Chicco *et al. BioData Mining* (2025) 18:40
https://doi.org/10.1186/s13040-025-00455-8

BioData Mining

RESEARCH Open Access

DBSCAN and DBCV application to open medical records heterogeneous data for identifying clinically significant clusters of patients with neuroblastoma

Davide Chicco[1,2*], Luca Oneto[3] and Davide Cangelosi[4]

*Correspondence:
davidechicco@davidechicco.it

[1] Università di Milano-Bicocca, Milan, Italy
[2] University of Toronto, Toronto, Ontario, Canada
[3] Università di Genova, Genoa, Italy
[4] IRCCS Istituto Giannina Gaslini, Genoa, Italy

Abstract

Neuroblastoma is a common pediatric cancer that affects thousands of infants worldwide, especially children under five years of age. Although recovery for patients with neuroblastoma is possible in 80% of cases, only 40% of those with high-risk stage four neuroblastoma survive. Electronic health records of patients with this disease contain valuable data on patients that can be analyzed using computational intelligence and statistical software by biomedical informatics researchers. Unsupervised machine learning methods, in particular, can identify clinically significant subgroups of patients, which can lead to new therapies or medical treatments for future patients belonging to the same subgroups. However, access to these datasets is often restricted, making it difficult to obtain them for independent research projects. In this study, we retrieved three open datasets containing data from patients diagnosed with neuroblastoma: the Genoa dataset and the Shanghai dataset from the Neuroblastoma Electronic Health Records Open Data Repository, and a dataset from the TARGET-NBL renowned program. We analyzed these datasets using several clustering techniques and measured the results with the DBCV (Density-Based Clustering Validation) index. Among these algorithms, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was the only one that produced meaningful results. We scrutinized the two clusters of patients' profiles identified by DBSCAN in the three datasets and recognized several relevant clinical variables that clearly partitioned the patients into the two clusters that have clinical meaning in the neuroblastoma literature. Our results can have a significant impact on health informatics, because any computational analyst wishing to cluster small data of patients of a rare disease can choose to use DBSCAN and DBCV rather than utilizing more common methods such as *k*-Means and Silhouette coefficient.

Keywords: Neuroblastoma, Electronic health records, EHRs, Childhood cancer, Open data, Unsupervised machine learning, DBSCAN, DBCV, Clustering

Introduction

Neuroblastoma is a childhood cancer that affects approximately 6,000 infants and contributes to 15% of cancer-related deaths in children worldwide. Neuroblastoma forms in the tissues of human nerves, is the most prevalent extracranial solid tumor in children, and usually affects patients under five years old [1]. The survival rate for kids with high-risk neuroblastoma is only 40%, while it can reach 80% for low-risk neuroblastoma. Medical treatments for patients with neuroblastoma include surgery, chemotherapy, radiation therapy, and immunotherapy [1].

Scientific research can help medical doctors understand the development and progression of the disease and can be conducted in several ways. Bioinformatics and computational biology research on genomics data, for example, can help unveil the genes most involved in neuroblastoma diagnosis and prognosis [2–7]. On the other hand, data derived from electronic health records (EHRs), collected after hospital laboratory exams (such as blood tests), can be an useful asset for computational analyses [8].

In the past, researchers applied computational statistics and machine learning to several datasets of EHRs of patients with neuroblastoma.

Regarding machine learning, we performed a computational intelligence study on the data of Italian Registry of Peripheral Neuroblastoma (RINB) [9], where we applied a supervised machine learning approach to detect the most predictive clinical variables for the outcome of patients with neuroblastoma. In that case, however, the data could not be released publicly for privacy reasons.

Sometimes, fortunately, study authors are allowed to release EHRs data of patients with neuroblastoma openly online for free, without any restrictions. Barbara Banelli and colleagues [10] investigated the role of 17 genes of the Protocadherin B cluster (PCDHB) on a dataset of genomics and EHRs of 121 patients with stage-4 neuroblastoma. They applied computational statistical methods for a survival analysis on this dataset, by employing the SPSS proprietary software. SPSS is the programming language employed also by Yangyan Ma and coauthors [11] in their study, where they analyze a dataset of EHRs and genetics information to identify the most relevant prognostic factors for 169 patients with stage-3 or stage-4 neuroblastoma.

Shunsuke Kimura et al. [12] analyzed a subset of data from the well-known TARGET-NBL project [13, 14]. They used both genetics data and EHRs to perform a computational statistical analysis through the R open source programming language. The main outcomes of this study regard genomics and the roles of some genes (or gene clusters) over others.

All these three studies have a particular asset: their authors released their neuroblastoma dataset openly and for free, so that it could be analyzed by anyone else in the world, following the FAIR principles [15]. We cleaned the dataset of Barbara Banelli and colleagues [10] and the dataset of Yangyan Ma and coauthors [11], and we described them thorougly and released them in our Neuroblastoma Electronic Health Records Open Data Repository [16, 17], as *dataBB2013* and *dataYM2018* respectively.

Unsupervised machine learning methods, such as clustering, applied on data of EHRs can identify clinically significant groups of patients based on their medical features. These clusters, in turn, can be useful to identify significant subgroups of patients that need particular treatments or therapies.

In the present study, we decided to apply several clustering algorithms to the Genoa dataset of Barbara Banelli and colleagues [10], to the Shanghai open dataset of Yangyan Ma and coauthors [11], and to a subset of the TARGET-NBL stage-4 dataset of Shunsuke Kimura et al. [12]. Among the 94 patient profiles of that dataset, we removed three rows having unknown diagnostic category, and kept only the 93 rows with diagnostic category equal to neuroblastoma or nodular ganglioneuroblastoma, which is a variant of neuroblastoma surrounded by ganglion cells. We describe in detail this dataset in "Datasets" section.

Other resources for EHRs data of patients diagnosed with pediatric neuroblastoma exist. The International Neuroblastoma Risk Group (INRG) has released and currently maintains the INRG Data Commons [18, 19], launched within the Pediatric Cancer Data Commons [20, 21], a global data collection initiative coordinated by University of Chicago (Chicago, Illinois, USA). The INRG Data Commons is a database of thousands of EHRs of patients with this oncological disease, but its access is restricted: researchers who want to analyze these data need to submit a proposal, that needs to be evaluated by an INRG committee who might approve it or not.

On the contrary, the Genoa, Shanghai, and TARGER-NBL datasets analyzed in this study are completely open, unrestricted, public, and can be be analyzed by anyone worldwide.

When high-quality health datasets are available, they can be analyzed using either supervised or unsupervised methods. Supervised approaches are employed when a gold standard piece of information is available, while unsupervised learning methods are used when there is no clear ground truth. Unsupervised problems are more complex but often more useful for investigation in medical sciences. In fact, patients frequently arrive at the hospital without clear information regarding their prognosis, diagnosis, or condition. This context is more appropriately framed within an unsupervised framework rather than a supervised one. Cutting-edge biomedical research is unsupervised.

Clustering algorithms are unsupervised computational methods which can split data into significant groups, called *clusters*, that otherwise would probably be unnoticeable to human beings. The research question we investigate in this study concerns the efficacy of modern computational clustering methods: is there a clustering method capable of grouping patients from these three datasets into two clinically meaningful clusters? We tried eight different methods, and DBSCAN was the only technique that succeeded in this task.

We organize the rest of the article as follows. After this Introduction, we describe the TARGER-NBL dataset and we outline the Genoa and Shanghai datasets in "Datasets" section. We then briefly describe the clustering computational algorithms employed in "Methods" section and the study results in "Results" section. Finally, we outline a discussion and the main conclusions of this work, including limitations and future developments, in "Discussion" and "Conclusions" sections. We represent all the steps of our study, both the manual ones and the computational ones, in Fig. 1.

Datasets

In this clustering study, we analyze three independent, open, deidentified datasets of electronic health records: the Genoa dataset [10], the Shanghai dataset [11], and the TARGET-NBL stage 4 dataset [12–14].

We initially conducted a dataset search, by looking for scientific articles on neuroblastoma electronic medical records which included a public dataset in their

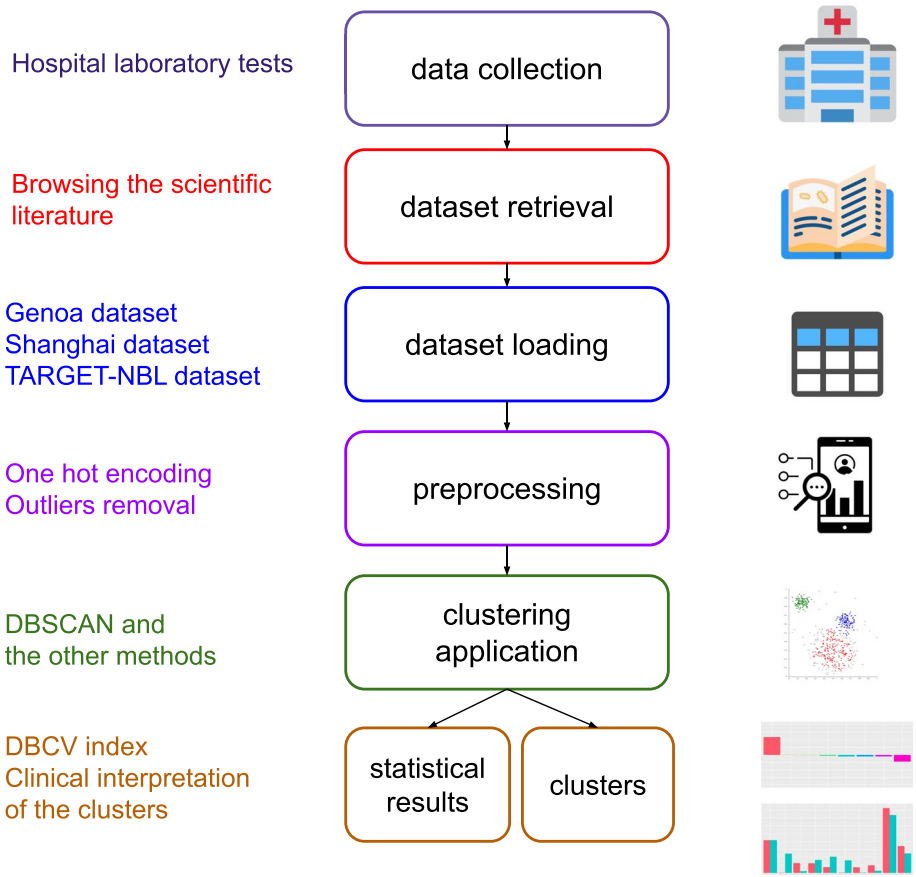

Fig. 1 Schematic representation of our study process. The data collection phase was conducted by the original datasets curators in their hospitals [10, 11, 13]. We conducted the dataset retrieval via scientific literature search engines. The steps from dataset loading to statistical results and clusters refer to the computational analysis presented in this study. All the illustration images were released online publicly under a Creative Commons license: the hospital icon from IconScout.com, the book icon from IconScout.com, the table icon from Wikimedia Commons, the barchart icon from IconScout.com, the clusters image from Wikimedia Commons

supplementary information (Fig. 1). We found five public datasets which we cleaned and released in our repository [16, 17] online for free. Among these five datasets, three were too small for any computational analysis, and thus we discarded them: dataCK2018 with 20 patients [22], dataEV2013 with 19 patients [23], and dataYBC2019 with 7 patients [24].

The only two datasets with a sufficient number of patients were *dataBB2013* containing data from 121 patients, which we have renamed here the Genoa dataset, and *dataYM2018* consisting of data from 169 patients, which we have renamed here the Shanghai dataset.

The Genoa dataset was collected at Gaslini hospital in Genoa (Italy, EU) between 1990 and 2004 [10], and contains data from 121 single patients, each having 11 clinical variables The Shanghai dataset was collected at Children's Hospital of Fudan University in Yangpu (Shanghai, China) between 2010 and 2015 [11], and consists of data of 169 single patients. Each patient profile includes 13 clinical features. We described the Genoa dataset and the Shanghai dataset precisely in the [16] article.

The TARGET-NBL stage 4 dataset contains data from 91 patients collected at the Children's Hospital of Philadelphia (Pennsylvania, USA) and at other hospitals in the USA, gathered over several years, primarily from 2005 to 2017. This dataset consists of 16 clinical variables, some having missing values. We report a quantitative description of this dataset in Table 1 and Table 2.

Our clustering study has been possible because the original curators of these datasets decided to release these datasets publicly online without restrictions, after obtaining the consent from the Institutional Review Board (IRB) of their hospital, following the open science best practices [25, 26].

Methods

Preprocessing We initially applied the clustering methods to the three datasets without removing any outliers from them. This way, we obtained satisfactory results on the Genoa and on the TARGET-NBL datasets, but not on the Shanghai dataset. On the Shanghai original dataset, we had to remove the most different 41% outliers: we computed the average record value for all the patients, we ranked them in descending order, and then we removed the top 20% and the bottom 20%. By doing so, we removed 69 outliers and ultimately obtained valuable clustering results. We used one-hot encoding to handle the *first event* variable of this dataset; all the other features are numeric and therefore do not need this step (Fig. 1).

Given the high heterogeneity of this dataset [11], which is further amplified by the general heterogeneity of neuroblastoma data [27–29], it was necessary to remove 41% of data points from the Shanghai dataset. This step allowed DBSCAN to identify relevant clusters among the remaining data. Without this adjustment, DBSCAN assigns all data points to the noise clusters.

Clustering algorithms Before applying the clustering algorithms, we decided to set the number of clusters to two, similar to what has been done in other neuroblastoma studies

Table 1 TARGET-NBL stage-4 patients dataset, quantitative characteristics of the binary features. NA: missing values. More information on this dataset can be found in [12–14]

Feature	Meaning	Value	#	%
Differentiating grade	undiff. or poorly diff.	0	72	79.121
Differentiating grade	differentiated	1	6	6.593
Differentiating grade	missing	NA	13	14.286
Histology	unfavorable	0	85	93.407
Histology	favorable	1	2	2.198
Histology	missing	NA	4	4.395
MYCN amplification	false	0	68	74.725
MYCN amplification	true	1	23	25.275
Sex	woman	0	36	39.560
Sex	man	1	55	60.440
Vital status	alive	0	38	41.758
Vital status	dead	1	53	58.242

Table 2 TARGET-NBL stage-4 patients dataset, quantitative characteristics of the numerical features. First event: 0 event, 1 progression, 2 relapse, 3 death. std. dev.: standard deviation. MKI: Mitotic Karyorrhectic Index. Percent Necrosis: percentage of dead tumor cells within the cancerous tissue, which can be an important factor in the pathology and prognosis of the disease. Percent Tumor: percentage of tumor cells present in a biopsy or surgical specimen compared to normal cells. Percent Tumor vs Stroma: percentage of tumor cells to stromal cells within a tumor sample. The stroma is the supportive tissue surrounding the tumor cells, which includes connective tissue, blood vessels, and immune cells. Ploidy Value: number of sets of chromosomes in the tumor cells. Neuroblastoma can be classified as either diploid (normal chromosome number) or aneuploid (abnormal chromosome number). Aneuploid tumors often indicate a more aggressive disease and are associated with poorer outcomes. In contrast, diploid tumors may have a better prognosis. Supenhancer group: 0 *ATRX*, 1 *MES*, and 2 *MYCN*, as defined in [12]. More information on this dataset can be found in [12–14]

Feature	Median	Mean	Range	Std. dev.	#NAs
Age at diagnosis days	1328.000	1618.813	[550, 6021]	1112.958	
Event Free Survival Time in Days	709.000	1130.341	[87, 4948]	1030.418	
First event	2.000	1.746	[0, 3]	0.842	28
MKI	1.000	1.000	[0, 2]	0.799	18
Overall Survival Time in Days	1330.000	1498.923	[87, 4948]	1050.606	
Percent Necrosis	10.000	16.383	[0, 90]	18.583	27
Percent Tumor	80.000	71.098	[5, 97.5]	23.518	25
Percent Tumor vs Stroma	80.000	70.841	[5, 98]	23.076	28
Ploidy Value	1.030	1.218	[1, 3]	0.377	
Superenhancer group	0.000	0.681	[0, 2]	0.880	
Years from diagnosis to last follow-up	4.000	4.055	[0, 14]	2.911	

involving clustering analyses [30–34]. In our work, we employed several clustering algorithms using the `scikit-learn` Python library [35], each based on distinct principles and requiring specific hyperparameter tuning for optimal performance.

k-Means [36] is a popular clustering method that partitions data into a predefined number of clusters by minimizing within-cluster variance. The primary parameter for this algorithm is the number of clusters, which must be specified in advance. *k*-Means assigns points to clusters iteratively until convergence. Spectral Clustering [37] leverages the eigenvalues of a similarity matrix to reduce dimensionality before applying clustering. Key parameters include the number of clusters, the kernel coefficient γ gamma for the radial basis function (RBF) kernel, and the number of eigenvectors used in the clustering process. This method is particularly effective for non-linear data separations.

Agglomerative Clustering [38], including Ward's method [39] and other linkage strategies such as complete, average, and single linkage, performs hierarchical clustering by successively merging pairs of clusters. Important parameters include the number of clusters, the type of linkage criterion used, and the distance metric applied, such as Euclidean, Manhattan, or cosine. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [40] builds a clustering feature tree, making it suitable for large datasets. Its key parameters include the number of clusters, the threshold for forming sub-clusters, and the branching factor, which determines the maximum number of sub-clusters within a node.

Gaussian Mixture [41] models data as a combination of multiple Gaussian distributions, allowing for flexible cluster shapes. Essential parameters include the number of components (clusters) and the covariance type, which can be full, tied, diagonal, or spherical, reflecting different assumptions about the cluster shapes.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [42] forms clusters based on dense regions of points. Its two key parameters are the maximum distance between points in a neighborhood and the minimum number of points required to form a dense cluster. It can identify clusters of varying shapes and sizes while labeling outliers.

Affinity Propagation [43] identifies clusters by exchanging messages between points until convergence. The damping parameter, which controls the extent of message propagation, is crucial for avoiding oscillations during the clustering process.

MeanShift [44] locates high-density regions by shifting data points toward density peaks. The bandwidth parameter, which determines the size of the search window, is critical for accurate clustering.

In most of these algorithms (for example, *k*-Means), the numberof clusters is specified a priori (in our case, 2). For algorithms where this parameter cannot be set in advance (for example, DBSCAN), we disregarded parameter combinations that did not yield the desired number of clusters. We report the values of all the optimized hyperparameters for these algorithms in Table 3.

We selected eight of the most common clustering methods employed in the health informatics literature, for which an open source Python implementation is available [35]. Among these algorithms, DBSCAN is known to be particularly effective in biomedical datasets [45, 46]. This step refers to the clustering application phase in the flowchart of Fig. 1.

Evaluation metric Finally, these clustering methods have been evaluated using the Density-Based Clustering Validation (DBCV) metric [47] (with the Felipe Alves Siqueira's Python implementation [48]), which assesses clustering quality by balancing density-based validation criteria. The code iterates through various parameter combinations, aiming to maximize clustering performance.

Here we decided not to utilize common metrics for clustering internal assessments (such as Silhouette coefficient [49], Davies-Bouldin index [50], Calinski-Harabasz index [51], Dunn index [52], Shannon entropy [53], and Gap statistic [54]) because these indexes work only on convex-shaped clusters, and not on concave-shaped clusters. DBSCAN, in fact, can produce both convex or concave clusters, which can be correctly assessed by the DBCV score.

We decided to employ several density-based clustering algorithms and to use the DBCV index for evaluation because of their capability to handle non-convex or irregularly-shaped clusters.

We then studied the clusters assigned by the top performing method, DBSCAN, which is also the only method which produced sufficient results (Fig. 1).

Table 3 Optimized hyperparameters for the tested algorithms on the two datasets analyzed. These hyperparameters refer to the functions implemented in `scikit-learn` Python library [35]

Genoa dataset
Affinity Propagation: damping = 0.97465
Agglomerative Clustering: linkage = single, metric = euclidean
BIRCH: branching factor = 4, threshold = 0.88862
DBSCAN: epsilon = 0.792016, min samples = 8
Gaussian Mixture: covariance type = full
k-Means: init = k-means++, n init = auto, max iter = 300, algorithm = lloyd
Spectral Clustering: gamma = 0.88862, n components = 2
Ward: linkage = ward, metric = euclidean
Shanghai dataset
Agglomerative Clustering: linkage = single, metric = L1
BIRCH: branching factor = 2, threshold = 0.888623
DBSCAN: epsilon = 0.526646, min samples = 4
Gaussian Mixture: covariance type = diag
k-Means: init = k-means++, n init = auto, max iter = 300, algorithm = lloyd
Spectral Clustering: gamma = 0.388815, n components = 2
Ward: linkage = ward, metric = euclidean
TARGET-NBL stage-4 dataset
Agglomerative Clustering: linkage = average, metric = cosine
BIRCH: branching factor = 4, threshold = 0.888624
DBSCAN: epsilon = 0.35019, min samples = 4
Gaussian Mixture: covariance type = full
k-Means: init = k-means++, n init = auto, max iter = 300, algorithm = lloyd
Spectral Clustering: gamma = 0.0001, n components = 9
Ward: linkage = ward, metric = euclidean

Results

We applied all the unsupervised clustering methods described in the previous section to the three datasets described in the Datasets section. We optimized the hyperparameters of the algorithms (Table 3) and tested the algorithms' configurations through the DBCV index [47].

All the algorithms obtained negative values of DBCV, except DBSCAN which attained $DBCV = +0.5968$ on the Genoa dataset, $DBCV = +0.49256$ on the Shanghai dataset, and $DBCV = +0.86032$ on the TARGET-NBL (Fig. 2). The worst and minimum value of DBCV is -1, while the best and maximum value of DBCV is $+1$. Some of the algorithms (Mean-Shift algorithm on the Genoa dataset and Affinity Propagation on both datasets) achieved $DBCV = -\infty$, indicating the possible presence of a technical bug in the implementation of the Python `DBCV` package [48].

The top performing DBSCAN method identified two clusters in each of the three datasets having different sizes (Table 4).

We then scrutinized the results obtained by DBSCAN and observed the feature partitions of the patient clusters identified by this algorithm (Fig. 3a).

In the Genoa dataset results, three clinical features completely partition the dataset patients into two clusters: INRG risk classification, outcome, and *MYCN* amplification. The INRG risk classification feature clearly partitions the data into cluster 0 for

high risk and into cluster 1 for intermediate or low risk. Similarly, the outcome variable assigns all the *dead of disease* patients into cluster 0, and all the patients with the *alive in complete remission* or *alive with disease* in cluster 1. Also the *MYCN* gene amplification clearly separates the patients into the two clusters: all patients having *MYCN* amplification were assigned to the cluster 0, and all the patients without were set to cluster 1.

Some other clinical variables showed average differences among the two clusters (age at diagnosis, progression free survival months, overall survival months, ferritin), but did not partition the patients into two completely separated groups (Fig. 3a). These results show the partitioning power of the three clinical variables INRG risk classification, outcome, and *MYCN* gene amplification.

Regarding the Shanghai dataset, a few variables indicated differences in the composition of the two clusters of patients: age, months time of overall survival and outcome (Fig. 3b). Among these three medical variables, only outcome completely separated the patients into two clusters. All the alive patients and all the patients that were lost during follow-up were assigned to cluster 0, and all the patients dead of disease were assigned to cluster 1.

Only the sex variable completely discriminated the two clusters of patients in the TARGET-NBL stage-4 dataset (Fig. 3c): all the women were included in cluster 0, while all the men were put in cluster 1. Other variables, however, indicated a clear separation between the two clusters. DBSCAN assigned to cluster 0 the majority of patients with more years from diagnosis to last follow-up, more overall survival days, and higher percentage of necrosis. The patients assigned by DBSCAN to cluster 1, instead, had more severe first neuroblastoma events and a higher MKI score. The patients of the two clusters had several other minor differences in other clinical features (age at diagnosis, percentage of tumor, percentage of tumor versus stroma, and ploidy value), but we consider these differences irrelevant (Fig. 3c).

For all the three datasets analyzed, DBSCAN assigned some patients to the noise −1 cluster: 33 patients out of 121 in the Genoa dataset, 88 out of 100 in the Shanghai dataset (where 69 patients were already removed during the preprocessing phase), and 82 out of 91 in the TARGET-NBL dataset. The algorithm considered these data points as outliers and inserted them into no cluster.

Tests of robustness To verify the robustness of the results we obtained, we performed a subsampling analysis without repetitions. For each of the ten iterations, we randomly selected 90% of the data points, applied DBSCAN, and saved the results measured

Table 4 Composition of the clusters identified by DBSCAN. removed: number of patients removed during preprocessing. Each cell represents a number of patients

			DBSCAN		
Dataset	**Size**	**Removed**	**Noise cluster**	**Cluster 0**	**Cluster 1**
Genoa dataset	121		33	76	12
Shanghai dataset	169	69	88	6	6
TARGET-NBL stage-4 dataset	91		82	5	4

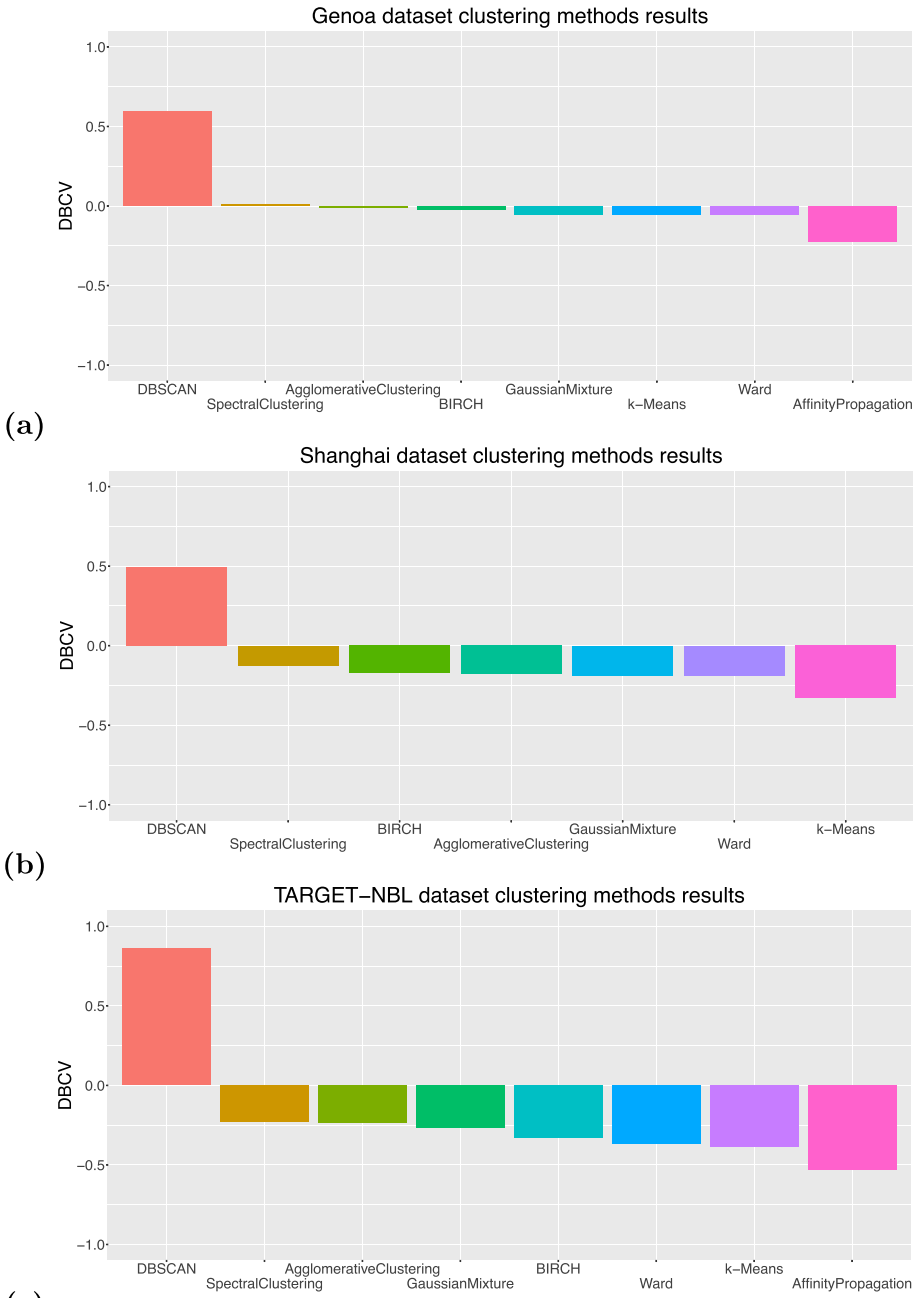

Fig. 2 Clustering results obtained by the clustering algorithms on the two dataset. **a** Upper image: Genoa dataset. **b** Mid image: Shanghai dataset. **c** Bottom image: TARGET-NBL dataset. We employed the DBCV (Density-Based Clustering Validation) index for clustering internal assessment. DBCV index interval: $[-1; +1]$, with -1 meaning worst possible clustering and $+1$ meaning perfect clustering. We also tried Mean-Shift algorithm on the Genoa and TARGET-NBL datasets, Affinity Propagation on the Genoa and Shanghai datasets, but our scripts generated $-\infty$ for these cases, probably for some code implementation bugs

though the DBCV index. We reported these results in Table 5. These outcomes confirm the robustness of our approach, indicating even an average improvement compared to the results obtained on the whole datasets.

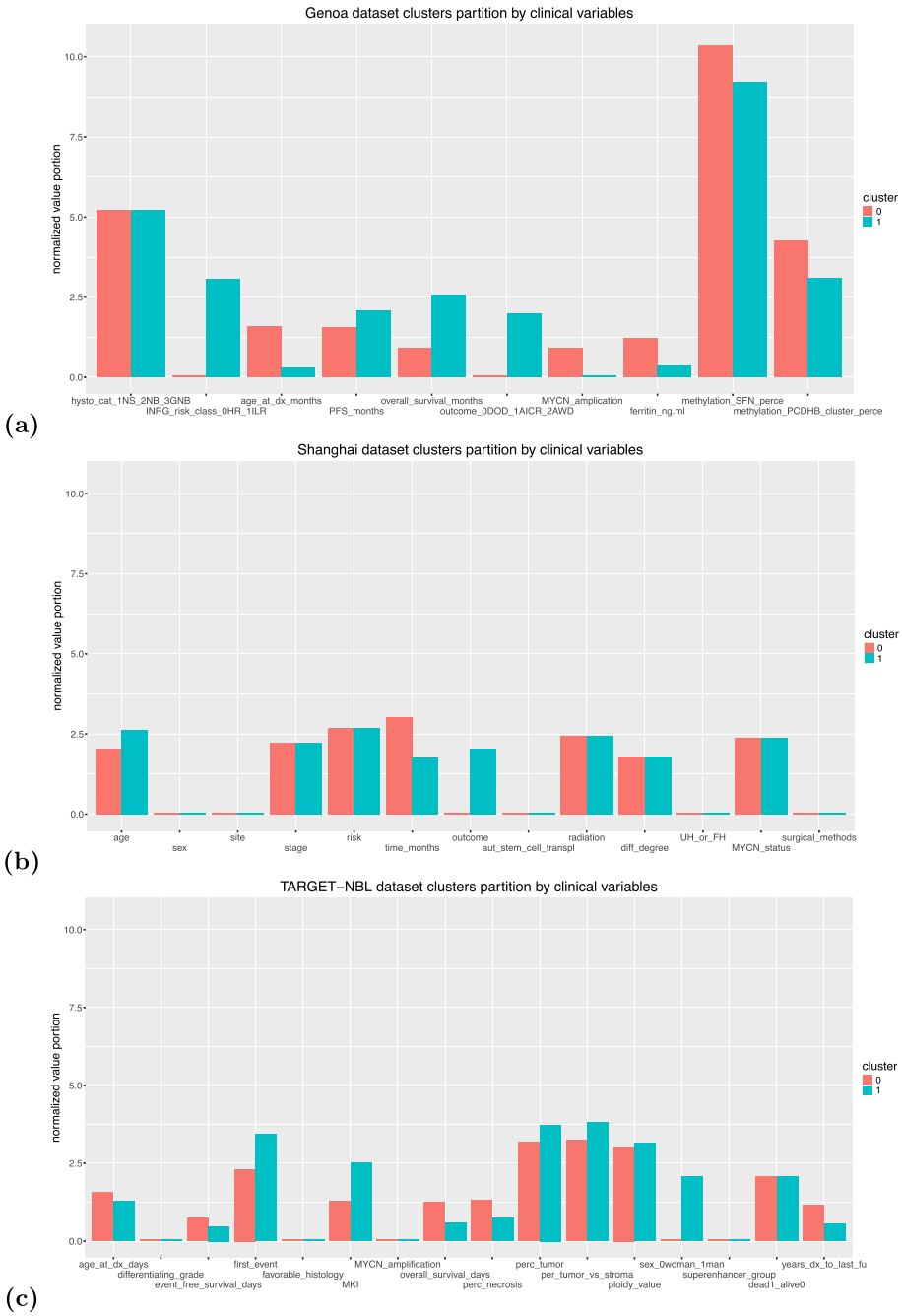

Fig. 3 Partition of the clinical features among the two clusters identified by DBSCAN. **a** Top image: representation of the normalized values of the clinical variables of the Genoa dataset in the subset of patients of the 0 cluster (red bars) and in the subset of patients of the 1 cluster (green bars). **b** Mid image: representation of the normalized values of the clinical variables of the Shanghai dataset in the subset of patients of the 0 cluster (red bars) and in the subset of patients of the 1 cluster (green bars). **c** Bottom image: representation of the normalized values of the clinical variables of the TARGET-NBL dataset in the subset of patients of the 0 cluster (red bars) and in the subset of patients of the 1 cluster (green bars). Each bar represents the average value for that specific factor for that specific cluster. We listed the meaning of the clinical variables in "Datasets" section and in [16]

Discussion

Neuroblastoma is a rare type of childhood cancer characterized by significant heterogeneity in its clinical presentation and in the underlying biological mechanisms driving its onset and development [9]. Clinical parameters such as age, stage, and MYCN amplification status are used at diagnosis to assign a risk group. Risk assignment is used to determine the most appropriate treatment according to international guidelines, such as the INRG pretreatment risk assignment system [55]. Despite multimodal treatments, disease outcomes remain poor for high-risk patients.

The scientific community has conducted numerous studies in recent years, proposing new treatments, prognostic factors, and therapeutic targets [5, 9, 56–59]. Supervised and unsupervised approaches, such as classification and clustering, are two most commonly encountered knowledge-discovery techniques [60]. Supervised approaches have been widely reported in the literature to analyze multi-omics data, EHRs and medical images to accurately stratify patients with neuroblastoma and to predict prognosis [61]. However, to the best of our knowledge, unsupervised approaches have not previously been reported for neuroblastoma EHRs analyses. The problem of clustering in general deals with partitioning a data set consisting of n points embedded in m-dimensional space into k distinct set of clusters of similar data points [60]. Traditional clustering algorithms use distance metrics such as Euclidean distance to assess similarity among data points. Although these metrics are suitable for features with purely numeric values, they fail to capture the similarity of data elements when attributes are categorical or mixed [60]. Clustering mixed data sets into meaningful groups is a well-known challenging task [60]. Discretization and dummy coding are straightforward and intuitive methods for creating a homogeneous dataset consisting solely of categorical data, enabling the application of classical techniques. However, these methods can distort the original data, potentially introducing bias. In the literature, a variety of clustering algorithms have been specifically designed to handle mixed data [62].

Previously published studies exploring appropriateness of unsupervised machine-learning methods for "heterogeneous" or "mixed" data used simulated and large real world datasets [62]. Pediatric datasets represented emblematic use cases for testing unsupervised clustering methods on mixed data because pediatric disease can be heterogeneous and the number of patients enrolled in the study rarely exceeds one thousand patients. The three datasets used in the present retrospective study are publicly available [10–14].

Datasets are composed of mixed EHRs features covering patients with neuroblastoma of all risk groups, as is the case of the Genoa and Shanghai datasets, as well as, the subset

Table 5 Results obtained by DBSCAN on ten subsampled datasets. DBSCAN refers to the optimized hyperparameters listed in Table 3. DBCV index interval: [−1; +1], the higher the better

Dataset	Average DBCV index	Standard Deviation
Genoa dataset	0.858	±0.0024
Shanghai dataset	0.779	±0.0212
TARGET-NBL stage-4 dataset	0.941	±0.0006

of high-risk patients, as is the case of the TARGET-NBL dataset. The dataset with the highest number of patients is the Shanghai dataset with 169 patients. Therefore, feasibility of unsupervised machine-learning methods for small, heterogeneous and mixed datasets remains to be demonstrated. Our analyses demonstrated that DBSCAN was the unique method able to identify clusters with a clear segregation in the NB datasets. Heterogeneity of the datasets brought the DBSCAN algorithm to cluster a large number of patients of the Shanghai and TARGET-NBL datasets into the noise group. Previous studies have highlighted the effectiveness of a modified version of the DBSCAN algorithm for mixed data analysis, but none have tested the method on real world pediatric datasets [63].

Our density-based spatial clustering analysis assigned significant portions of the Genoa dataset to the two clusters (63% to the first cluster and 10% to the second cluster), but only few patients in the other two datasets. In fact, for the TARGET-NBL dataset only 5% of patients were placed in the first cluster and only 4% of them were assigned into the second cluster, leaving 91% of patients in the noise cluster. For the Shanghai dataset, we initially used all the dataset, but no method could find any cluster in it. So we had to remove the top 20% and the bottom 20% outliers and to work only on 100 patients out of 69. BSCAN placed 6 of these patients' profiles in the first cluster and 6 in the second cluster, assigning the remaining 88 to the noise cluster. For this dataset, we assigned 92% of patients to no cluster.

We recognize that these clusters are small compared to those in other health informatics studies involving cluster analysis. However, with a rare disease such as neuroblastoma, where datasets are rare, data are small and so are the number of patients, we consider this result to be relevant for the medical significance of the clusters identified. In a landscape where seven traditional clustering algorithms found nothing, at least DBSCAN was able to find thus: even if small, these clusters make clinical sense, and therefore confirm the clustering capability of DBSCAN.

Moreover, neuroblastoma data are known to be extremely heterogeneous [27–29], making them difficult to analyze and process, especially in an unsupervised scenario. The consequence of this heterogeneity is that the the noise clusters identified by DBSCAN have a huge size, compared to te data clusters.

We used the pre-treatment risk groups, when available, and the outcome to evaluate the prognostic value of each cluster. Analysis of the main characteristics differentiating clusters revealed a clear association between clusters and prognosis in the Genoa and Shanghai datasets, thereby confirming the feasibility and potential utility of DBSCAN on small, heterogeneous and mixed data analysis.

Conclusions

Neuroblastoma is a rare cancer that affects around five thousand infants worldwide, and datasets for scientific research on this disease are scarce. In this study, we leveraged three open, unrestricted, public datasets of electronic health records of patients diagnosed with this pediatric cancer to identify clustering methods which can discriminate significant subgroups of patients. To do so, we took advantage of eight different unsupervised clustering methods and of the DBCV metric implemented in Python, and then analyzed the the medical meaning of the clusters identified.

Our results show that DBSCAN paired with DBCV implemented through an open source programming language, applied to open data, can produce significant results and outcomes that have clinical meaning. Health informatics researchers and analysts can now leverage our discoveries and, when conducting a clustering analysis on small dataset of EHRs of neuroblastoma, can choose to use DBSCAN rather than utilizing more traditional techniques *k*-Means and DBCV rather than utilizing more traditional metrics such as Silhouette coefficient. We highlight the fact that DBSCAN is not only the best performing method among the ones that we employed, but it is also the only one that was able to obtain sufficient results, according to the DBCV index.

Moreover, our study stands for its adherence to the principles of open science: we utilized only open source software code (in Python) to analyze only open data (of electronic medical records), and are publishing the results in an open access journal. Anyone with a technological device can read our findings, reuse our software code, and reutilize the datasets we analyzed here.

Regarding limitations, we need to report that unfortunately we employed datasets with many different features: it would have been better to use datasets having all the same variables, but it was impossible. To the best of our knowledge, no other open EHRs datasets than the three we employed exist in the scientific literature nowadays. Moreover, our clustering analysis did not identify any relevant clinical feature that might impact neuroblastoma treatment or research. In the future, we plan to repeat a similar unsupervised computational analysis on medical records' data of patients with glioblastoma [64] and other diseases.

Abbreviations

ATRX	Protein-coding gene, ATRX Chromatin Remodeler
BHLH	Basic helix-loop-helix
BIRCH	Balanced iterative reducing and clustering using hierarchies
CSV	Comma-separated values
DBCV	Density-Based Clustering Validation
DBSCAN	Density-based spatial clustering of applications with noise
Dx	Diagnosis
EHRs	Electronic health records
INRG	International Neuroblastoma Risk Group
INSS	International Neuroblastoma Staging System
LDH	Lactic acid dehydrogenase level
MES	Mesenchymal
MYCN	Protein-coding *MYCN* Proto-Oncogene, BHLH Transcription Factor
NB, NBL	Neuroblastoma
PCDHB	Protocadherin Beta Cluster
RBF	Radial basis function
RINB	Italian Registry of Peripheral Neuroblastoma
TARGET	Therapeutically Applicable Research to Generate Effective Treatments

Acknowledgements
The authors thank the original datasets curators who decided to release the three datasets publicly online for free.

Clinical trial
This study is not a clinical trial.

Authors' contributions
D.Ch. conceived the study, made the plots, and wrote most of the article. L.O. performed the computational analysis and contributed to the writing of the manuscript. D.Ca. performed the interpretation of the clinical results and contributed to the writing of the manuscript. All authors approved the current version for submission.

Funding
The work of D.Ch. 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) (project code F/310240/01-04/X56) programme within the framework "Innovation Agreements" (Accordi per l'Innovazione) and is partially supported by Ministero dell'Università

e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to 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 Genoa and Shanghai datasets used in this project are openly, publicly available for free under the CC BY-NC 4.0 license at: https://davidechicco.github.io/neuroblastoma_EHRs_data

The TARGET-NBL stage-4 dataset employed in this study is openly, publicly available for free under the CC BY-NC 4.0 license at: https://figshare.com/articles/dataset/Clinical_and_mutational_information_of_analyzed_94_Stage_4_neuroblastoma_cases_in_TARGET_cohort_/13609375

Our Python software code is publicly available at: https://colab.research.google.com/drive/1NJ3HaljQos0JNRjt4DGF-FSEOcSYVO7c?usp=sharing.

Declarations

Ethics approval and consent to participate

Permission to collect and analyze the data of patients involved in this study has been obtained from hospital ethical committees by the original datasets' curators [10, 11, 13].

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 16 April 2025 Accepted: 3 June 2025
Published online: 12 June 2025

References

1. Nong J, Su C, Li C, Wang C, Li W, Li Y, et al. Global, regional, and national epidemiology of childhood neuroblastoma (1990–2021): a statistical analysis of incidence, mortality, and DALYs. eClinicalMedicine. 2025;79. https://doi.org/10.1016/j.eclinm.2024.102964.
2. Cangelosi D, Pelassa S, Morini M, Conte M, Bosco MC, Eva A, et al. Artificial neural network classifier predicts neuroblastoma patients' outcome. BMC Bioinformatics. 2016;17(12):83–93. https://doi.org/10.1186/s12859-016-1194-3.
3. Francescatto M, Chierici M, Rezvan Dezfooli S, Zandonà A, Jurman G, Furlanello C. Multi-omics integration for neuroblastoma clinical endpoint prediction. Biol Direct. 2018;13:1–12. https://doi.org/10.1186/s13062-018-0207-8.
4. Melaiu O, Chierici M, Lucarini V, Jurman G, Conti LA, De Vito R, et al. Cellular and gene signatures of tumor-infiltrating dendritic cells and natural-killer cells predict prognosis of neuroblastoma. Nat Commun. 2020;11(1):5992. https://doi.org/10.1038/s41467-020-19781-y.
5. Cangelosi D, Morini M, Zanardi N, Sementa AR, Muselli M, Conte M, et al. Hypoxia predicts poor prognosis in neuroblastoma patients and associates with biological mechanisms involved in telomerase activation and tumor microenvironment reprogramming. Cancers. 2020;12(9):2343. https://doi.org/10.3390/cancers12092343.
6. Cangelosi D, Brignole C, Bensa V, Tamma R, Malaguti F, Carlini B, et al. Nucleolin expression has prognostic value in neuroblastoma patients. eBioMedicine. 2022;85. https://doi.org/10.1016/j.ebiom.2022.104300.
7. Chicco D, Sanavia T, Jurman G. Signature literature review reveals AHCY, DPYSL3, and NME1 as the most recurrent prognostic genes for neuroblastoma. BioData Min. 2023;16(1):7. https://doi.org/10.1186/s13040-023-00325-1.
8. King J, Patel V, Jamoom EW, Furukawa MF. Clinical benefits of electronic health record use: national findings. Health Serv Res. 2014;49(1pt2):392–404. https://doi.org/10.1111/1475-6773.12135.
9. Chicco D, Haupt R, Garaventa A, Uva P, Luksch R, Cangelosi D. Computational intelligence analysis of high-risk neuroblastoma patient health records reveals time to maximum response as one of the most relevant factors for outcome prediction. Eur J Cancer. 2023;193:113291. https://doi.org/10.1016/j.ejca.2023.113291.
10. Banelli B, Merlo DF, Allemanni G, Forlani A, Romani M. Clinical potentials of methylator phenotype in stage 4 high-risk neuroblastoma: an open challenge. PLoS ONE. 2013;8(5):e63253. https://doi.org/10.1371/journal.pone.0063253.
11. 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.
12. Kimura S, Sekiguchi M, Watanabe K, Hiwatarai M, Seki M, Yoshida K, et al. Association of high-risk neuroblastoma classification based on expression profiles with differentiation and metabolism. PLoS ONE. 2021;16(1):e0245526. https://doi.org/10.1371/journal.pone.0245526.
13. Childhood Cancer Clinical Data Commons (C3DC). C3DC Studies dbGaP accession: phs000467, 2024. https://clinicalcommons.ccdi.cancer.gov./phs000467. URL visited on 18th December. Accessed 5 June 2025.
14. National Cancer Institute, Center for Cancer Genomics. TARGET's Study of Neuroblastoma, 2024. https://www.cancer.gov/ccg/research/genome-sequencing/target/studied-cancers/neuroblastoma#targets-neuroblastoma-nbl-project. URL visited on 18th December.
15. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3(1). https://doi.org/10.1038/sdata.2016.18.
16. Chicco D, Cerono G, Cangelosi D. A survey on publicly available open datasets derived from electronic health records (EHRs) of patients with neuroblastoma. Data Sci J. 2022;21:17. https://doi.org/10.5334/dsj-2022-017.

17. Chicco D, Cangelosi D, Cerono G. Neuroblastoma Electronic Health Records Open Data Repository, 2025. https://davidechicco.github.io/neuroblastoma_EHRs_data/. URL visited on 6th January.
18. Volchenboum SL, Cox SM, Heath A, Resnick A, Cohn SL, Grossman R. Data commons to support pediatric cancer research. Am Soc Clin Oncol Educ Book. 2017;37:746–52. https://doi.org/10.1200/EDBK_175029.
19. International Neuroblastoma Risk Group. INRG Data Commons, 2025. https://inrgdb.org/. URL visited on 6th January.
20. Plana A, Furner B, Palese M, Dussault N, Birz S, Graglia L, et al. Pediatric cancer data commons: federating and democratizing data for childhood cancer research. JCO Clin Cancer Inform. 2021;5:1034–43. https://doi.org/10.1200/cci.21.00075.
21. Volchenboum S, Cohen E, Furner B, the Pediatric Cancer Data Commons Team. Pediatric Cancer Data Commons, 2025. https://commons.cri.uchicago.edu/. URL visited on 6th January.
22. Kim C, Choi YB, Lee JW, Yoo KH, Sung KW, Koo HH. Excellent treatment outcomes in children younger than 18 months with stage 4 MYCN nonamplified neuroblastoma. Korean J Pediatr. 2018;61(2):53. https://doi.org/10.3345/kjp.2018.61.2.53.
23. Villamón E, Berbegall AP, Piqueras M, Tadeo I, Castel V, Djos A, et al. Genetic Instability and Intratumoral Heterogeneity in Neuroblastoma with MYCN Amplification Plus 11q Deletion. PLoS ONE. 2013;8(1):e53740. https://doi.org/10.1371/journal.pone.0053740.
24. Choi YB, Son MH, Cho HW, Ma Y, Lee JW, Kang ES, et al. Safety and immune cell kinetics after donor natural killer cell infusion following haploidentical stem cell transplantation in children with recurrent neuroblastoma. PLoS ONE. 2019;14(12):e0225998. https://doi.org/10.1371/journal.pone.0225998.
25. Pisani E, Aaby P, Breugelmans JG, Carr D, Groves T, Helinski M, et al. Beyond open data: realising the health benefits of sharing data. BMJ. 2016;355. https://doi.org/10.1136/bmj.i5295.
26. Kostkova P, Brewer H, De Lusignan S, Fottrell E, Goldacre B, Hart G, et al. Who owns the data? Open data for healthcare. Front Public Health. 2016;4:7. https://doi.org/10.3389/fpubh.2016.00007.
27. Mora J, Cheung NKV, Gerald WL. Genetic heterogeneity and clonal evolution in neuroblastoma. Br J Cancer. 2001;85(2):182–9. https://doi.org/10.1054/bjoc.2001.1849.
28. Rodriguez-Fos E, Planas-Fèlix M, Burkert M, Puiggròs M, Toedling J, Thiessen N, et al. Mutational topography reflects clinical neuroblastoma heterogeneity. Cell Genomics. 2023;3(10):100402. https://doi.org/10.1016/j.xgen.2023.100402.
29. Lundberg KI, Treis D, Johnsen JI. Neuroblastoma heterogeneity, plasticity, and emerging therapies. Curr Oncol Rep. 2022;24(8):1053–62. https://doi.org/10.1007/s11912-022-01270-8.
30. Han Y, Ye X, Wang C, Liu Y, Zhang S, Feng W, et al. Integration of molecular features with clinical information for predicting outcomes for neuroblastoma patients. Biol Direct. 2019;14(1). https://doi.org/10.1186/s13062-019-0244-y.
31. Baali I, Acar DAE, Aderinwale TW, HafezQorani S, Kazan H. Predicting clinical outcomes in neuroblastoma with genomic data integration. Biol Direct. 2018;13(1). https://doi.org/10.1186/s13062-018-0223-8.
32. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, et al. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. Front Genet. 2018;9. https://doi.org/10.3389/fgene.2018.00477.
33. van Noesel MM, van Bezouw S, Voûte PA, Herman JG, Pieters R, Versteeg R. Clustering of hypermethylated genes in neuroblastoma. Genes Chromosom Cancer. 2003;38(3):226–33. https://doi.org/10.1002/gcc.10278.
34. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Versteeg R, et al. A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients. Mol Cancer. 2010;9(1):185. https://doi.org/10.1186/1476-4598-9-185.
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
36. Steinley D. K-means clustering: a half-century synthesis. Br J Math Stat Psychol. 2006;59(1):1–34. https://doi.org/10.1348/000711005X48266.
37. Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst. 2001. https://proceedings.neurips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html.
38. Ackermann MR, Blömer J, Kuntze D, Sohler C. Analysis of agglomerative clustering. Algorithmica. 2014;69:184–215. https://doi.org/10.1007/s00453-012-9717-4.
39. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J Classif. 2014;31:274–95. https://doi.org/10.1007/s00357-014-9161-z.
40. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Rec. 1996;25(2):103–14. https://doi.org/10.1145/235968.233324.
41. Yang MS, Lai CY, Lin CY. A robust EM clustering algorithm for Gaussian mixture models. Pattern Recognit. 2012;45(11):3950–61. https://doi.org/10.1016/j.patcog.2012.04.031.
42. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst. 2017;42(3):1–21. https://doi.org/10.1145/3068335.
43. Dueck D. Affinity propagation: clustering data by passing messages. University of Toronto; 2009. http://hdl.handle.net/1807/17755.
44. Cheng Y. Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Mach Intell. 1995;17(8):790–9. https://doi.org/10.1109/34.400568.
45. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. IEEE Rev Biomed Eng. 2010;3:120–54. https://doi.org/10.1109/RBME.2010.2083647.
46. Wiwie C, Baumbach J, Röttger R. Comparing the performance of biomedical clustering methods. Nat Methods. 2015;12(11):1033–8. https://doi.org/10.1038/nmeth.3583.
47. Moulavi D, Jaskowiak PA, Campello RJ, Zimek A, Sander J. Density-based clustering validation. In: Proceedings of SDM24 – the 2014 SIAM International Conference on Data Mining. SIAM; 2014. pp. 839–847. https://doi.org/10.1137/1.9781611973440.96.
48. Felipe Alves Siqueira. Fast Density-Based Clustering Validation (DBCV), 2024. https://github.com/FelSiq/DBCV. URL visited on 16th December.

49. 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.
50. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979;PAMI-1(2):224–227. https://doi.org/10.1109/tpami.1979.4766909.
51. Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Stat Theory Methods. 1974;3(1):1–27. https://doi.org/10.1080/03610927408827101.
52. Dunn JC. Well-separated clusters and optimal fuzzy partitions. J Cybern. 1974;4(1):95–104. https://doi.org/10.1080/01969727408546059.
53. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
54. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B Stat Methodol. 2001;63(2):411–23. https://doi.org/10.1111/1467-9868.00293.
55. Cohn SL, Pearson ADJ, London WB, Monclair T, Ambros PF, Brodeur GM, et al. The International Neuroblastoma Risk Group (INRG) classification system: an INRG task force report. J Clin Oncol. 2009;27(2):289. https://doi.org/10.1200%2FJCO.2008.16.6785.
56. Uva P, Bosco MC, Eva A, Conte M, Garaventa A, Amoroso L, et al. Connectivity map analysis indicates PI3K/Akt/mTOR inhibitors as Potential anti-hypoxia drugs in neuroblastoma. Cancers. 2021;13(11):2809. https://doi.org/10.3390/cancers13112809.
57. Morini M, Cangelosi D, Segalerba D, Marimpietri D, Raggi F, Castellano A, et al. Exosomal microRNAs from longitudinal liquid biopsies for the prediction of response to induction chemotherapy in high-risk neuroblastoma patients: a proof of concept SIOPEN study. Cancers. 2019;11(10):1476. https://doi.org/10.3390/cancers11101476.
58. Morini M, Raggi F, Bartolucci M, Petretto A, Ardito M, Rossi C, et al. Plasma-derived exosome proteins as novel diagnostic and prognostic biomarkers in neuroblastoma patients. Cells. 2023;12(21):2516. https://doi.org/10.3390/cells12212516.
59. Barco S, Lavarello C, Cangelosi D, Morini M, Eva A, Oneto L, et al. Untargeted LC-HRMS based-plasma metabolomics reveals 3-O-Methyldopa as a new biomarker of poor prognosis in high-risk neuroblastoma. Front Oncol. 2022;12. https://doi.org/10.3389/fonc.2022.845936.
60. Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl Eng. 2007;63(2):503–27. https://doi.org/10.1016/j.datak.2007.03.016.
61. Jahangiri L. Predicting neuroblastoma patient risk groups, outcomes, and treatment response using machine learning methods: a review. Med Sci. 2024;12(1):5. https://doi.org/10.3390/medsci12010005.
62. Preud'homme G, Duarte K, Dalleau K, Lacomblez C, Bresso E, Smaïl-Tabbone M, et al. Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Sci Rep. 2021;11(1). https://doi.org/10.1038/s41598-021-83340-8.
63. Liu X, Yang Q, He L. A novel DBSCAN with entropy and probability for mixed data. Clust Comput. 2017;20(2):1313–23. https://doi.org/10.1007/s10586-017-0818-3.
64. Cerono G, Melaiu O, Chicco D. Clinical feature ranking based on ensemble machine learning reveals top survival factors for glioblastoma multiforme. J Healthc Informat Res. 2024;8(1):1–18. https://doi.org/10.1007/s41666-023-00138-1.

Publisher's Note