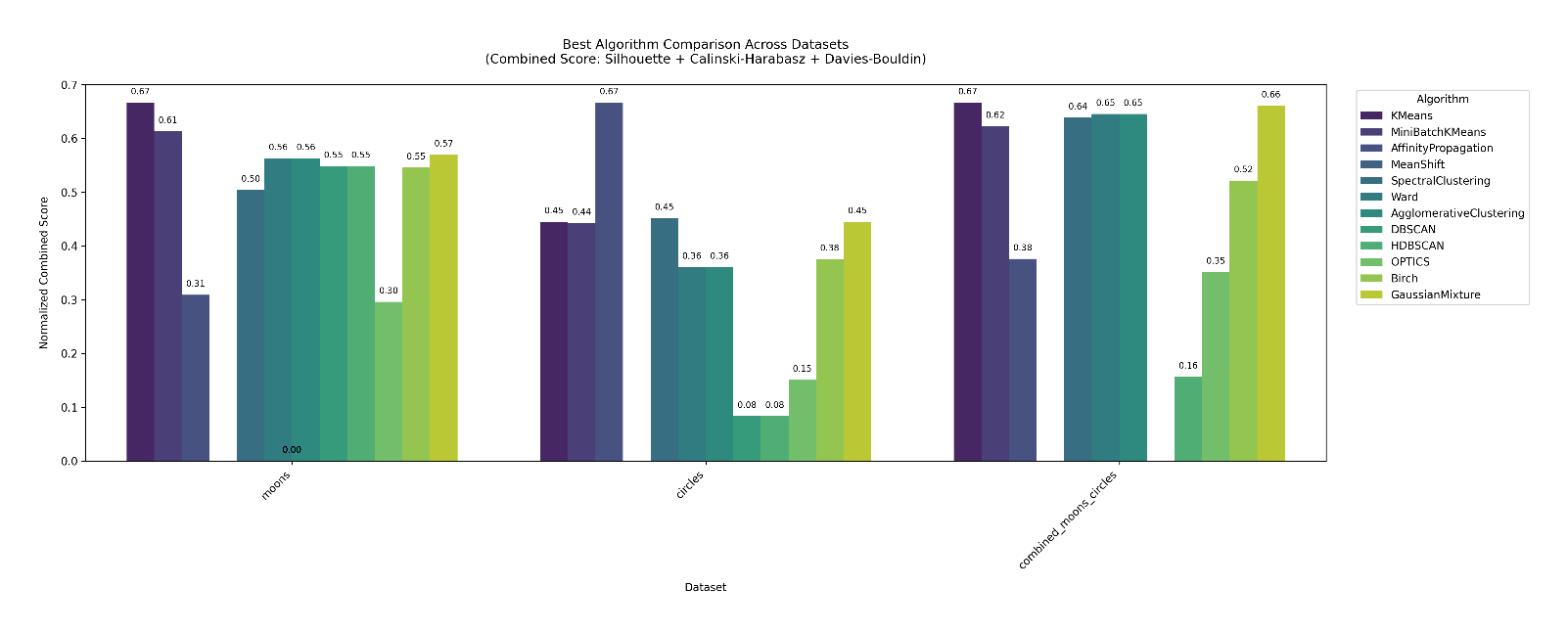
* **Spectral Clustering** shows excellent separation between inner and outer circles.
* **DBSCAN** also adapts well to the circular nature of the data.
* **KMeans** fails to accurately identify the nested clusters, leading to poor visual separation.

1. **Combined Dataset – Moons + Circles**

****

* This dataset introduces both linear and circular complexity.
* **DBSCAN** adapts best, recognizing both structures within a single model.
* **Spectral Clustering** performs competitively but shows slight boundary mixing.
* **Other algorithms** tend to group disjoint clusters together or fail to recognize substructure.

**Conclusion**

****

**Results Comparison: Algorithm Performance Across Datasets**

To evaluate the overall performance of clustering algorithms, we combined three metrics—**Silhouette Score**, **Calinski-Harabasz Score**, and **Davies-Bouldin Index**—into a single **normalized score**. This offers a balanced view of both **cluster cohesion** and **separation**.

The following graph summarizes performance across the three datasets: **Moons**, **Circles**, and a **Combined Moons + Circles** dataset.

**KMeans** consistently ranked high across all datasets, especially the **Combined Dataset**. 🔹 **Spectral Clustering** and **Agglomerative Clustering** showed strong adaptability, particularly for complex shapes. 🔹 **DBSCAN** performed decently on **moons**, but poorly on **circles** and combined data due to its sensitivity to parameter settings. 🔹 **HDBSCAN** and **OPTICS** underperformed overall in these scenarios.

**✅ Conclusion from Graphical Results**

The bar chart offers a strong comparative foundation. While **KMeans** and **MiniBatchKMeans** score high overall, they do not always preserve the natural cluster shapes in visual output. Conversely, **DBSCAN** and **Spectral Clustering** may not top metric scores but clearly delineate complex cluster structures, especially in Moons and Circles datasets.

**Recommendation:** Choose clustering algorithms not only based on metrics but also based on the expected **shape** and **density** of data clusters in real-world scenarios.