DBSCAN clustering applied to electronic health records' data from patients with glioblastoma

Davide Chicco[1,2] and Luca Oneto[3]

[1] Università di Milano-Bicocca, Milan, Italy
[2] University of Toronto, Ontario, Canada
davidechicco@davidechicco.it
ORCID: 0000-0001-9655-7142
[3] Università di Genova, Genoa, Italy
luca.oneto@unige.it
ORCID: 0000-0002-8445-395X

Abstract. Glioblastoma is an aggressive brain cancer that kills approximately one hundred thousand people worldwide every year. Unfortunately, treatment and therapy for patients with this disease are complicated and have limited efficacy in improving individuals' chances of survival. Electronic health records (EHRs) contain patient information collected routinely at hospitals through medical visits and laboratory tests, providing an interesting source of data for computational analyses. Clustering is an area of unsupervised machine learning where an algorithm partitions data according to certain statistical properties or rules, thereby identifying hidden patterns and correlations that would otherwise be difficult to notice. In this study, we applied several clustering techniques to three open datasets of data derived from electronic health records, which included clinical, genetic, and administrative features of patients diagnosed with glioblastoma, considering two possible clusters. We evaluated our clustering results with the Density-Based Clustering Validation (DBCV) index, a relatively new score capable of accurately assessing both convex-shaped and concave-shaped clusters. Among the methods tested, Density-based Spatial Clustering of Applications with Noise (DBSCAN) yielded the best results across all three datasets. We then analyzed the features of the clusters identified by DBSCAN and found that cytosolic Hsp70 protein, sex, and brain subventricular zone resulted were significant distinguishing the two clusters across the three datasets.

Keywords: clustering · unsupervised machine learning · machine learning · glioblastoma · electronic health records · EHRs

1 Introduction

Glioblastoma is an aggressive type of brain cancer that originates from glial cells, which are supportive cells in the nervous system [1]. Glioblastomas are

To appear in the Proceedings of the 1st Workshop on Responsible Healthcare using Machine Learning (RHCML 2025) of the ECML PKDD 2025 conference

characterized by rapid growth, and they can infiltrate surrounding brain tissue, making them difficult to treat. Symptoms may include headaches, seizures, cognitive changes, and neurological deficits, depending on the tumor's location in the brain [1].

Data derived from electronic health records (EHRs) of patients having this cancer type can be used for computational analyses that, in turn, can lead to significant discoveries in medical sciences. In the past, we used supervised machine learning models on EHRs data of three open glioblastoma curated datasets to infer the most prognostic clinical factors [5].

In the present study, we reuse the same three open curated datasets for an unsupervised analysis, aimed at identifying clusters of patients which might have a particular medical relevance. In the past, several studies employed computational clustering techniques to analyze glioblastoma data, but mainly within the bioinformatics domain [2,16,6]. We could not find any study regarding unsupervised analyses of data from electronic medical records of patients with glioblastoma. We fill this gap by presenting this study aimed at detecting groups of patients having particular medical meaning.

2 Datasets

We analyzed three open datasets derived from electronic health records (EHRs) of patients diagnosed with glioblastoma, that we called Munich2019 dataset [8,9], Tainan2020 dataset [17], and Utrecht2019 [4]. A description of the clinical features of these datasets can be found in our previous study [5].

The Tainan2020 dataset contains data from 84 patients, each having 9 clinical features. The Utrecht2019 dataset consists of data from 647 patients and 7 variables. The Munich2019 dataset is made of data from 60 patients, each having 7 features as well.

3 Methods

We applied several unsupervised clustering methods to the three medical datasets of patients with glioblastoma, and DBSCAN [15] obtained the best results, measured as density-based clustering validation (DBCV index) [11] The DBCV index is a density-based version of the Silhouette coefficient [13], which is a common metric employed for clustering internal assessment. The original Silhouette coefficient can work well when assessing convex clusters, but can mislead when employed to evaluate concave or nested clusters. The DBCV index solves this problem by considering the density of the clusters in its formula, and generates a score in the $[-1; +1]$, where -1 means disastrous clustering, 0 means a clustering no better than random chance, and $+1$ means perfect clustering.

Differently from other clustering methods such as k-means or hierarchical clustering, DBSCAN assigns data points not only to real clusters but also to a noise cluster. This noise cluster consists of data points that do not belong to any real cluster, according to the DBSCAN partitioning.

We decided to set the number of clusters to two because it is a common number of clusters used in several glioblastoma biomedical informatics studies in the past [20,12,10]

4 Results

DBSCAN outperformed the other clustering algorithms we employed by obtaining a DBCV index of +0.963 in the Munich2019 dataset, of +0.923 in the Tainan2020 dataset, and of +0.961 in the Utrecht2019 dataset, respectively (Figure 1). On the three analyzed datasets, DBSCAN assigned to real clusters around 28% of patients in the Munich2019 dataset, around 31% of patients in the Tainan2020 dataset, and around 24% of patients in the Utrecht2019 dataset [11] (Figure 1).

dataset	**DBCV**	**# cl 1**	**# cl 2**	**# noise cl**	**# pts cl 1 & 2**	**%**
Munich2019	0.963	7	10	43	60	28.33
Tainan2020	0.923	13	13	58	84	30.95
Utrecht2019	0.961	74	80	493	647	23.80

Table 1: ***Results obtained by DBSCAN****. # cl 1: number of patients assigned by DBSCAN to the first cluster. # cl 2: number of patients assigned by DBSCAN to the second cluster. cl 1 & 2 %: percentage of patients assigned by DBSCAN to the first or the second cluster, and not assigned to the noise cluster. # pts.: number of patients. Hyperparameters: 2 clusters for DBSCAN in each test and the following epsilon values and minimal points Munich2019 dataset epsilon: 0.232, minimal samples: 4. Tainan2020 dataset epsilon: 0.286, minimal samples: 4. Utrecht2019 dataset epsilon: 0.429, minimal samples: 64.*

We then analyzed the content of the two clusters detected by DBSCAN in the three datasets, and observed the proportions of their clinical variables.

In the Munich2019 dataset, DBSCAN divided the patients by the cytosolic Hsp70 protein (major stress-inducible heat shock protein 70) level: patients with a high value of this factor were assigned to cluster 1, and patients with a low value to cluster 0 (Figure 1)a. DBSCAN divided the Tainan2020 dataset based on the sex variable: female patients were put in the cluster 0 and male patients in the cluster 1 (Figure 1)b. The patients of the Utrecht2019 dataset, instead, were partitioned based on the subventricular zone variable (Figure 1)c.

5 Discussion and conclusions

The current study has several assets. To the best of our knowledge, our project is the first clustering study using open medical data records of patients with glioblastoma and employing only open source software libraries. We found no

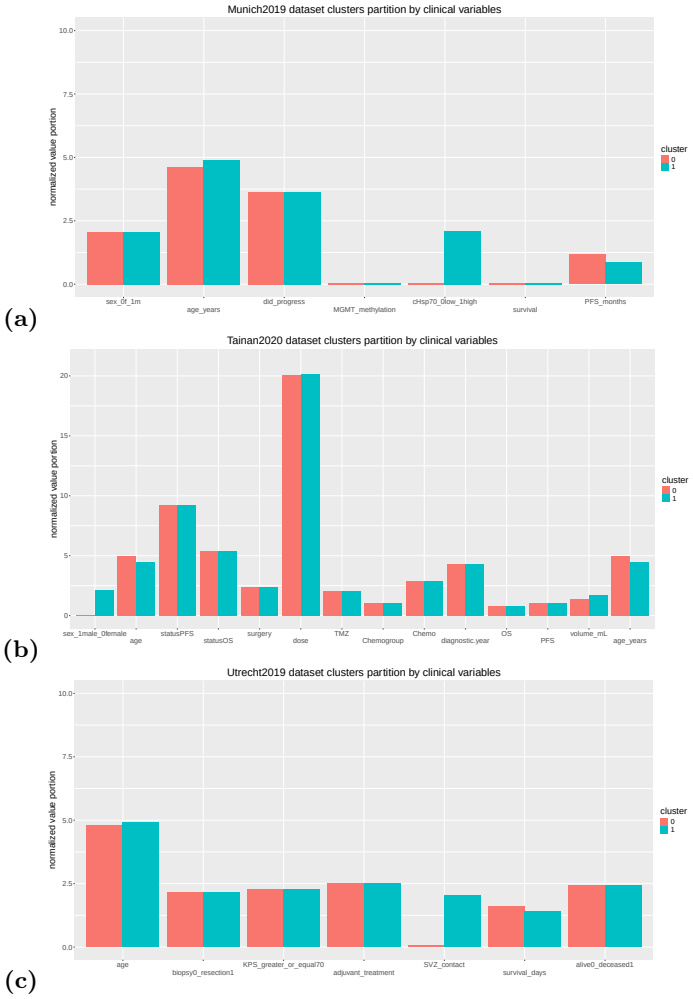

Fig. 1: ***Partition of the clinical features among the two clusters identified by DBSCAN****. Representation of the normalized values of the clinical variables of teach dataset in the subset of patients of the 0 cluster (red bars) and in the subset of patients of the 1 cluster (green bars). (a) Top image: Munich2019 dataset. (b) Mid image: Tainan2020 dataset. (c) Bottom image: Utrecht2019 dataset. We listed the meaning of the clinical variables in [5].*

other article describing the application of a fully-unsupervised approach data of this particular brain tumor.

Our results demonstrate that DBSCAN clustering, paired with the DBCV index, can identify groups of patients with significant medical traits among data derived from electronic health records. Moreover, our results indicate that some

clinical features can be more useful than others to partition the data of patients in a medically significant way: cystolic Hsp70 protein for the Munich2019 dataset, sex for the Tainan2020 dataset, and brain subventricular zone for the Utrecht2019 dataset. Each of these three clinical factors is known to have a significant role in glioblastoma [18,7,3]. These promising preliminary results appear to proof the capability of DBSCAN and DBCV to identify clusters of patients that have a medical meaning, paving the way to further analyses.

6 Generalizable insights about responsible application of machine learning in healthcare

This study represents an indicative case of responsible use of machine learning in healthcare: the algorithm and the metric employed (DBSCAN and DBCV index), in fact, can be clearly interpreted and explained to anyone, even without a deep knowledge on machine learning. Our approach can be be considered fair because our clustering methods do not produce biased outcomes based on sensitive attributes such as race, sex, gender, or socioeconomic status. Moreover, the privacy of patients is preserved by the anonymity of data: nobody can trace the identity of patients from the data, even in the remote case they wanted to. The datasets were collected by the original data curators after obtaining the informed consents from the patients and the authorization of the ethical committees of the corresponding hospitals [8,9,17,4]. The anonymous datasets were then released openly following the FAIR principles [19]. Regarding explainability and interpretability, we decided to use a modern clustering algorithm (DBSCAN) whose functioning is known and can be explained to anyone. DBSCAN, in fact, is not a black-box model [14]. Our project is fully patient-centric: our primary goal was on improving patient outcomes, and we did it by identifying the main clinical features that can discriminate the clusters of patients in the three analyzed datasets. Of course, these practices can be generalized to any biomedical informatics research project.

Additional information

List of abbreviations DBCV: Density-Based Clustering Validation. DBSCAN: density-based spatial clustering of applications with noise. FAIR: findability, accessibility, interoperability, and reusability. KPS: Karnofsky Performance Scale. MGMT: Methylated-DNA–protein-cysteine methyltransferase. mL: milliliters. OS: overall survival. PFS: progression-free survival. SVZ: sub ventricular zone.

Ethics approval and consent to participate Permission to collect and analyze the data of the patients involved in this study was obtained from the ethical committees by the original data curators, as stated in the original articles [8,9,17,4].

Conflict of interests The authors declare they have no conflict of interests.

Funding 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) (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.

Acknowledgments The authors acknowledge the usage of Ecosia AI Chat for English proof-reading and grammar correction of the article's text.

Data availability The datasets used in this study are publicly available under the CC BY 4.0 license at the following URLs.

– Munich2019 dataset [8,9]: https://figshare.com/articles/dataset/Clinical_data_of_individual_patients_/14201600?file=26782502
– Tainan2020 dataset [17]: https://figshare.com/articles/dataset/S1_Data_-/12312737?file=22696553
– Utrecht2019 dataset [4]: https://figshare.com/articles/dataset/Adverse_prognosis_of_glioblastoma_contacting_the_subventricular_zone_Biological_correlates/9972809?file=17979143

References

1. Anjum, K., Shagufta, B.I., Abbas, S.Q., Patel, S., Khan, I., Shah, S.A.A., Akhter, N., ul Hassan, S.S.: Current status and future therapeutic perspectives of glioblastoma multiforme (GBM) therapy: a review. Biomedicine & Pharmacotherapy **92**, 681–689 (2017), DOI URL: https://doi.org/10.1016/j.biopha.2017.05.125
2. Baheti, B., Innani, S., Nasrallah, M., Bakas, S.: Prognostic stratification of glioblastoma patients by unsupervised clustering of morphology patterns on whole slide images furthering our disease understanding. Frontiers in Neuroscience **18** (2024), DOI URL: http://doi.org/10.3389/fnins.2024.1304191
3. Beiriger, J., Habib, A., Jovanovich, N., Kodavali, C.V., Edwards, L., Amankulor, N., Zinn, P.O.: The subventricular zone in glioblastoma: genesis, maintenance, and modeling. Frontiers in Oncology **12**, 790976 (2022), DOI URL: https://doi.org/10.3389/fonc.2022.790976
4. Berendsen, S., van Bodegraven, E., Seute, T., Spliet, W.G., Geurts, M., Hendrikse, J., Schoysman, L., Huiszoon, W.B., Varkila, M., Rouss, S., Bell, E.H., Kroonen, J., Chakravarti, A., Bours, V., Snijders, T.J., Robe, P.A.: Adverse prognosis of glioblastoma contacting the subventricular zone: biological correlates. PLOS One **14**(10), e0222717 (2019), DOI URL: https://doi.org/10.1371/journal.pone.0222717
5. Cerono, G., Melaiu, O., Chicco, D.: Clinical feature ranking based on ensemble machine learning reveals top survival factors for glioblastoma multiforme. Journal of Healthcare Informatics Research **8**(1), 1–18 (Sep 2023), DOI URL: http://doi.org/10.1007/s41666-023-00138-1
6. García-Gómez, J.M., Gómez-Sanchis, J., Escandell-Montero, P., Fuster-Garcia, E., Soria-Olivas, E.: Sparse manifold clustering and embedding to discriminate gene expression profiles of glioblastoma and meningioma tumors. Computers in Biology and Medicine **43**(11), 1863–1869 (2013), DOI URL: https://doi.org/10.1016/j.compbiomed.2013.08.025

7. Gittleman, H., Ostrom, Q.T., Stetson, L.C., Waite, K., Hodges, T.R., Wright, C.H., Wright, J., Rubin, J.B., Berens, M.E., Lathia, J., Connor, J.R., Kruchko, C., Sloan, A.E., Barnholtz-Sloan, J.S.: Sex is an important prognostic factor for glioblastoma but not for nonglioblastoma. Neuro-Oncology Practice **6**(6), 451–462 (2019), DOI URL: https://doi.org/10.1093/nop/npz019
8. Lämmer, F., Delbridge, C., Würstle, S., Neff, F., Meyer, B., Schlegel, J., Kessel, K.A., Schmid, T.E., Schilling, D., Combs, S.E.: Cytosolic Hsp70 as a biomarker to predict clinical outcome in patients with glioblastoma. PLOS One **14**(8), e0221502 (2019), DOI URL: https://doi.org/10.1371/journal.pone.0221502
9. Lämmer, F., Delbridge, C., Würstle, S., Neff, F., Meyer, B., Schlegel, J., Kessel, K.A., Schmid, T.E., Schilling, D., Combs, S.E.: Correction: Cytosolic Hsp70 as a biomarker to predict clinical outcome in patients with glioblastoma. PLOS One **16**(3), e0248612 (2021), DOI URL: https://doi.org/10.1371/journal.pone.0248612
10. Maher, E.A., Brennan, C., Wen, P.Y., Durso, L., Ligon, K.L., Richardson, A., Khatry, D., Feng, B., Sinha, R., Louis, D.N., Quackenbush, J., Black, P.M., Chin, L., DePinho, R.A.: Marked genomic differences characterize primary and secondary glioblastoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. Cancer Research **66**(23), 11502–11513 (2006), DOI URL: https://doi.org/10.1158/0008-5472.can-06-2072
11. Moulavi, D., Jaskowiak, P.A., Campello, R.J., Zimek, A., Sander, J.: Density-based clustering validation. In: Proceedings of SDM24 – the 2014 SIAM International Conference on Data Mining. pp. 839–847. SIAM (2014), DOI URL: https://doi.org/10.1137/1.9781611973440.96
12. Rayfield, C.A., Grady, F., De Leon, G., Rockne, R., Carrasco, E., Jackson, P., Vora, M., Johnston, S.K., Hawkins-Daarud, A., Clark-Swanson, K.R., Whitmire, S., Gamez, M.E., Porter, A., Hu, L., Gonzalez-Cuyar, L., Bendok, B., Vora, S., Swanson, K.R.: Distinct phenotypic clusters of glioblastoma growth and response kinetics predict survival. JCO Clinical Cancer Informatics **2**, 1–14 (2018), DOI URL: https://doi.org/10.1200/cci.17.00080
13. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (Nov 1987), DOI URL: http://doi.org/10.1016/0377-0427(87)90125-7
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019), DOI URL: https://doi.org/10.1038/s42256-019-0048-x
15. Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X.: DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems **42**(3), 1–21 (2017), DOI URL: https://doi.org/10.1145/3068335
16. Shen, R., Mo, Q., Schultz, N., Seshan, V.E., Olshen, A.B., Huse, J., Ladanyi, M., Sander, C.: Integrative subtype discovery in glioblastoma using iCluster. PLOS One **7**(4), e35236 (2012), DOI URL: https://doi.org/10.1371/journal.pone.0035236
17. Shieh, L.T., Guo, H.R., Ho, C.H., Lin, L.C., Chang, C.H., Ho, S.Y.: Survival of glioblastoma treated with a moderately escalated radiation dose—Results of a retrospective analysis. PLOS One **15**(5), e0233188 (2020), DOI URL: https://doi.org/10.1371/journal.pone.0233188
18. Thorsteinsdottir, J., Stangl, S., Fu, P., Guo, K., Albrecht, V., Eigenbrod, S., Erl, J., Gehrmann, M., Tonn, J.C., Multhoff, G., Schichor, C.: Overexpression of cytosolic, plasma membrane bound and extracellular heat shock protein 70

(Hsp70) in primary glioblastomas. Journal of Neuro-Oncology **135**, 443–452 (2017), DOI URL: https://doi.org/10.1007/s11060-017-2600-z

19. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3**(1) (Mar 2016), DOI URL: http://doi.org/10.1038/sdata.2016.18
20. Zhang, G., Xu, X., Zhu, L., Li, S., Chen, R., Lv, N., Li, Z., Wang, J., Li, Q., Zhou, W., Yang, P., Liu, J.: A novel molecular classification method for glioblastoma based on tumor cell differentiation trajectories. Stem Cells International **2023**(1), 2826815 (2023), DOI URL: https://doi.org/10.1155/2023/2826815