Anfossi and Chicco *Discover Computing* (2026) 29:212
https://doi.org/10.1007/s10791-026-10094-0

Discover Computing

BRIEF REPORT **Open Access**

An easy guide to the Davies-Bouldin index for unsupervised internal clustering evaluation

Alessandro Anfossi[1] and Davide Chicco[1,2*]

*Correspondence:
Davide Chicco
davide.chicco@unimib.it
[1]Università di Milano-Bicocca, Milan, Italy
[2]University of Toronto, Toronto, Ontario, Canada

Abstract
This study provides a practical guide to using the Davies-Bouldin Index (DBI), a popular metric for internally evaluating clustering results. While previous studies applied DBI across results on various datasets and generated by several algorithms, no comprehensive guidelines exist for effective use of this coefficient. We systematically analyze DBI's behavior through experiments on both synthetic and real-world medical datasets, including results obtained through *k*-means, DBSCAN, and hierarchical clustering. Our results highlight DBI's strengths and limitations, such as sensitivity to cluster imbalance, non-compact clusters, and high-dimensional data. Based on these findings, we propose actionable recommendations and best practices for unsupervised machine learning researchers, including optimal algorithm choices and parameter settings. This study bridges the gap between theoretical knowledge and practical application, offering a clear *manual of use* for the Davies-Bouldin Index in diverse clustering tasks.

Keywords Davies-Bouldin index, Clustering evaluation, Clustering algorithms, *k*-means, Clustering metrics, Unsupervised machine learning

1 Introduction

Clustering is a data analysis technique that groups data points into clusters according to their similarity, typically without predefined labels. Different algorithms, such as hierarchical clustering [1], DBSCAN [2], and k-means [3], produce different cluster structures depending on their assumptions and parameters. Evaluating the quality of these results is therefore essential. Clustering indices are quantitative metrics designed to assess how well a dataset has been partitioned, helping compare methods and choose suitable parameters, such as the number of clusters. Some examples are the Dunn Index [4], the Silhouette Index [5], the Califiski-Harabasz Index [6] and the Gap Statistic [7]. Despite this variety, practitioners still lack clear practical guidance on how to interpret and apply these measures in real scenarios.

Even if several metrics exist, no consensus on which one performs better has been reached in the scientific community yet. In this article we will approach the Davies-Bouldin Index [8]. Introduced by David L. Davies and Donald W. Bouldin in 1979, the index provides a measure of the quality of clustering by assessing the trade-off between intra-cluster compactness and inter-cluster separation, as in the examples given. Its simplicity and efficiency make it a popular choice for evaluating the performance of clustering algorithms.

We conducted a literature review to evaluate how the Davies-Bouldin Index (DBI) is currently applied in the biomedical field and how it is generally regarded as a clustering evaluation index. We focused on comparisons with other indices, such as the Silhouette and Dunn indices. We sourced articles from databases such as Google Scholar, PubMed, and Web of Science [9], by considering all the biomedical informatics studies involving the DBI coefficient published in reliable journals. We summarized the gathered information in Tables 1 and 2, if one is interested the material is freely available online. Articles were selected in sequential order until the target quota was reached, provided they met the predefined inclusion criteria, namely a minimum citation threshold, adequate journal reputation within the field, and relevance to the review topic.

In Table 1 the majority of articles employed the Davies-Bouldin Index (DBI) either as a clustering evaluation metric [10–14] or as a reference for dimensionality reduction [15]. Although other uses emerged [16, 17], they were too infrequent (one in twenty articles) to include in the table. Interested readers can find some noteworthy examples in the supplementary materials [18–20].

In most cases, the DBI was compared with the Dunn, Silhouette, and Calinski-Harabasz indices [21–26]. Many of the articles listed in Table 1 are methodological studies focused on evaluating index performances, which explains the frequent use of artificial datasets [27–29].

We also documented the types of inter-cluster measurements used. This aspect often goes unreported and was typically limited to a simple mean distance from the centroid. In many instances, researchers did not consider whether alternative metrics might be more appropriate for their specific datasets.

Table 2 summarizes information regarding twenty biomedical studies that employed the Davies-Bouldin Index. While the index found a range of uses, a great majority focused on clustering evaluation [30–33]. Between the subjects were a wide range of clinical situations with a relative majority of studies on carcinoma cases [34–37].

In almost half the studies the Davies-Bouldin Index was used alone [38–41] without the aid of other instruments, while in almost the other half it was confronted mainly with the Silhouette Index [42–45].

Table 1 Recap of the main findings of the scientific articles collected and studied

Technique	%	Indices Confronted	%	Datasets Used	%	Intracluster measurement used	%
Dim. Reduc	10	None	10	mixed	25	Variance	5
DBSCAN	10	I	15	artificial	35	not specified	20
SOM	15	Calinski-Harabasz	25	real	40	Centroids	75
k-means	40	Silhouette	30				
others	50	Davies-Bouldin Index-variation	30				
		Dunn Index	60				
		others	60				

Data referred to twenty different methodological studies on clustering indices. The percentage refers to the appearances per article. For example, "10" in the percentage column means it appeared in 10 percent of the 20 articles

Table 2 Use of DBI in the biomedical field: The table summarizes the use of Davies-Bouldin Index in the biomedical field, with percentages indicating the frequency of occurrences

Pathologies	%	Biomedical Problem	%	Cluster Scores (Per Appearance)	%
Healthy	20	Improve SEMG/EMG/MES signal classification (including spike sorting)	30	solo	45
Wide range of cancer types and subtypes	15	Improve diagnostics in respective field	30	Silhouette (SI)	40
Amputees/Upper Limb debilitated	10	Gene analysis for cancer correlation	15	Elbow method	15
Acute heart failure and comorbidities	10	Finding patterns in anamnesis to foresee outcome	15	Calinski-Harabasz (CHI)	15
Hypertension	5	Others	10	Jaccard index (JI)	5
Breast cancer	5			C-index (Hubert)	5
Burnout and Depression	5			Distance ratio (DRI)	5
Dyslexia	5			Bayesian Information Criterion (BIC)	5
Colorectal carcinoma	5			Dunn Index	5
Chronic pain	5			Stability Index	5
Diabetes and comorbidities	5			Fisher's Linear Discriminant Index (FLDI)	5
Parkinson disease	5				
Data from a lab rat	5				

Use Made of the Indices	%	Type of Data*	%
Clustering evaluation	60	SEMG/EMG	25
Feature space performance	25	Lab measurements of phenotypic expressions or symptom expressions	25
Dimension reduction	5	Omic and Genomic data	20
Harmonisation	5	Questionnaires (symbolic)	15
Template creation	5	Radiomic features	10
		Extracellular action potential (EAP)	5
		Myoelectric signals (MES)	5
		Blood pressure features	5

*Some articles use mixed types of data from these categories

We found a great variety in the type of information that was elaborated and the objectives of the studies [46–49].

Most importantly those articles were randomly selected among tens of others, and between those we did not find a guide to the Davies-Bouldin Index which we think can be the main contribution of this article.

Overall the articles had a positive perception of the index [25, 44].

Most importantly, although numerous articles employ the Davies-Bouldin Index, we did not identify a methodological guide specifically focused on its practical interpretation and limitations. Existing surveys and comparative studies typically analyze multiple clustering validity indices at a high level, whereas the present work provides a focused and practice-oriented analysis of DBI. This article aims to guide the reader from the mathematical formulation of the index to its practical use, combining analysis of implementations with experiments on both artificial and real-world datasets.

In this study, we analyze which existing implementations in the R ecosystem are most appropriate for computing the Davies-Bouldin Index, providing both theoretical justification and practical criteria that can guide the selection of analogous implementations in other environments. We then present illustrative examples on both synthetic and real-world datasets, designed to show how the mathematical structure of the index manifests across different scenarios. The main analysis is developed in the theoretical part of the Methods section, while the Discussion highlights how the key behaviours observed in the experiments translate into practical recommendations for interpreting DBI values and informing clustering choices. The paper is primarily intended for applied researchers and practitioners who use clustering validation indices in practice, while also offering methodological insights for readers interested in the theoretical behaviour and limitations of DBI.

2 Methods

2.1 Study design and methodological framework

We documented each step of this study, following good practices as stated in cited literature [50]. Every information regarding it is public and accessible (Additional information). We used a seed for each generation of a dataset for reproducibility.

We developed the experiments in R, and we organized the code and data repository according to best practices [51, 52]. We logged regular updates in a scientific diary, documenting experimental decisions and observations. Our whole software codebase is available on GitHub for transparency and reproducibility [53], and updates were committed systematically to maintain a versioned and accessible repository structure.

The methodological workflow was structured to investigate the Davies-Bouldin Index (DBI) from complementary perspectives, moving from theoretical analysis to controlled simulations and finally to real-world applications.

First, we provide a formal definition of DBI and discuss theoretical limitations. Second, we analyze implementation aspects in the R ecosystem to ensure reproducible and consistent computation of the index. Third, we design synthetic experiments aimed at isolating specific behaviours of DBI under controlled conditions, including centroid-related issues and high-dimensional scenarios. Finally, we evaluate DBI across multiple clustering algorithms on real biomedical datasets to assess its practical behaviour in applied settings.

This multi-stage design allows us to connect theoretical properties, simulation-based insights, and empirical observations into a coherent practical guide.

2.2 DBI formulation and theoretical analysis

The Davies-Bouldin Index is defined as:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} D_i \tag{1}$$

(lowest and best value: 0; highest and worst value: $+\infty$)

where K is the number of clusters, and D_i refers to cluster number i

$$D_i \equiv \max_{j \neq i} D_{ij} \tag{2}$$

calculated as:

$$D_{ij} = \frac{S_i + S_j}{M_{ij}} \tag{3}$$

Here, M_{ij} represents the similarity measure between clusters i and j, and S_i is the dispersion of cluster i. The goal is to minimize DBI to achieve an optimal clustering configuration.

The dispersion S_i of cluster i is defined as:

$$S_i = \left(\frac{1}{T_i} \sum_{l=1}^{T_i} \|\mathrm{X}_l - \mathrm{A}_i\|^p \right)^{\frac{1}{p}} \tag{4}$$

where: T_i is the number of points in cluster i, X_l represents each point in cluster i, A_i is the centroid of cluster i, and p is a parameter that determines the distance metric.

The similarity measure M_{ij} between clusters i and j is defined as:

$$M_{ij} = \left(\sum_{k=1}^{n} \|a_{ik} - a_{jk}\|^p \right)^{\frac{1}{p}} \tag{5}$$

where: a_{ik} and a_{jk} are the k-th components of the centroids of clusters i and j, n is the dimensionality of the data points, and p is the same parameter used in the dispersion measure, defining the distance metric.

DBI from Eq. 1 has a value between 0 and $+\infty$ where 0 is the best possible outcome and $+\infty$ is the worst.

This value is the average result of the D_i's (Eq. 2) which is like an evaluation of the individual clusters. These individual evaluations depend on whether the cluster is close to a very sparse cluster or if in general is too close to another cluster. If the individual evaluations are "good" then probably the clusters have very distant centroids and vice versa.

For an easier interpretation of the behaviour of DBI, we also introduced a normalized version of it: The Davies-Bouldin Index is defined as:

$$normDBI = \frac{1}{1 + DBI} \tag{6}$$

(lowest and worst value: 0; highest and best value: 1)

We employed this normalized variant only in a subset of experiments reported in the supplementary material. The normalization was introduced solely to place DBI on a bounded scale, allowing visual comparison with bounded indices such as Silhouette. It should be interpreted as a visualization aid rather than as a new clustering validity measure, and it does not modify the underlying DBI values used in our analysis. Consequently, all methodological conclusions remain unchanged when using the standard DBI, and the normalization should not be interpreted as enabling direct comparisons across different datasets.

2.2.1 Illustrative examples of DBI's theoretical limitations

The two main flaws of the Davies-Bouldin Index are its reliance on centroids to measure cluster separation [14] and the way D_i (Eq. 2) is calculated by selecting the worst-case scenario $\max_j D_{ij}$ for each cluster. This often results in clusters having the same local evaluation due to symmetry; for example, two clusters with centroids very close to each other will likely have the same D_i.

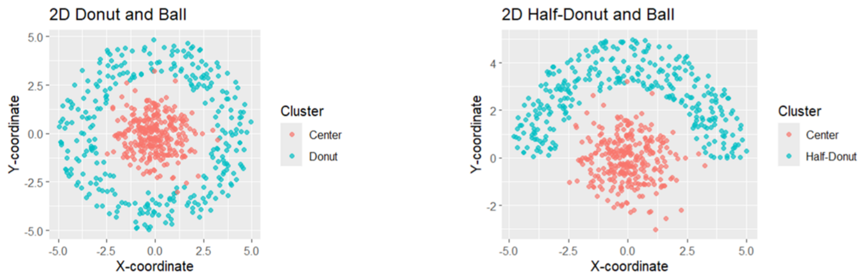
This occurs due to Eq. 5, which measures the distance between two clusters as the distance between their centroids, and Eq. 3, which shows that the ratio is derived from the sum of the two dispersions (Eq. 4) and the proximity of the two centroids.

Figure 1a illustrates a clear example of the "centroids problem" for the Davies-Bouldin Index. The index assigns a score of approximately 64, which is not a strong result. Since we will present examples in different spaces, we encourage readers to consider the cluster sizes, as they factor into the index calculation. In Fig. 1a, both clusters are contained within $[-5, 5] \times [-5, 5]$).

In Fig. 1b, the index provides a more reasonable result for a more realistic scenario. However, the score is still inadequate. In Fig. 1c, where the clustering is clearly worse than in Fig. 1b, the index assigns only a slightly higher score. Given the significantly larger interval in which the data are distributed, this small gap between evaluations should be considered almost negligible. Furthermore, in Fig. 1d, the Davies-Bouldin Index assigns an even lower score than in Fig. 1b, further demonstrating that centroid distance is not a reliable measure of cluster separation

This may not seem like a major issue at first, but we encourage the reader to keep this limitation in mind when using the Davies-Bouldin Index. This is especially important when analyzing high-dimensional datasets, where clusters may be intertwined—similar to the scenario shown in Fig. 2a. In such cases, it becomes difficult to detect these overlaps, as human intuition struggles to visualize beyond three dimensions.

A more subtle issue with the Davies-Bouldin Index is its definition of D_i as $\max_j D_{ij}$. This problem arises as the number of clusters increases. Figure 2b–d illustrate three different clusterings of the same dataset. Although all are clearly poor clusterings (assuming we define 'bad' clustering based on known criteria), the Davies-Bouldin Index



Davies-Bouldin Index: 64.799 (a) Davies-Bouldin Index: 2.166 (b)

Davies-Bouldin Index: 2.937 (c) Davies-Bouldin Index: 1.447 (d)

Fig. 1 Examples for the centroids problem: (**a**) 2D Donut and Ball; (**b**) 2D Half-Donut and Ball; (**c**) Two Highly Mixed Clusters; (**d**) Two Highly Mixed Clusters on a Line. Here we notice the behavior of the Davies-Bouldin index in cases of density-based concave clusters and in cases of mixed overlapping clusters

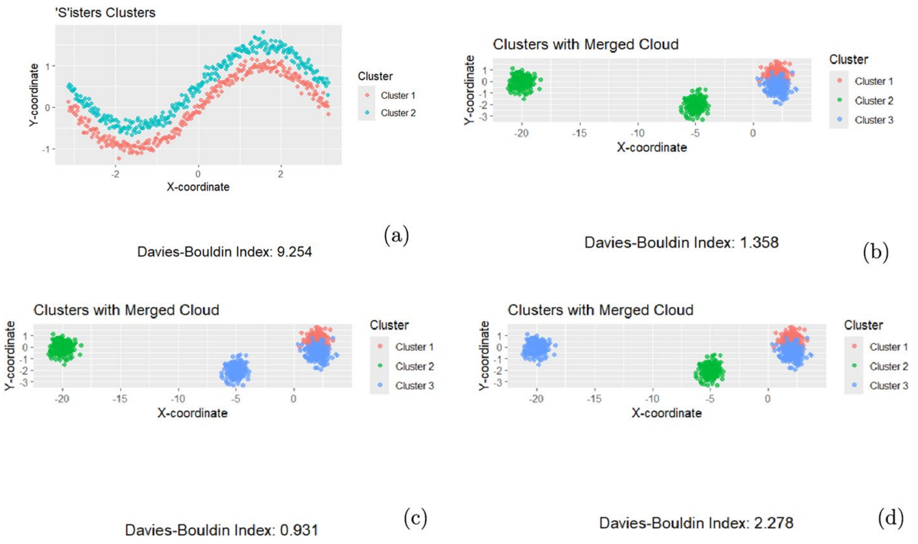


Fig. 2 Examples, second set: (**a**) ’S’isters Clusters; (**b**) Clusters with merged cloud case1; (**c**) Clusters with merged cloud case2; (**d**) Clusters with merged cloud case3. Here again we notice the behavior of the Davies-Bouldin index in cases of density-based concave clusters and in cases of mixed overlapping clusters

Table 3 Set of DBI implementations: Details of implementation choices and reasons for approval or refusal. We selected clv.Davies.Bouldin()

Implementation	Package	Monthly Downloads	Evaluation (Reason)	References
index.DB()	clusterSim	4,537	Approved (Easy to use, highly used)	[54]
clv.Davies.Bouldin()	clv	2,543	Selected (More accurate)	[55]
DB.IDX()	UniversalCVI	337	Considered (Implements clustering)	[56]
DavBou()	MGMM	309	Discarded (Unknown problem)	[57]
DB_weightedIdx()	Radviz	302	Considered (Difficult to use)	[58]
davies_bouldin_score()	ClusterStability	247	Approved (Easy to use)	[59]
db_indexR()	SOMEnv	160	Considered (Mean of the index)	[60]
check_DB()	ulrb	143	Considered (Strongly field-related)	[61]
DBindex()	chickn	51	Discarded (Package removed)	[62]
ClusterDaviesBouldinIndex()	FCPS	N/A	Considered (Reapplies index.DB)	[63]

“Discarded” means we could not investigate the script. “Considered” means we found the implementation not useful for general use. “Approved” means we made a sanity check on it. Data are referenced as of 14th May 2024, as documented at RDocumentation.org

assigns significantly different scores to each. This discrepancy occurs because a single poorly formed cluster can influence the score multiple times.

For example, in Fig. 2d, the centroids of green cluster 2 and blue cluster 3 are very close, and green cluster 2 is highly sparse. As a result, the sparsity of green cluster 2 is counted twice. In contrast, in Fig. 2b, a very sparse cluster is only penalized once. This leads to a paradox: despite the clusterings in Fig. 2c, d being very similar, their index scores differ significantly.

2.3 Implementation and computational setup in R

One of the goals of this study has been to select the most informative and effective DBI functions among the ones available for the R open source programming language (Table 3).

Table 4 Vectors associated with each example

Name	Vector
A	(1 2)
B	(2 3)
C	(3 2)
D	(2 1)
E	(1 3)
F	(3 1)
G	(1 1)
H	(3 3)

Table 5 Examples of clustering results with two clusters

Example	Cluster 1	Cluster 2
1	A C	B D
2	A B	C D
3	E E	F F
4	E F	E F
5	E G	F H
6	B E G	D F H

In the experiment we used `clv.Davies.Bouldin()` function from the `clv` package with setting "centroid" (Additional information section). We implemented the normalized DBI (Eq. 6) through our R `DBnormalize()` function. The following are the steps, we took in order to choose the implementation desired:

- Review of the monthly downloads of the packages
- Check at source code
- Check at documentation
- Test on easy examples for "approved" implementations
- Comparison of the results on all the examples resulting in a "selected" implementation

The sanity check setting is the following: Table 4 shows the vectors associated with each name, and Table 5 shows the example clusters. We applied the three implementations, `clv.Davies.Bouldin()`, `index.DB()` and `davies_bouldin_score()`, in each of the six cases and confronted them with Davies-Bouldin Index results we calculated by hand.

2.4 Synthetic experiments

2.4.1 Clouds experiment

We conducted an initial experiment generating synthetic *cloud* datasets to exemplify Davies-Bouldin Index's performance. We generated two scenarios: a single cohesive cluster (non-separated clouds) and two distinct clusters (well-separated clouds). We used the `mvrnorm()` function (MASS package [64]) to generate the clouds. We applied k-means, with two centers, and applied the index, we provided visual representations of the data distributions. Notably, points are randomly generated around fixed centers with predefined variance.

2.4.2 Manipulation of matrices of zeros and ones

The Zero–Ones experiment aimed to challenge Davies-Bouldin Index's capacity to evaluate clustering in a high-dimensionality scenario.

It consisted in constructing a $2n \times n$ matrix. First n rows of the matrix were ones [1,...,1] and the other n were zeroes [0,...,0]. Each row represented a datapoint. This matrix was divided into two clusters using k-means with $k = 2$. We then transformed a row, alternating between ones and zeros, into a random vector generated with `runif()` from R `stats` package.

Davies-Bouldin Index scores were calculated at each iteration.

The synthetic experiments described above are designed to isolate specific theoretical behaviours of DBI under controlled conditions. The real-world biomedical datasets are then used to assess whether these behaviours persist in practical, high-dimensional clinical scenarios.

2.5 Real-world evaluation across clustering algorithms

For practical insights, we tested the Davies-Bouldin Index on multiple real-world datasets of electronic health records data collected from patients with various diseases. Each dataset was clustered using three methods: k-means (with varying number of cluster centers), hierarchical clustering (varying linkage criteria), and DBSCAN (varying the epsilon and minimum point parameters). Then we calculated the Davies-Bouldin Index for each clustering result to evaluate its effectiveness in real-world clustering tasks with diverse data structures. We stored clustering configurations and results for reproducibility, and we computed statistical summaries to assess the Davies-Bouldin Index's stability and performance across algorithms (”Data and software availability”). We used five open datasets of published studies [65–69].

Neuroblastoma dataset The dataset includes 169 neuroblastoma patients (from a cohort of 275) and 13 clinical and demographic variables. It records age (in months), sex, tumor site, cancer stage, and risk classification, which help assess disease progression and severity [66].

Treatment details include autologous stem cell transplantation, radiation therapy, and surgical methods, along with the tumor's degree of differentiation. The MYCN status, a key prognostic factor linked to tumor aggression, is also documented. Follow-up duration (in months) tracks patient monitoring, while the outcome variable indicates survival status.

Collected between 2010 and 2015 at the Children's Hospital of Fudan University in Shanghai, China, the dataset is a valuable resource for studying neuroblastoma prognosis. Most variables are binary, except for follow-up duration, which is measured in months.

Diabetes type one dataset The study initially considered 77 Japanese young adult patients with childhood-onset type one diabetes. Ten patients were left out—six due to missing data and four because they were under 20, which could have affected bone metabolism results [68].

With 20 clinical characteristics recorded for each. It includes metrics such as age, diabetes duration, body mass index (BMI), total daily insulin dose (TDD), basal and bolus insulin levels, HbA1c (a marker of blood glucose control), estimated glomerular filtration rate (eGFR, indicating kidney function), and percentage of body fat. Other measures include adiponectin (a hormone related to metabolism), free testosterone levels, skeletal muscle index (SMI), grip and knee extension strength, gait speed, and osteocalcin (OC and ucOC, markers of bone turnover). Additionally, weight in kilograms, a binary insulin regimen indicator, and sex (0 for men, 1 for women) are included.The study was conducted by researchers from Tokushima University and Osaka University in Japan.

Depression and Heart Failure dataset The dataset consists of 425 patients, each represented by 15 clinical characteristics [69]. It focuses on heart failure and depression, with data including age, gender, PHQ-9 depression scores, systolic blood pressure, estimated glomerular filtration rate, ejection fraction, serum sodium, and blood urea nitrogen levels. Additionally, it tracks the etiology of heart failure, history of diabetes, and whether BNP/NT-BNP levels were elevated. Outcomes such as time from heart failure to death or hospitalization and whether the patient died or was hospitalized are also recorded. It combines binary variables to values around the tens.

Cardiac Arrest dataset The dataset consists of 416 patients who experienced cardiac arrest [67]. It contains 10 clinical characteristics recorded during the study. These include whether the patient survived (Exitus), their sex, age, whether they received endotracheal intubation, functional status, whether they had asystole, whether the cardiac arrest occurred at home, whether a bystander was present, the time in minutes, and whether the cause was cardiogenic.

We had to remove some rows due to missing information in the table at our disposal (rows: 52, 203, 221, 234, 298, 391).

Sepsis dataset The dataset contains information on 1257 patients diagnosed with sepsis, a life-threatening condition caused by the body's extreme response to infection. It includes 16 clinical characteristics, such as age, sex, severity scores (APACHE II and SOFA), inflammatory markers (CRP, WBCC, NeuC, LymC), and platelet-related indicators (PLTC, MPV). Additionally, the dataset records the length of stay in the ICU (LOS-ICU) and patient mortality outcomes [65].

3 Results

3.1 Sanity check tests

Here the results from six base examples used to choose the implementation of the Davies-Bouldin Index. And the results from the clouds experiments used to give an example of the behavior of the index.

3.1.1 Test examples

In Table 6 we show the results from the sanity check tests. The `clv.Davies.Bouldin()` matched the calculation made by hand.

Both `index.DB()` and `davies_bouldin_score()` performed correctly on examples 2, 3 and 5. `index.DB()` presented two small issues: Does not contemplate the worst case scenario, cluster centroids overlapping, writing NaN instead of $+\infty$ on examples 1 and 4; and makes a calculation error of ~ 0.05 on example 6.

Table 6 Final test results: Results of three different implementations of the Davies-Bouldin Index on six examples compared with the result calculated by hand

Example	clv.Davies.Bouldin()	index.DB()	davies_bouldin_score()	Actual result
1	$+\infty$	NaN	0	$+\infty$
2	1	1	1	1
3	0	0	0	0
4	$+\infty$	NaN	0	$+\infty$
5	1	1	1	1
6	1.369607	1.414214	1.369607	1.369607

Table 7 Davies-Bouldin Index results: Separated clouds case

Davies-Bouldin Index score	n
0.281	100
0.288	1000
0.291	10,000

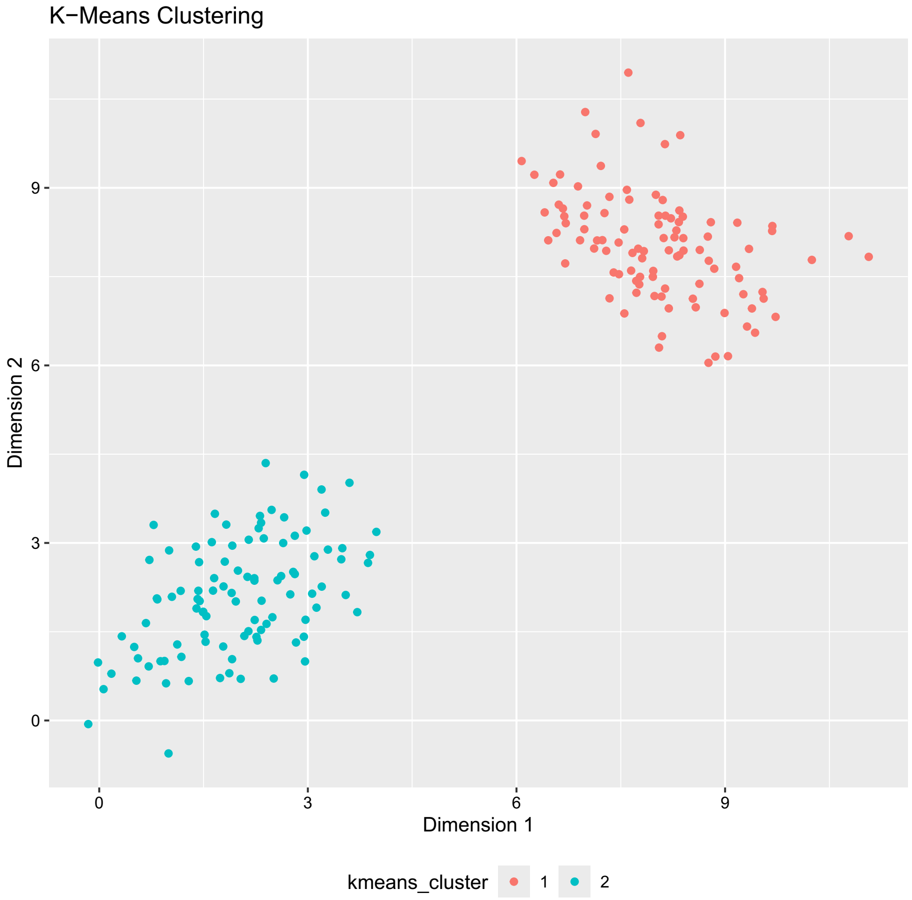


Fig. 3 Well-separated clouds with n=100 points each clustered with *k*-means algorithm, k=2

davies_bouldin_score performed correctly on example 6 but confuses the worst case scenario ($+\infty$) with the best case scenario (0), every cluster has all the elements having the same value and the clusters are well separated, on example 1 and 4.

3.1.2 Well separated clouds

We show in Table 7 the clv.Davies.Bouldin results for clusters of well separated clouds of data.

We show in Figs. 3, 4, and 5, the plots of the *k*-means clustering on well separated clouds of data respectively for n equal to 100, 1,000, and 10,000.

k-means gave similar performances with 0.281 for $n = 100$, 0.288 for $n = 1,000$ and 0.291 for $n = 10,000$, according to the Davies-Bouldin index.

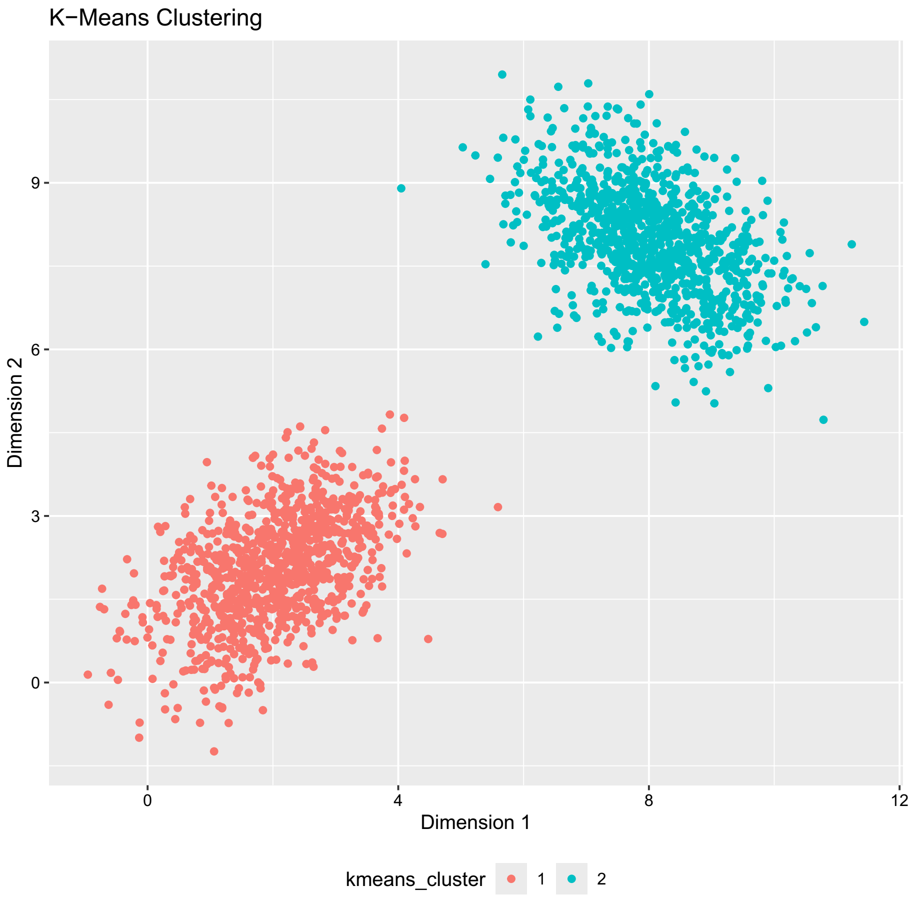


Fig. 4 Well-separated clouds with n=1,000 points each clustered with *k*-means algorithm, k=2

3.1.3 Non-separated clouds

We show in Table 8 the clv.Davies.Bouldin results for clusters of non-separated clouds of data.

We show in Figs. 6, 7 and 8, the plots of the *k*-means clustering on non-separated clouds of data respectively for n equal to $100,\ 1000,\ and\ 10000$. *k*-means gave similar performances with 0.880 for $n = 100$, 0.951 for $n = 1,000$ and 0.930 for $n = 10,000$, according to the Davies-Bouldin index.

Similar results were obtained from multiple iterations of both Well-separated clouds and Non-separated clouds experiments

3.2 Matrices of zeros and ones - k means and dimensionality

Figure 9 shows the Davies-Bouldin Index scorings of the *k*-means clustering in each step of the iteration for $n = 5$. The Index gives a score of 0 at step zero and goes worsening almost scaling at each step.

Figure 9 also shows how many points are in each cluster at each step for $n = 5$. *k*-means makes clusters with proportions contained between 40% and 60% for seven out of ten steps.

Figure 10 shows the Davies-Bouldin Index scorings of the *k*-means clustering in each step of the iteration for $n = 10$. The Index gives a score of 0 at step zero and goes

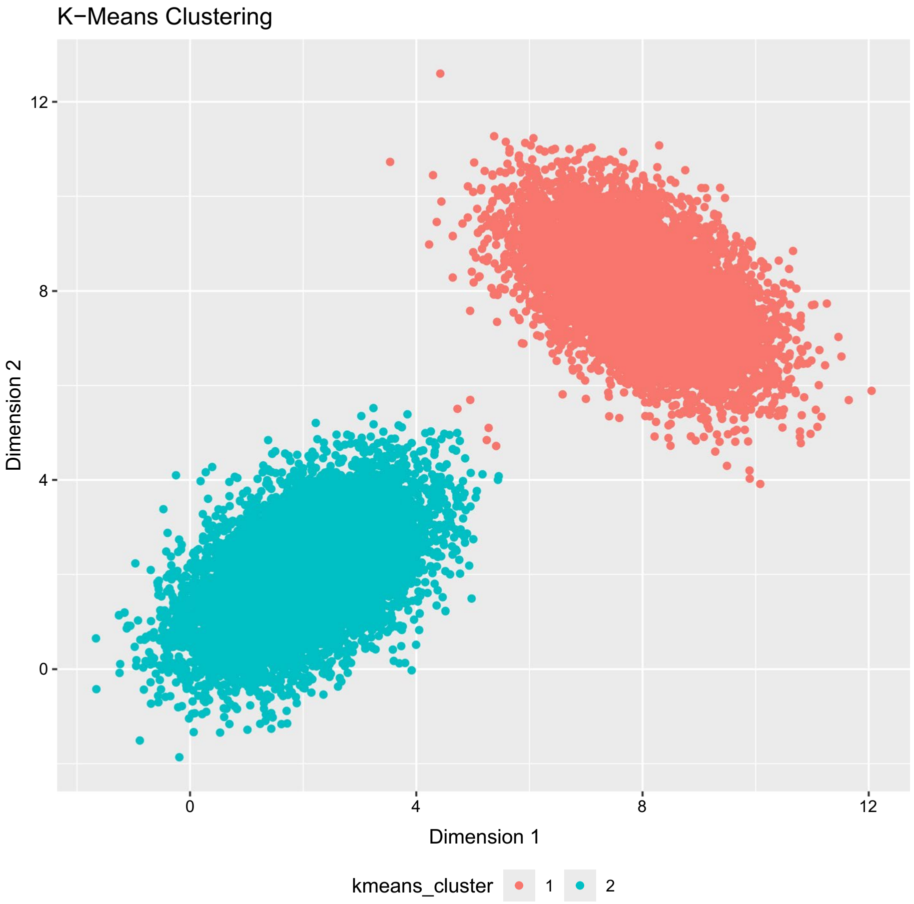


Fig. 5 Well-separated clouds with n=10,000 points each clustered with *k*-means algorithm, k=2

Table 8 Davies-Bouldin Index results: Non-separated clouds case

Davies-Bouldin Index score	n
0.880	100
0.951	1000
0.930	10,000

worsening almost scaling at each step. Between steps 17 and 18 there is a leap of ~ 0.6 before the scores starts to lower again.

Figure 10 also shows how many points are in each cluster at each step for n=10. *k*-means makes clusters with growing disproportion until step 17, from step 18 the two clusters are more similar in proportion.

Figure 11 shows the Davies-Bouldin Index scoring of the *k*-means clustering in each step of the iteration for $n = 20$. The Index gives a score of 0 at step zero and goes worsening reaching a plateau around step 30. Between step 35 and 36 there is a leap of more than two points in the score. And except for step 38, where the score is ~ 0.5, the score goes worsening.

Figure 11 also shows how many points are in each cluster at each step for $n = 20$. *k*-means makes clusters with growing disproportion until step 35, from step 36 the two clusters are more similar in proportion. Except for step 38.

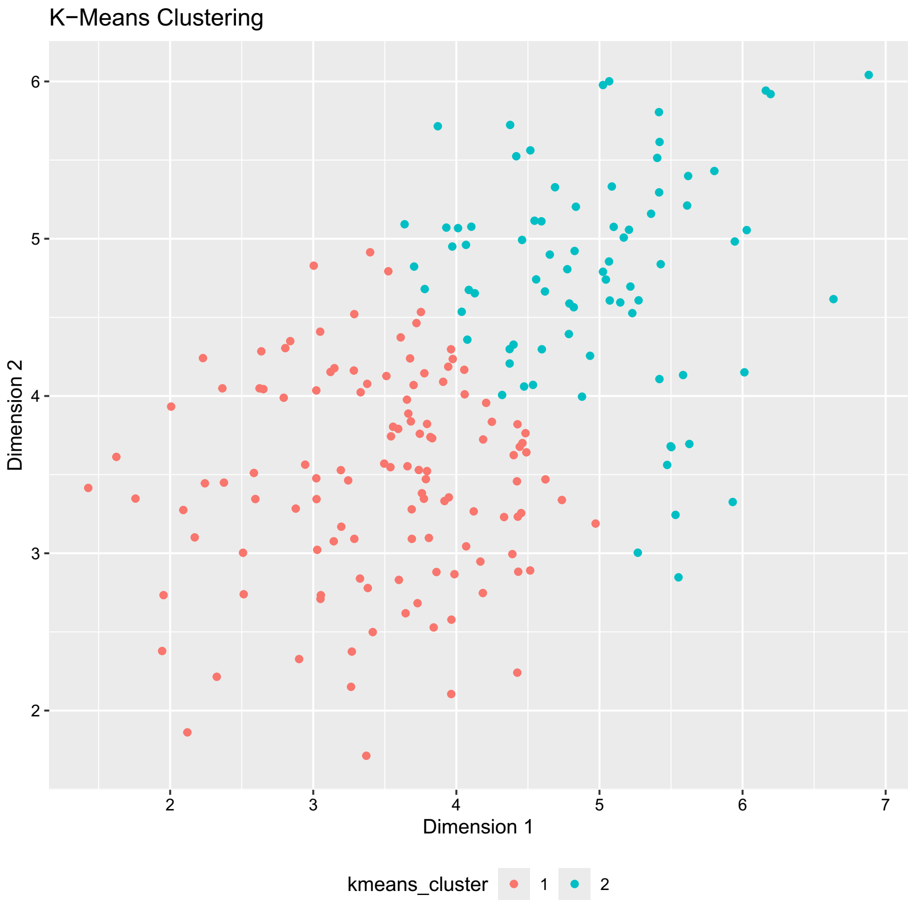


Fig. 6 Non-separated clouds with n=100 points each clustered with *k*-means algorithm, k=2

3.3 EHRs–results for multiple algorithms

In this section, we present the clustering outcomes based on the Davies-Bouldin Index (DBI) and cluster distributions for the datasets analyzed. Multiple clustering algorithms were applied, including different configurations of DBSCAN (small, medium, large epsilon), hierarchical clustering (single, complete, and average linkage), and *k*-means with 2, 3, and 5 centers.

Cluster 0 represents the outliers in the DBSCAN clustering and is not considered a cluster in the Davies-Bouldin Index scoring.

Depression and Heart Failure dataset For this dataset [69], the DBSCAN algorithm with a medium and larger epsilon yielded the lowest Davies-Bouldin Index score, ~ 0.2, forming three clusters. A smaller epsilon meant lower result of Davies-Bouldin Index. In all three cases the outliers were between 20 and 25\% of the data.

Hierarchical clustering with single linkage produced the highest score, ~ 2, and formed a single large cluster with $\sim 95\%$ of the data.

k-means achieved results between 0.5 and 0.8 forming clusters with great proportion variation having the bigger and smaller clusters differ at least of 25% of size.

Hierarchical clustering at complete and average linkage gave similar results to *k*-means both in score and in cluster disproportions forming 4 clusters both (Fig. 12).

Neuroblastoma dataset We applied several algorithms to this dataset [66, 70]. The DBSCAN algorithm yielded the highest Davies-Bouldin Index score, ~ 0.61 with a

Fig. 7 Non-separated clouds with n=1,000 points each clustered with *k*-means algorithm, k=2

smaller epsilon, forming three clusters. Growing the epsilon size the algorithm produced five clusters, two bigger than 25% of the dataset and three smaller than 10%, which resulted in a better score according to the Davies-Bouldin Index, $\sim 57.5\%$ for a medium epsilon and $\sim 52.5\%$ for a bigger one. In all three cases the outliers were between $20-25\%$ of the data.

Hierarchical clustering with single linkage produced the lowest score, ~ 0.35, and formed mostly a single large cluster with $\sim 90\%$ of the data.

k-means got results between 0.5 and 0.55 forming clusters with great proportion variation having the bigger and smaller clusters reaching a 25% difference of size.

Hierarchical clustering with complete and average linkage gave similar results to *k*-means both in score and in cluster disproportions forming 4 clusters both (Fig. 13).

Sepsis and SIRS dataset The DBSCAN algorithm yielded the highest Davies-Bouldin Index score in this dataset [67], higher than 1.6 with a smaller epsilon, forming two clusters. Growing the epsilon size the algorithm produced the same number of clusters, growing Cluster 1 size by at least ten points from $\sim 60\%$ to more than 75% with a medium epsilon and $\sim 80\%$ with a bigger epsilon, which resulted in a higher score for both configurations according to the Davies-Bouldin Index, between 0.65% and 0.75%. The outliers group size went from almost 38% to less than 10% as the epsilon grew.

Hierarchical clustering produced a single large cluster, of more than 90% of the dataset, in all three cases; getting a higher grade from Davies-Bouldin Index with complete

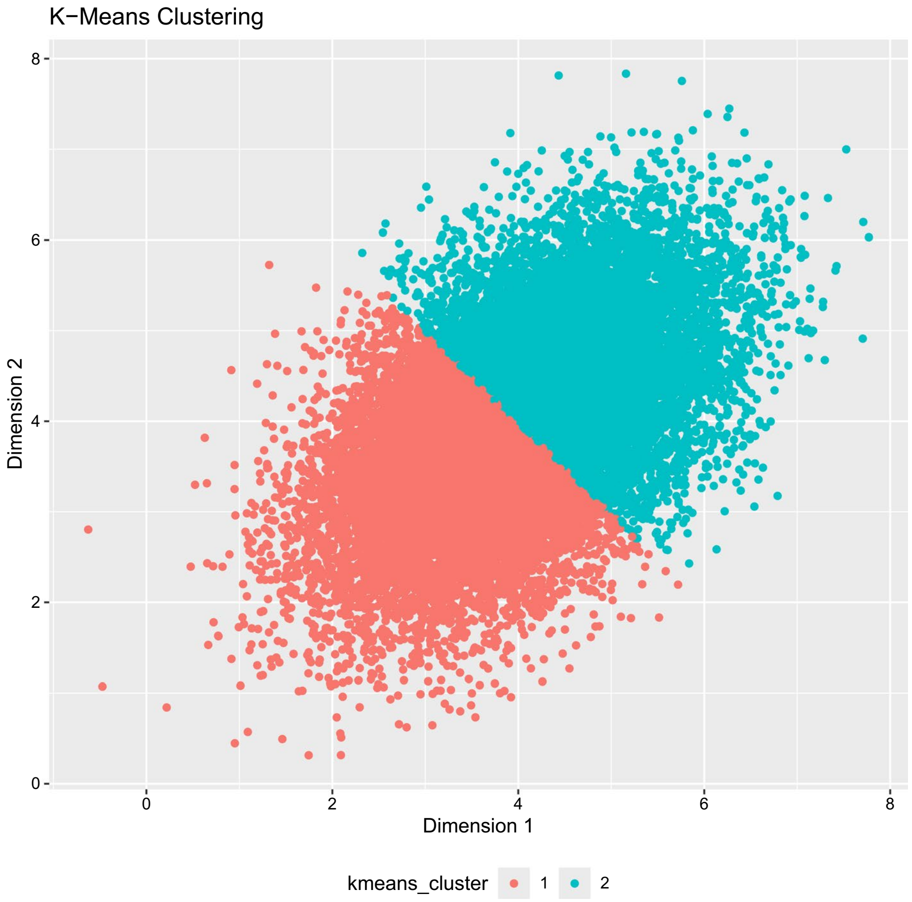


Fig. 8 Non-separated clouds with n=10,000 points each clustered with *k*-means algorithm, k=2

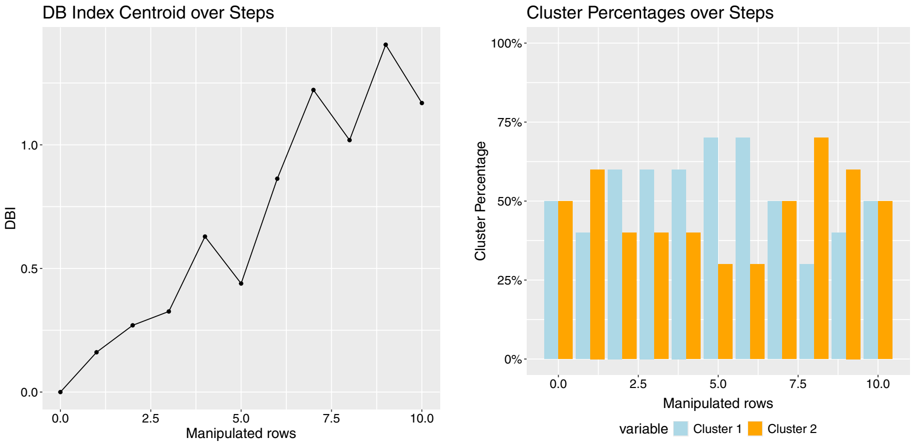


Fig. 9 Matrices of zeros and ones. For n=5. Plot showing DBI values as the number of manipulated rows increases (left) and the percentages of cluster memberships as the number of manipulated rows increases (right)

and average linkage close to a 0.5% score and single linkage close to 0.25% score. *k*-means clustering formed clusters with varying size in all three cases, a higher than 25% difference in size between the bigger and smaller cluster. The Davies-Bouldin Index appointed a score around 0.8 to all three the clustering settings; Appointing 2, 3 and 5 centers made little difference according to the Index (Fig. 14).

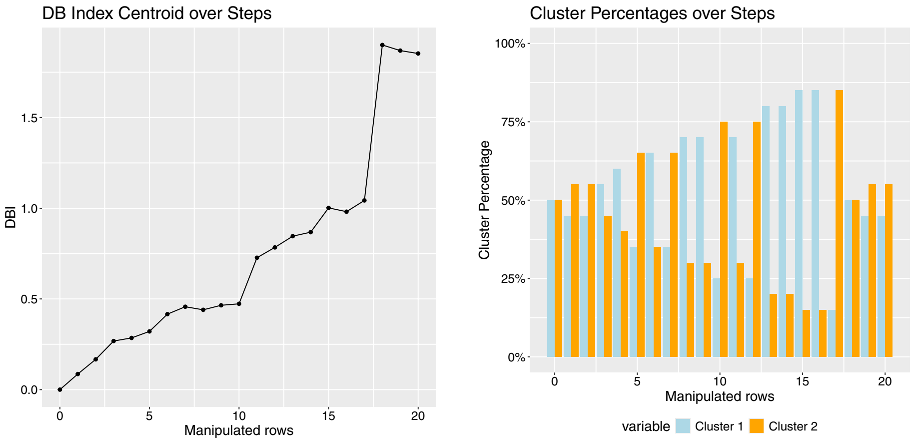


Fig. 10 Matrices of zeros and ones. For n=10. Plot showing DBI values as the number of manipulated rows increases (left) and the percentages of cluster memberships as the number of manipulated rows increases (right)

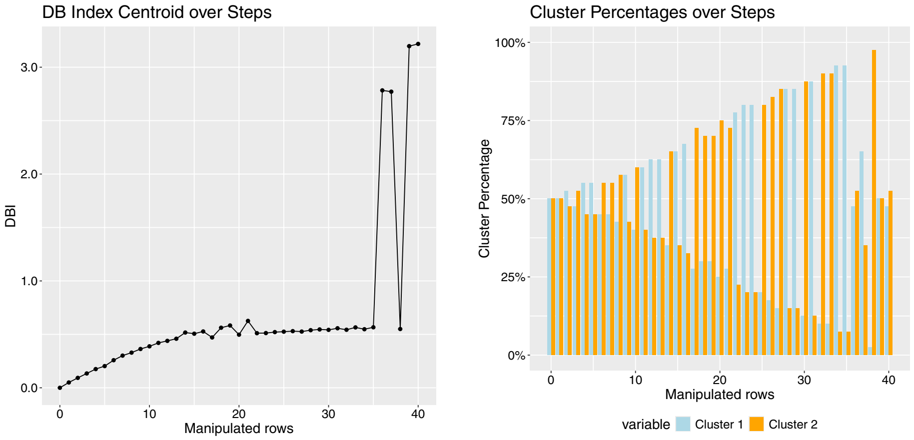


Fig. 11 Matrices of zeros and ones. For n=20. Plot showing DBI values as the number of manipulated rows increases (left) and the percentages of cluster memberships as the number of manipulated rows increases (right)

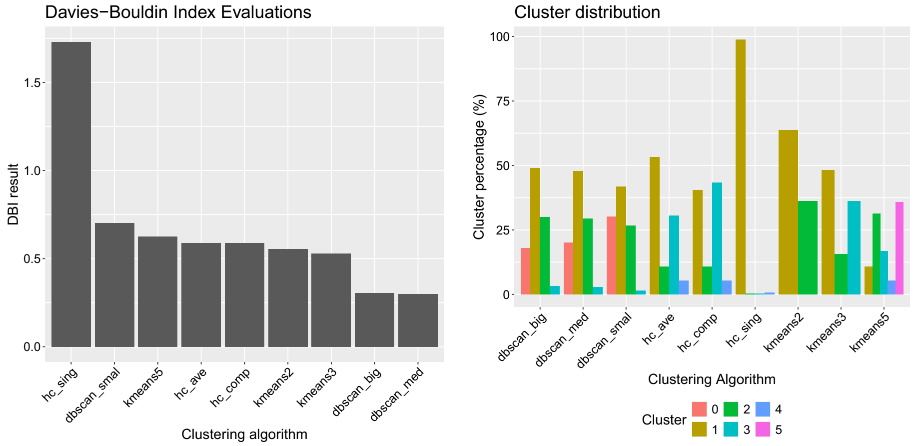


Fig. 12 Evaluation and clustering percentage results on the depression and heart failure dataset. Davies-Bouldin Index scores for different configurations of three clustering algorithms: DBSCAN, Hierarchical clustering and *k*-means. Cluster sizes in percentage by clustering algorithm (Cluster 0 are outliers). DBI results (left) and cluster memberships (right)

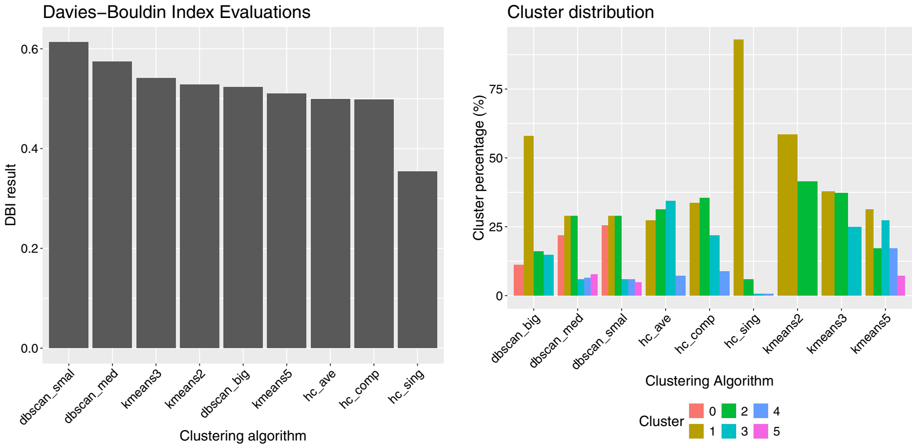


Fig. 13 Evaluation and clustering percentage results on the neuroblastoma dataset. Davies-Bouldin Index scores for different configurations of three clustering algorithms: DBSCAN, Hierarchical clustering and *k*-means. Cluster sizes in percentage by clustering algorithm (Cluster 0 are outliers). DBI results (left) and cluster memberships (right)

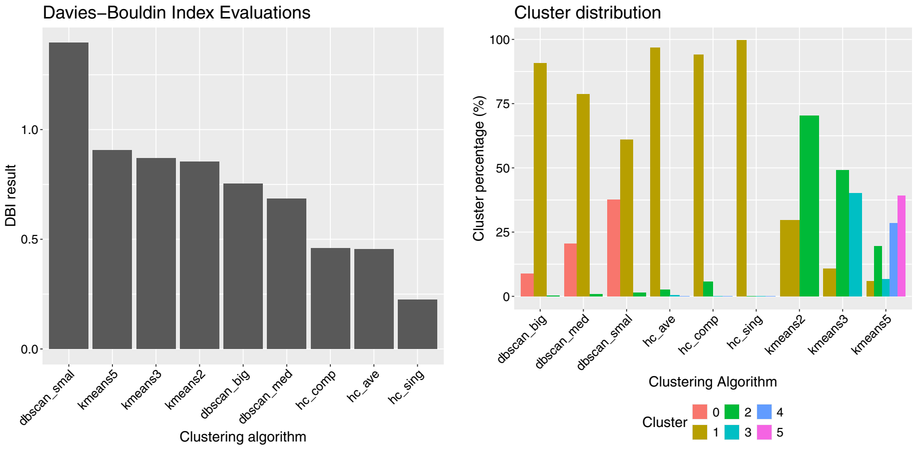


Fig. 14 Evaluation and clustering percentage results on the sepsis and SIRS dataset. Davies-Bouldin Index scores for different configurations of three clustering algorithms: DBSCAN, Hierarchical clustering and *k*-means. Cluster sizes in percentage by clustering algorithm (Cluster 0 are outliers). DBI results (left) and cluster memberships (right)

Diabetes type one dataset In this dataset [68] the DBSCAN algorithm with a medium and a bigger epsilon achieved around 0.9 Davies-Bouldin Index score, forming both a cluster with around 70% of the dataset size and outliers being 20% of the dataset. The algorithm formed three and four clusters, all sized $\sim 5\%$ of the dataset, respectively in the bigger epsilon and medium epsilon cases. The outliers group almost doubled in size in the smaller epsilon case, with around 37% of the dataset. A smaller epsilon corresponds to a reduction to two clusters with cluster one around 60% of the dataset size; and a ~ 1.25 Davies-Bouldin Index score.

Hierarchical clustering produced a single large cluster, of $\sim 80\%$ and $\sim 95\%$ of the dataset, with average and single linkage. Forming three much smaller clusters in both those cases. Hierarchical clustering with average linkage scored around 0.95 in the Davies-Bouldin Index scale; while with single linkage scored less than 0.75. Hierarchical clustering with complete linkage formed clusters of varying size. The bigger cluster taking less than half of the dataset, Cluster 3 taking $\sim 35\%$, Cluster 2 taking $\sim 15\%$ and Cluster 4 less than 5% of the dataset size. The *k*-means clustering formed clusters with varying size in all three cases, a higher than 25% difference in size between the bigger and smaller cluster. The Davies-Bouldin Index appointed very different scores in these cases: ~ 1.25 to 2 centers configuration, almost 1.75 in 3 centers configuration and around 1.40 to 5centers configuration (Fig. 15).

Cardiac arrest dataset DBSCAN behaved similarly in all three configurations in this scenario [65], both according to Davies-Bouldin Index score both in clusters numbers and proportions. All three configurations scored between 0.7 and 0.8 in the Davies-Bouldin scale; and all three formed one big cluster with more than 70% of the data and three to four smaller clusters with at best 5% of the data. The outlier's Cluster 0 reaches no higher than the 20% of the cluster size in all those cases.

Hierarchical clustering with average linkage scored 0.6 in the Davies-Bouldin Index scale. Forming 1 bigger cluster containing up to 65% of the dataset, one cluster with around 30% of the dataset, and two smaller clusters with some percentile point of size.

Hierarchical clustering with complete linkage scored around 0.83 in the Davies-Bouldin Index scale. Forming 2 bigger clusters with around 40% of the dataset each,and two smaller clusters with $\sim 10\%$ and $\sim 6\%$ of the dataset.

Hierarchical clustering with single linkage scored 0.86 in the Davies-Bouldin Index scale. Forming one big cluster with almost all the dataset, leaving three small clusters with at best a couple percentile points of the dataset size.

k-means algorithm produced clusters of similar sizes in all three cases, both the three and 5 centers configurations forming a smaller cluster of around 5% of the dataset size. The scoring in the Davies-Bouldin Index scale is also similar with all three cases scoring between 0.85 and 0.9 (Fig. 16).

4 Discussion and conclusions

Insights from the theoretical analysis and synthetic experiments provided a framework for interpreting the real-world clustering results. They helped us understand, for example, why hierarchical clustering behaved unexpectedly in the presence of noise or why DBI scores for K-means were sometimes less informative, guiding our reasoning as we analyzed the results. The patterns observed in the real-world datasets were consistent

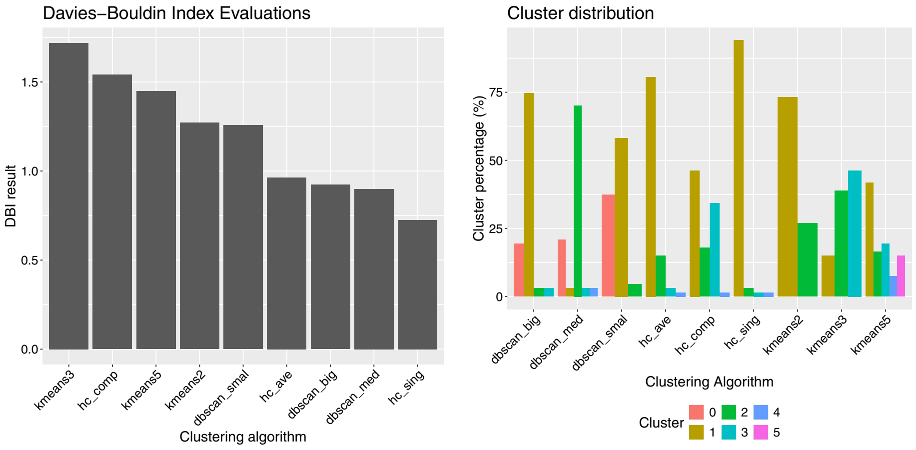


Fig. 15 Evaluation and clustering percentage results on the diabetes type one dataset. Davies-Bouldin Index scores for different configurations of three clustering algorithms: DBSCAN, Hierarchical clustering and *k*-means. Cluster sizes in percentage by clustering algorithm (Cluster 0 are outliers). DBI results (left) and cluster memberships (right)

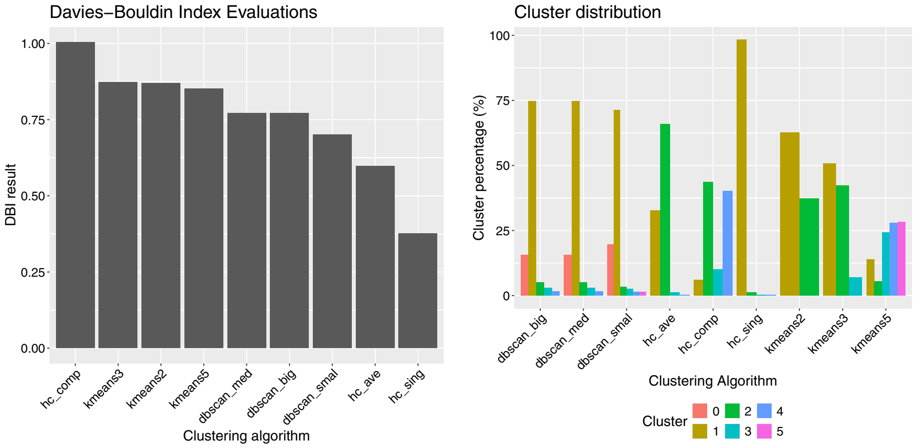


Fig. 16 Evaluation and clustering percentage results on the cardiac arrest dataset. Davies-Bouldin Index scores for different configurations of three clustering algorithms: DBSCAN, Hierarchical clustering and *k*-means. Cluster sizes in percentage by clustering algorithm (Cluster 0 are outliers). DBI results (left) and cluster memberships (right)

with the behaviors identified in the synthetic experiments, confirming that DBI's limitations—such as underestimating poorly separated or irregularly shaped clusters—can influence the interpretation of clustering outcomes. This study's main contribution is to equip practitioners with a lens through which to interpret DBI results, allowing them to recognize subtle or elusive patterns that might otherwise be overlooked.

While widely used and well-regarded in scientific research [25, 44], our results confirm that DBI provides only partial information about clusters, ignoring both cluster magnitude and, to a large extent, cluster shape.

Lower (higher) DBI scores should not be interpreted as evidence of superior (inferior) clustering quality, but rather as an indication of how well the clustering structure matches the assumptions embedded in the index, namely the separation between cluster centroids and the average dispersion of points around their centroid.

For example, in cases such as a small ball within a donut-shaped cluster (Fig. 1a), DBI may rate DBSCAN correctly splitting them as poor clustering, while *k*-means splitting both shapes may appear reasonable according to DBI, clearly illustrating the index's limitations. These issues are especially pronounced in high-dimensional datasets, where intertwined clusters can bring centroids closer together, affecting DBI scores.

Another example is how in both cloud experiments, centroid distances and within-cluster variance remain constant across runs, as points are randomly generated around fixed centers with predefined variance. As the number of generated points increases, the law of large numbers causes the observed quantities to converge toward these imposed values. Because DBI is explicitly based on centroid separation and intra-cluster dispersion, and because k-means consistently partitions the data according to the same geometric assumptions, the resulting clusters remain spatially similar across experiments, leading DBI to assign comparable scores.

Our experiments highlight that DBI is most informative when applied to DBSCAN results. DBSCAN identifies dense regions while ignoring sparsity, and DBI complements this by evaluating intra- and inter-cluster compactness. In contrast, the interpretation of DBI scores can be less informative for *k*-means and hierarchical clustering, where cluster formation criteria aligns with DBI's assumptions or produces trivial solutions (for example, all points in one cluster as also observed in literature [26]).

We also recommend monitoring the proportion of outliers, as DBI does not account for them.

We also have to report several limitations of this study. The literature review focused on a subset of studies, and our datasets were primarily biomedical with varying data scales. A broader dataset variety, more clustering algorithms, and additional comparisons with other indices could strengthen future studies. Furthermore, the DBI implementations tested allowed only partial flexibility in parameter tuning.

Ultimately, this article contributes a systematic, application-oriented perspective on DBI, bridging the gap between theoretical understanding and practical use. By providing detailed examples, highlighting limitations, and offering actionable guidelines, this study serves as a manual of use for DBI in diverse clustering tasks.

It seems to us a good practice complementing DBI with indices such as the Silhouette Index, which looks robust and better supported both theoretically and in the literature [25], as it seems to provide a more reliable measure of inter-cluster distance [71]. Additionally, we suggest considering the ’I’ Index [21], which we propose renaming as the ’Bandyopadhyay-Maulik Index’ or ’Ban-Mau Index’ to improve its recognizability; further studies are required to validate this proposal. Finally, a simple modification to DBI could involve replacing the ’max’ operator in the formula with an ’average’, potentially enhancing its performance.

This study provides researchers and practitioners with a concrete framework to effectively apply DBI in diverse clustering scenarios. Moreover, in the future we plan to design a density-based version of the Davies-Bouldin index to employ to assess nested clusters. If the DBCV index can be considered a density-based variant of the Silhouette coefficient [72, 73], we believe a density-based variant of DBI can be effective as well.

Abbreviations

BIC	Bayesian Information Criterion
CHI	Calinski-Harabasz Index
DBCV	Density-Based Clustering Validation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DBI	Davies-Bouldin Index
DRI	Distance Ratio Index
EAP	Extracellular Action Potential
EHRs	Electronic Health Records
EMG	Electromyography
FLDI	Fisher’s Linear Discriminant Index
HC	Hierarchical Clustering
JI	Jaccard Index
MES	Myoelectric Signals
PAMI	Pattern Analysis and Machine Intelligence
SEMG	Surface Electromyography
SIRS	Systemic Inflammatory Response Syndrome
SOM	Self-Organizing Map

Acknowledgements

No one to mention.

Author contributions

Alessandro Anfossi: formal analysis, investigation, methodology, resources, software, visualization, writing-original draft, writing-review and editing. Davide Chicco: conceptualization, investigation, methodology, project administration, supervision, writing-original draft, writing-review and editing.

Funding

Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. The work of D.C. is partially funded by the Italian Ministero Italiano delle Imprese e del Made in Italy under the Digital Intervention in Psychiatric and Psychologist Services (DIPPS) program (project code F/310240/01-04/X56), within the framework “Innovation Agreements” (Accordi per l’Innovazione) and is partially supported by the Ministero dell’Università e della Ricerca of Italy under the “Dipartimenti di Eccellenza 2023-2027” ReGAInS grant assigned to the Dipartimento di Informatica, Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The deidentified datasets derived from EHRs and used in this project are publicly available under the Creative Commons CC BY 4.0 license at: - Cardiac arrest dataset https://figshare.com/articles/dataset/Mortality_after_ out-of-hospital_cardiac_arrest_in_a_Spanish_Region/4876247?file=8166893 - Neuroblastoma dataset https://doi.org/10.7717/peerj.5665/supp-5 - Diabetes type one dataset https://figshare.com/articles/dataset/Circulating_ osteocalcin_as_a_bone-derived_hormone_is_inversely_correlated_with_body_ fat_in_patients_with_type_1_diabetes/8079389?file=15057092 - Sepsis and SIRS dataset https://figshare.com/articles/dataset/_C_Reactive_ Protein_and_Hemogram_Farameters_for_the_Non_Sepsis_Systemic_Inflammatory_ Response_Syndrome_and_Sepsis_What_Do_They_Mean_/1644426?file=2637248 - Depression and heart failure dataset https://figshare.com/articles/dataset/Comorbid_ Depression_and_Heart_Failure_A_Community_Cohort_Study/3916224?file=6130425

Declarations

Ethics approval

Permission to collect and analyze the data of the patients involved in this study was obtained and approved by the Institutional Review Boards through the original dataset curators, as stated in the original dataset articles [65–69].

Consent to participate

Informed consent for participation in the studies involving patient data was obtained from the Institutional Review Boards by the original dataset curators, as stated in the original dataset articles [65–69].

Consent for publication

Not applicable.

Software availability

Our R software code is publicly available under the GPL−3.0 license at: https://github.com/AleAnfossi/DBI-guide

Competing interests

The authors declare no competing interests.

Received: 11 October 2025 / Accepted: 1 April 2026

References

1. Nielsen F. Hierarchical Clustering. In Introduction to HPC with MPI for Data Science, pages 195–211. Springer; 2016. https://doi.org/10.1007/978-3-319-21903-5_8.
2. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96. AAAI Press. 1996. pp. 226–231. https://www.semanticscholar.org/paper/A-Density-Based-Algorithm-for-Discovering-Clusters-Ester-Kriegel/5c8fe9a0412a078e30eb7e5eeb0068655b673e86.
3. MacQueen JB. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley. 1967. pp. 281–297. University of California Press. https://www.semanticscholar.org/paper/Some-methods-for-classification-and-analysis-of-MacQueen/ac8ab51a86f1a9ae74dd0e4576d1a019f5e654ed.
4. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybernet. 1973;3(3):32–57. https://doi.org/10.1080/01969727308546046.
5. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
6. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3(1):1–27. https://doi.org/10.1080/03610927408827101.
7. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the Gap Statistic. J R Statist Soc Ser B (Statistical Methodology). 2001;63(2):411–23. https://doi.org/10.1111/1467-9868.00293.
8. Davies DL, Bouldin DW. A cluster separation measure. In IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1. No. (2). 1979. pp. 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.
9. Thompson EA, Gann LB, Cressman ENK. Learning to successfully search the scientific and medical literature. Cell Stress Chaperon. 2019;24(2):289–93. https://doi.org/10.1007/s12192-019-00984-2.
10. Kovács F, Legány C, Babos A. Cluster validity measurement techniques. In Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. 2006. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c4f9df3c66105382d05e58ec35faa8d435f55c91.
11. Mughnyanti M, Efendi S, Zarlis M. Analysis of determining centroid clustering x-means algorithm with Davies-Bouldin index evaluation. In IOP Conference Series Materials Science and Engineering vol. 725. No. (1). 2020. p. 012128. https://doi.org/10.1016/0005-2795(75)90109-9.
12. Chhabra A, Masalkovaité K, Mohapatra P. An overview of fairness in clustering. IEEE Access. 2021;9:130698–720. https://doi.org/10.1109/ACCESS.2021.3114099.
13. Saha S, Bandyopadhyay S. Performance evaluation of some symmetry-based cluster validity indexes. IEEE Trans Syst Man Cybernet Part C (Applications and Reviews). 2009;39(4):420–5. https://doi.org/10.1109/TSMCC.2009.2013335.
14. Gultom S, Sriadhi S, Martiano M, Simarmata J. Comparison analysis of k-means and k-medoid with Euclidean, Canberra, and Chebyshev distance for big data clustering. In IOP Conference Series Materials Science and Engineering vol. 420. No. 1. 2018. p. 012092. https://doi.org/10.1088/1757-899X/420/1/012092.
15. Koch FC, Sutton GJ, Voineagu I, Vafaee F. Supervised application of internal validation measures to benchmark dimensionality reduction methods in scrna-seq data. Brief Bioinform. 2021;22(6):bbab304. https://doi.org/10.1093/bib/bbab304.

16. Di Caro L, Frias-Martinez V, Frias-Martinez E. Analyzing the role of dimension arrangement for data visualization in Radviz. In: Zaki MJ, Yu JX, Ravindran B, Pudi V, editors. Advances in Knowledge Discovery and Data Mining. Heidelberg: Springer Berlin Heidelberg; 2010. pp. 125–132. https://doi.org/10.1007/978-3-642-13672-6_13.
17. Stein B, Meyer S, Wißbrock E. On cluster validity and the information need of users. Artif Intell Appl. 2003;373:216–21. https://doi.org/10.1016/0005-2795(75)90109-9.
18. Petrovic SV. A comparison between the Silhouette index and the Davies-Bouldin index in labelling ids clusters. In Proceedings of the 11th Nordic Workshop of Secure IT Systems. 200653–64. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2db00f73fc6b97ebe12e97cfdaefbb2fefc253b.
19. Günter S, Bunke H. Validation indices for graph clustering. Pattern Recogn Lett. 2003;24(8):1107–13. https://doi.org/10.1016/S0167-8655(02)00257-X.
20. Kryszczuk K, Hurley P. Estimation of the number of clusters using multiple clustering validity indices. In El Gayar N, Kittler J, Roli F, editors. Multiple Classifier Systems. Heidelberg: Springer Berlin Heidelberg; 2010. pp. 114–123. https://doi.org/10.1016/0005-2795(75)90109-9.
21. Bandyopadhyay S, Maulik S. Nonparametric genetic clustering: comparison of validity indices. IEEE Trans Syst Man Cybernet Part C (Applications and Reviews). 2001;31(1):120–5. https://doi.org/10.1109/5326.923275.
22. Bandyopadhyay S, Maulik U. Performance evaluation of some clustering algorithms and validity indices. IEEE Trans Pattern Anal Mach Intell. 2002;24(12):1650–4. https://doi.org/10.1109/TPAMI.2002.1114856.
23. Bolshakova N, Azuaje F. Cluster validation techniques for genome expression data. Signal Process. 2003;83(4):825–33. https://doi.org/10.1016/S0165-1684(02)00475-9.
24. Hashimoto W, Nakamura T, Miyamoto S. Comparison and evaluation of different cluster validity measures including their kernelization. SCIS ISIS. 2008;2008:399–403. https://doi.org/10.14864/softscis.2008.0.399.0
25. Rendón E, Abundez IM, Gutiérrez C, Zagal SD, Arizmendi A, Quiroz EM, Arzate HE. A comparison of internal and external cluster validation indexes. In Proceedings of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications, AMERICAN-MATH'11/CEA'11 Puerto Morelos, Mexico. World Scientific and Engineering Academy and Society (WSEAS). 2011. pp. 158–163. https://doi.org/10.1109/ICSET53708.2021.9612525.
26. Yang C, Wan B, Gao X. Effectivity of internal validation techniques for gene clustering. In Nicos Maglaveras, Ioanna Chouvarda, Vassilis Koutkias, and Rüdiger Brause, editors, Biological and Medical Data Analysis. Heidelberg: Springer Berlin Heidelberg; 2006. pp. 49–59. https://doi.org/10.1007/11946465_5.
27. Baarsch J, Celebi ME. Investigation of internal validity measures for k-means clustering. Lecture Notes Eng Comput Sci. 2012;2195:471–6. https://doi.org/10.1016/0005-2795(75)90109-9.
28. Mali K, Mitra S. Clustering and its validation in a symbolic framework. Pattern Recogn Lett. 2003;24(14):2367–76. https://doi.org/10.1016/S0167-8655(03)00066-7.
29. Kärkkäinen I, Fränti P. Minimization of the value of Davies-Bouldin index. Department of Computer Science. University of Joensuu. 2000. http://cs.uef.fi/sipu/pub/DBI-SPC2000.pdf.
30. Gan Y, Kakiashvili T, Koczkodaj WW, Li F. A note on relevance of diagnostic classification and rating scales used in psychiatry. Comput Methods Programs Biomed. 2013;112(1):16–21. https://doi.org/10.1016/j.cmpb.2013.05.016.
31. Urban S, Błaziak M, Jura M, Iwanek G, Ponikowska B, Horudko J, et al. Machine learning approach to understand worsening renal function in acute Heart Failure. Biomolecules. 2022;12(11):1616. https://doi.org/10.3390/biom12111616.
32. Urban S, Błaziak M, Jura M, Iwanek G, Zdanowicz A, Guzik M, et al. Novel phenotyping for acute Heart Failure–unsupervised machine learning-based approach. Biomedicines. 2022;10(7):1514. https://doi.org/10.3390/biomedicines10071514.
33. Mullin S, Zola J, Lee R, Hu J, MacKenzie B, Brickman A, et al. Longitudinal k-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. J Biomed Inform. 2021;122:103889. https://doi.org/10.1016/j.jbi.2021.103889.
34. Chiu Y-C, Chen H-IH, Gorthi A, Mostavi M, Zheng S, Huang Y, et al. Deep learning of pharmacogenomics resources: moving towards precision oncology. Brief Bioinform. 2019;21(6):2066–83. https://doi.org/10.1093/bib/bbz144.
35. Leng D, Zheng L, Wen Y, Zhang Y, Lianlian W, Wang J, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biol. 2022;23(1):171. https://doi.org/10.1016/0005-2795(75)90109-9.
36. Handhayani T, Hiryanto L. Intelligent kernel k-means for clustering gene expression. Procedia Comput Sci. 2015;59:171–7. https://doi.org/10.1016/j.procs.2015.07.544.
37. Xiao G, Guan R, Cao Y, Huang Z, Xu Y. KISL: knowledge-injected semi-supervised learning for biological co-expression network modules. Front Genet. 2023. https://doi.org/10.3389/fgene.2023.1151962.
38. Wang G, Wang Z, Chen W, Zhuang J. Classification of surface EMG signals using optimal wavelet packet method based on Davies-Bouldin criterion. Med Biol Eng Comput. 2006;44(10):865–72. https://doi.org/10.1007/s11517-006-0100-y.
39. Xi X, Tang M, Luo Z. Feature-level fusion of surface electromyography for activity monitoring. Sensors. 2018;18(2):614. https://doi.org/10.3390/s18020614.
40. Hatamikia S, Maghooli K, Nasrabadi AM. The emotion recognition system based on autoregressive model and sequential forward feature selection of electroencephalogram signals. J Med Signals Sens. 2014;4:194–201. https://doi.org/10.4103/2228-7477.137777.
41. Whitney HM, Hui Li Yu, Ji PL, Giger ML. Harmonization of radiomic features of breast lesions across international dce-mri datasets. J Med Imaging. 2020;7(1):012707. https://doi.org/10.1117/1.JMI.7.1.012707.
42. Ombao H, Fiecas M, Ting C-M, Low YF. Statistical models for brain signals with properties that evolve across trials. Neuroimage. 2018;180:609–18. https://doi.org/10.1016/j.neuroimage.2017.11.061.
43. Li J, Huang J, Jiang T, Liping T, Cui L, Cui J, et al. A multi-step approach for tongue image classification in patients with diabetes. Comput Biol Med. 2022;149:105935. https://doi.org/10.1016/j.compbiomed.2022.105935.
44. Appadurai JP, Bhargavi R, Rani DBV. Prediction of dyslexia severity levels from fixation and saccadic eye movement using machine learning. Biomed Signal Process Control. 2023;79:104094. https://doi.org/10.1016/j.bspc.2022.104094.
45. Tsoi KKF, Chan NB, Yiu KKL, Poon SKS, Lin B, Ho K. Machine learning clustering for blood pressure variability applied to systolic blood pressure intervention trial (sprint) and the Hong Kong community cohort. Hypertension. 2020;76(2):569–76. https://doi.org/10.1161/HYPERTENSIONAHA.119.14213.
46. Zardoshti-Kermani M, Wheeler BC, Badie K, Hashemi RM. EMG feature evaluation for movement control of upper extremity prostheses. IEEE Trans Rehabil Eng. 1995;3(4):324–33. https://doi.org/10.1109/86.481972.

47. Sato T, Suzuki T, Mabuchi K. Fast automatic template matching for spike sorting based on Davies-Bouldin validation indices. In 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2007. pp. 3200–3203. https://doi.org/10.1109/IEMBS.2007.4353010.
48. Surangsrirat D, Sri-iesaranusorn P, Chaiyaroj A, Vateekul P, Bhidayasiri R. Parkinson's disease severity clustering based on tapping activity on mobile device. Sci Rep. 2022;12(1):3142. https://doi.org/10.1038/s41598-022-06572-2.
49. Oskoei MA, Hu H. Ga-based feature subset selection for myoelectric classification. In 2006 IEEE International Conference on Robotics and Biomimetics. 2006. pp. 1465–1470. https://doi.org/10.1109/ROBIO.2006.340145.
50. Chicco D. Ten quick tips for machine learning in computational biology. BioData Mining. 2017;10(35):1–17. https://doi.org/10.1016/0005-2795(75)90109-9.
51. Schnell S. Ten simple rules for a computational biologist's laboratory notebook. PLoS Comput Biol. 2015;11(9):1–5. https://doi.org/10.1371/journal.pcbi.1004385.
52. Noble WS. A quick guide to organizing computational biology projects. PLoS Comput Biol. 2009;5(7):1–5. https://doi.org/10.1371/journal.pcbi.1000424.
53. Blischak JD, Davenport ER, Wilson G. A quick introduction to version control with git and github. PLoS Comput Biol. 2016;12(1):1–18. https://doi.org/10.1371/journal.pcbi.1004668.
54. Walesiak M, Dudeck A. clustersim: searching for optimal clustering procedure for a data set. https://doi.org/10.32614/cran.package.clustersim. Accessed: 2025-01-23.
55. Nieweglowski L. clv: Cluster validation techniques. https://doi.org/10.32614/cran.package.clv. Accessed: 2025-01-23.
56. Wiroonsri N, Preedasawakul O. Universalcvi: hard and soft cluster validity indices. https://doi.org/10.32614/cran.package.universalcvi. Accessed: 2025-01-23.
57. McCaw Z. Mgmm: Missingness aware gaussian mixture models. https://doi.org/10.32614/cran.package.mgmm. Accessed: 2025-01-23.
58. Abraham Y, Sauwen N. Radviz: Project multidimensional data in 2d space. https://doi.org/10.32614/cran.package.radviz. Accessed: 2025-01-23.
59. Lord E, Willems M, Lapointe F-J, Makarenkov V. Clusterstability: Assessment of stability of individual objects or clusters in partitioning solutions. https://doi.org/10.32614/cran.package.clusterstability. Accessed: 2025-01-23.
60. Licen S, Franzon M, Rodani T, Barbieri P. Somenv: Som algorithm for the analysis of multivariate environmental data. https://doi.org/10.32614/cran.package.somenv. Accessed: 2025-01-23.
61. Pascoal F, Branco P, Torgo L, Costa R, Magalhães C. ulrb: Unsupervised learning based definition of microbial rare biosphere. https://doi.org/10.32614/cran.package.ulrb. Accessed: 2025-01-23.
62. Permiakova O, Guibert Romain, and Thomas Burger. 'compressive' hierarchical kernel clustering toolbox. https://cran.r-project.org/package=chickn. Accessed: 2025-01-23.
63. Michael Thrun, Peter Nahrgang, Felix Pape, Vasyl Pihur, Guy Brock, Susmita Datta, Somnath Datta, Luis Winckelmann, Alfred Ultsch, and Quirin Stier. Fcps: Fundamental clustering problems suite. https://doi.org/10.32614/cran.package.fcps. Accessed: 2025-01-23.
64. Ripley B. Mass: Support functions and datasets for venables and ripley's mass. https://doi.org/10.32614/cran.package.mass . Accessed: 2025-01-23.
65. Requena-Morales R, Palazón-Bru A, Rizo-Baeza MM, Adsuar-Quesada JM, Gil-Guillén VF, Cortés-Castell E. Mortality after out-of-hospital cardiac arrest in a spanish region. PLOS One. 2017;12(4):1–10. https://doi.org/10.1371/journal.pone.0175818.
66. Ma Y, Zheng J, Feng J, Chen L, Dong K, Xiao X. Neuroblastomas in eastern China: a retrospective series study of 275 cases in a regional center. PeerJ. 2018;6:e5665. https://doi.org/10.7717/peerj.5665.
67. Gucyetmez B, Atalan HK. C-reactive protein and hemogram parameters for the non-Sepsis systemic inflammatory response syndrome and Sepsis: What do they mean? PLOS One. 2016;11(2):e0148699. https://doi.org/10.1371/journal.pone.0148699.
68. Takashi Y, Ishizu M, Mori H, Miyashita K, Sakamoto F, Katakami N, et al. Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 Diabetes. PLOS One. 2019;14(5):e0216416. https://doi.org/10.1371/journal.pone.0216416.
69. Jani BD, Mair FS, Roger VL, Weston SA, Jiang R, Alanna M. Chamberlain. Comorbid Depression and Heart Failure: a community cohort study. PLOS One. 2016;11(6):1–11. https://doi.org/10.1371/journal.pone.0158570.
70. Cerono G, Melaiu O, Chicco D. Clinical feature ranking based on ensemble machine learning reveals top survival factors for glioblastoma multiforme. J Healthc Inform Res. 2024;8(1):1–18. https://doi.org/10.1007/s41666-023-00138-1.
71. Chicco D, Campagner A, Spagnolo A, Ciucci D, Jurman G The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters. PeerJ Comput Sci. 2025;11:e3309. https://doi.org/10.7717/peerj-cs.3309.
72. Moulavi D, Jaskowiak PA, Campello RJGB, Zimek A, Sander J. Density-based clustering validation. In Proceedings of SDM24—the 2014 SIAM International Conference on Data Mining. 2014. pp. 839–847. SIAM. https://doi.org/10.1137/1.9781611973440.96.
73. Chicco D, Sabino G, Oneto L, Jurman G. 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. 2025;11:e3095. https://doi.org/10.7717/peerj-cs.3095.

Publisher's Note