# Investigating Alzheimer's Disease Using a Sequential Analytical Pipeline for High-Dimensional, Low-Sample Biomedical Data

Nadezhda Shchegoleva[1(✉)], Petr Tonka[1], and Natalia Zalutskaya[2]

[1] Saint Petersburg State University, St. Petersburg, Russia
n.shchegoleva@spbu.ru, st102620@student.spbu.ru
[2] Federal State Budgetary Institution «Bekhterev National Medical Research Psychiatry and Neurology Center», Saint-Petersburg, Russia

**Abstract.** Dementia, particularly Alzheimer's disease, is a growing global medical and economic problem exacerbated by increasing life expectancy and aging population. This study demonstrates the critical need for efficient data processing methods in biomedical research, focusing on scenarios where the number of observations is smaller than the dimensionality of the data. Based on a dataset of patients with various neurodegenerative diseases including Alzheimer's disease, this paper describes a comprehensive data processing pipeline including dimensionality reduction and clustering. The methodology is based on the combined use of principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) for dimensionality reduction, followed by clustering with optimized parameters. The study demonstrates how these methods can mitigate problems such as data sparsity and the curse of dimensionality, providing a pipeline to gain insights into potential biomarkers for early diagnosis of neurodegenerative diseases. Although this study does not directly interpret clustering results, it lays the groundwork for experts in the field to explore new hypotheses and improve diagnostic tools.

**Keywords:** Alzheimer's disease · dimensionality reduction · clustering · PCA · UMAP · HDBSCAN

## 1 Introduction

Dementia, particularly Alzheimer's disease (AD), is a critical global health issue, rapidly increasing in prevalence due to rising life expectancy and aging of the population [1, 2]. It represents a significant medical and economic burden worldwide [3, 4], posing substantial challenges to healthcare systems. While the scale of the problem is substantial, with millions affected globally and projections indicating continued growth [2], the urgent need for effective diagnosis and treatment methods is widely acknowledged.

Current diagnostic approaches involve various biomarkers such as cerebrospinal fluid (CSF) analysis of amyloid beta (Aβ), total tau (t-tau), and phosphorylated tau

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15894, pp. 340–355, 2026.
https://doi.org/10.1007/978-3-031-97648-3_23

(p-tau) accumulation, and positron emission tomography (PET) scans (amyloid PET and tau PET) [5]. However, these methods often have drawbacks, including high cost, complexity, and invasiveness [5–8]. In this regard, alternative diagnostic methods are being developed, for example, the development of biomarkers based on the patients' blood analysis, mainly also relying on the analysis of Aβ and tau pathologies, looks quite promising [5–9]. However, the amyloid cascade theory itself, built around the toxic action of amyloid-β, is still a subject of controversy and debate, so there remains a need to search for and investigate other factors of disease onset in parallel with the amyloid hypothesis [5].

Despite growing interest and advances in diagnostic strategies, research into Alzheimer's disease continues to face considerable methodological difficulties, particularly in relation to data acquisition and analysis. Biomedical studies in this domain typically generate large volumes of complex and heterogeneous data derived from a variety of sources, including clinical records, patient histories, neuropsychological testing, and instrumental methods such as electroencephalography (EEG) and magnetic resonance imaging (MRI). These multimodal data sources differ significantly in structure and often contain missing values, which adds layers of complexity to data integration and analysis workflows [14].

Moreover, the high dimensionality of such data – especially in the case of neuroimaging, where thousands of features can be extracted from a single scan – presents further analytical challenges [15]. These are compounded by the limited number of observations that are characteristic of clinical research, resulting in scenarios where the number of variables dramatically exceeds the number of observations. Such conditions hinder the applicability of conventional statistical and machine learning methods, necessitating the development and implementation of specialized approaches for robust and interpretable data analysis [10–12].

This paper proposes a methodological approach specifically tailored to analyze this type of complex medical data characterized by limited dataset size and high dimensionality. Based on a dataset of patients with various neurodegenerative diseases, including AD, our study details a comprehensive data processing pipeline. This pipeline includes preprocessing, a two-stage dimensionality reduction using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), and subsequent clustering with optimized parameters. This methodology is designed to effectively manage the curse of dimensionality while preserving meaningful underlying structures in the data. A main outcome of this study is to lay the technical foundation for a diagnostician's workplace by building the basis of a decision support system, enabling domain experts to explore new hypotheses about potential disease biomarkers for early diagnosis of neurodegenerative diseases.

## 2 Methods

### 2.1 Data Description

The dataset used in this study was obtained from the St. Petersburg Bekhterev Psychoneurological Research Institute. This institution-specific dataset contains detailed clinical information on patients diagnosed with various neuropsychiatric disorders. While the

original dataset is not publicly available due to the inclusion of non-anonymized patient information, the analysis for this study was performed on data accessed and processed within a secure, distributed private cloud environment designed for handling sensitive health data, including anonymized forms of the data. The dataset comprises demographic and background data (such as gender, age, education level, and work history), medical histories (anamneses), results from patient assessments, instrumental diagnostic findings, and clinical diagnoses provided by expert practitioners. The dataset encompasses six diagnostic categories: dementia, vascular dementia, depression, Alzheimer's disease, Parkinson's disease, and a control group.

The available assessments include laboratory blood tests, experimental psychological evaluations, neuropsychological testing, as well as neuroimaging data from MRI and EEG studies. Notably, MRI results were processed using voxel-based morphometry [13] and formatted into structured tabular data, consistent with the rest of the dataset. The dataset includes a total of 254 patients. The distribution of patients across diagnostic categories is presented in Table 1.

Next, the stages of processing and analyzing the dataset under consideration will be described. The complete process cycle is illustrated in Fig. 1.

**Table 1.** The distribution of patients by diagnoses

| Clinical Diagnosis | No. of Observations |
|---|---|
| Alzheimer's disease | 60 |
| Vascular Dementia | 58 |
| Control Group | 54 |
| Depression | 45 |
| Dementia | 36 |
| Parkinson's disease | 1 |

## 2.2 Data Preparation

Considering all available patient information, the sample contains 714 features—almost three times the number of records. For analysis, we focused on EEG and MRI data. The EEG results are divided into two parts: the first includes averaged electrode values, relative indices, and measurement correlations, while the second provides aggregated values per electrode. Initially, the subsample was split into three independent datasets: MRI (MRI), general EEG (EEG), and aggregated EEG (EEG_sumup).

Data quality checks were performed on each subsample. No partial missing values were detected. Incomplete rows (reflecting unperformed procedures) were discarded. The data were also examined for duplicate rows and negative values, with none identified.

Following data cleaning, two additional subsets were created by merging the original three: a mixed EEG subset (EEG + EEG_sumup) and a combined EEG and MRI subset

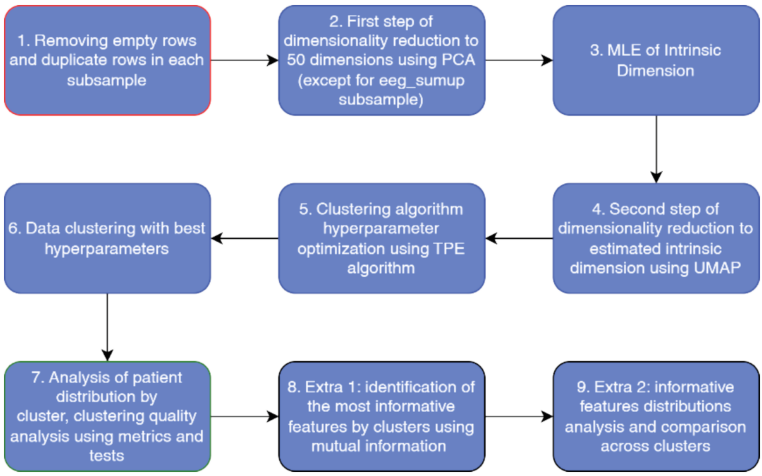


**Fig. 1.** Data preparation and analysis steps

(EEG + EEG_sumup + MRI). Only patients present in both source subsamples were retained (inner join). Table 2 summarizes the sizes of all five resulting subsamples.

**Table 2.** Final subsample sizes

| Subsample Name | No. of Samples | No. of Features |
|---|---|---|
| MRI | 117 | 315 |
| EEG | 176 | 192 |
| EEG_sumup | 177 | 15 |
| EEG + EEG_sumup | 172 | 207 |
| EEG + EEG_sumup + MRI | 94 | 522 |

**Table 3.** Distribution of patients by cluster using HDBSCAN

| Cluster Label | No. of cases | No. of Clinician Diagnoses in clusters | Clinician Diagnoses Across Clusters (%) | Clinician Diagnoses Within Clusters (%) |
|---|---|---|---|---|
| −1 | 31 | 6/6/6/7/6/0 | 14/17.1/23.1/19.4/17.1/0.0 | 19.4/19.4/19.4/22.6/19.4/0.0 |
| 0 | 73 | 15/13/13/18/13/1 | 34.9/37.1/50.0/50.0/37.1/100.0 | 20.5/17.8/17.8/24.7/17.8/1.4 |
| 1 | 4 | 1/1/0/1/1/0 | 2.3/2.9/0.0/2.8/2.9/0.0 | 25.0/25.0/0.0/25.0/25.0/0.0 |
| 2 | 25 | 8/2/3/5/7/0 | 18.6/5.7/11.5/13.9/20.0/0.0 | 32.0/8.0/12.0/20.0/28.0/0.0 |
| 3 | 7 | 3/2/0/1/1/0 | 7/5.7/0.0/2.8/2.9/0.0 | 42.9/28.6/0.0/14.3/14.3/0.0 |
| 4 | 9 | 3/3/2/0/1/0 | 7/8.6/7.7/0.0/2.9/0.0 | 33.3/33.4/22.2/0.0/11.1/0.0 |
| 5 | 27 | 7/8/2/4/6/0 | 16.3/22.9/7.7/11.1/17.1/0.0 | 25.9/29.6/7.4/14.8/22.2/0.0 |

**Table 4.** Distribution of patients by cluster using Agglomerative Clustering

| Cluster Label | No. of cases | No. of Clinician Diagnoses in clusters | Clinician Diagnoses Across Clusters (%) | Clinician Diagnoses Within Clusters (%) |
|---|---|---|---|---|
| 0 | 91 | 25/19/8/18/21/0 | 58.1/54.3/30.8/50.0/60.0/0.0 | 27.5/20.9/8.9/19.8/23.1/0.0 |
| 1 | 85 | 18/16/18/18/14/1 | 41.9/45.7/69.2/50.0/40.0/100.0 | 21.2/18.8/21.2/21.2/16.5/1.2 |

**Table 5.** Percentage labeled, metrics and cluster stability scores (EEG)

| Algorithm | Labeled (%) | DBCV | Silhouette | ARI | AMI |
|---|---|---|---|---|---|
| HDBSCAN | 82.39 | 0.47 | 0.32 | 0.56 | 0.68 |
| Agglomerative | 100.00 | 0.21 | 0.45 | 0.71 | 0.68 |
| Spectral | 100.00 | 0.17 | 0.40 | 0.74 | 0.79 |

### 2.3 Dimensionality Reduction with PCA

After the first step - selection of three separate subsamples and cleaning them from irrelevant strings - each of them was further subjected to the same processing and analysis steps, except for the EEG_sumup subsample, to which the second step was not applied due to its small dimensionality. The second step of data processing is primary dimensionality reduction using principal component analysis (PCA) with preliminary data normalization. Data normalization ensures that all features contribute equally, which is critical for distance-based methods. PCA is used to reduce the dimensionality of the original data to a set of principal components that preserve the underlying variance. This reduces the impact of the curse of dimensionality in the data. In addition, PCA combats noise and redundancy in the data, and forms a space of uncorrelated features [16]. This step subsequently not only speeds up the algorithms used in the subsequent steps, but also helps to stabilize their results. In this step, the number of principal components to be selected was set to a fixed value of 50 components for both subsamples. In the case of EEG, this number of principal components preserved 97% of the variance in the data, while for MRI - 90% of the variance.

### 2.4 Calculation of Internal Dimensionality of Data

To objectively determine the optimal number of dimensions for a low-dimensional representation that preserves the essential structure of the original space, we computed the intrinsic dimensionality of subsamples using the MLE estimator (Maximum Likelihood Estimator of Intrinsic Dimension) [17].

It is believed that in most cases any methods applied in very high dimensions work only because the data are not really truly high dimensional, but rather they are embedded in a high dimensional space, but can be effectively generalized to a space of much lower dimensionality. Evaluating the intrinsic dimensionality reveals the minimum number of dimensions sufficient to describe the underlying structure of the data [16, 17]. This approach helps to select the number of components that simultaneously minimizes noise and

redundancy, providing a more accurate representation of the real relationships between objects. This, in turn, improves subsequent analysis, for example, in clustering or building predictive models.

The choice of this estimator is justified by its simplicity and sufficiency within the framework of this work. It should be noted that the main problem of the estimator is the underestimation of the real dimensionality of the data when this dimensionality is actually large. However, as the experimental results showed, the estimated intrinsic dimensionality for each of the subsamples did not exceed 10, which is relatively small, especially when compared to the original dimensionality of the data.

Another problem can be the bias of the estimator results from the real internal dimensionality in case of calculating too large number of nearest neighbors for each of the points. This number is regulated by two hyperparameters – $k1$ and $k2$ – that set a range of values for the number of nearest neighbors computed during the algorithm. This problem is partially solved by simply averaging the estimators over a range of values $k = k1 \ldots k2$. The problem is also mitigated by choosing optimal values of $k1$ and 2, which can be chosen empirically based on the sample size and scale of values obtained by the estimator over an arbitrary range of values $k$.

### 2.5 Dimensionality Reduction with UMAP

After applying PCA and estimating the intrinsic dimensionality of the data, the use of UMAP [18] follows to further reduce the dimensionality. UMAP is often used in biomedical research and has proven itself in biomedical research as a versatile tool for creating low-dimensional representations [19–22], also due to the fact that it is nonparametric and, as a consequence, it does not require a large amount of data for its application.

By minimizing the noise and the number of irrelevant features using PCA, UMAP can model the internal structure of the data more effectively. By further reducing the dimensionality after applying PCA (50 features) and after applying UMAP equal to the value of internal dimensionality obtained in the previous step, algorithm better preserves the local and global relationships of the original data in the lower dimensionality space, which makes the subsequent data analysis results more reliable.

UMAP, as a nonlinear dimensionality reduction method, has a number of strengths. Unlike linear methods such as PCA, LDA (Linear Discriminant Analysis) or FA (Factor Analysis), it is able to account for complex, nonlinear relationships between objects, which is particularly relevant for biomedical data where the true structure may be hidden as a low-dimensional manifold. The nonlinearity of UMAP allows the local topology of the original space to be preserved, ensuring accurate recovery of local relationships between objects. This is critical for subsequent analysis, as it is in such a compact and informative representation that biologically relevant groups are easier to identify.

Within the scope of this work, the parameter for the number of nearest neighbors of the algorithm was chosen to be relatively large (30 neighbors) to avoid too much emphasis on local structures. The minimum distance between points in the low-dimensional space was chosen to be 0 in order to increase the density of clusters.

### 2.6 Clustering Algorithms

Given that the goal of the analysis, as stated above, is to generate new hypotheses about biomarkers of Alzheimer's disease, this analysis process is organized, relying as much as possible on the data themselves, rather than on human partitioning, to use clustering seems a logical choice. Clustering was chosen as the primary analysis tool for the reason that clustering, as a method of learning without a teacher, does not require proper data partitioning and is capable of searching the data for structure on its own. However, the automatic partitioning of the data does not mean that there is no need of validation of the results obtained by a specialist. On the contrary, the presence of clinical diagnoses and comparison of clustering results with the available partitioning can also be valuable information, especially if the obtained clusters overlap with the opinion of doctors. However, the absence of such overlaps can also be informative, but in such a case a more thorough analysis of the obtained clusters is needed to validate the value of the obtained partitions in the data.

Three clustering algorithms considered in this work are: HDBSCAN [23, 24], agglomerative clustering and spectral clustering [25]. The choice of these three methods is justified by the balance between the complementary principles underlying each algorithm, providing a more complete view of different aspects of the internal structure of the data. HDBSCAN and agglomerative clustering, among others, are able to determine the number of clusters in the data. Moreover, the use of multiple algorithms provides additional information about the stability of the clusters in the case where there are strong overlaps between the clustering results of the algorithms, i.e., stable partitions are formed. At the same time, a larger number of methods generates more hypotheses and adds reliability to the process of data analysis, since we do not have to rely on the correct operation of a single algorithm.

### 2.7 Hyperparameter Optimization

To evaluate the result of clustering, a more objective way than simply checking whether clusters correspond to clinical diagnoses is to select metrics that would reflect the quality of clustering and could be used to evaluate the results of the algorithms used. These same metrics could then be used to optimize the hyperparameters of the models fit by the algorithms. Two metrics were chosen as part of this work, DBCV [26] and Silhouette Score [27]. These metrics are the minimum set needed to get an overview of the internal structure of the data and to evaluate the clustering algorithms under consideration. DBCV considers both density and structural features of clusters, assuming their arbitrary shape (e.g., non-convex), and is also able to deal with noise points, which is important for HDBSCAN. Silhouette Score in turn assumes convex compact clusters. Using the two metrics together gives a general idea of the convexity/non-convexity of the clusters and the consistency of the algorithms' assumptions about the structure of the data. Both metrics, moreover, have the same range of values $[-1, 1]$ and in both cases higher values of the metric mean better clustering.

Based on these metrics, we can optimize the hyperparameters of each clustering method by maximizing each of the metrics. The Tree-Structured Parzen Estimator (TPE)

method [28] is well suited for this purpose. TPE is a versatile and widely used hyperparameter optimization method that outperforms simpler Random Search and Grid Search due to the probabilistic model of the target function after each iteration [29, 30]. TPE iteratively models the probability distribution of the best performing hyperparameters based on previous trials. It identifies priority regions of the hyperparameter space that are more likely to yield good (in terms of the metrics being optimized) clusters, which significantly reduces the number of evaluations required for convergence to optimal configurations. In convergence, the results of TPE are better than those of Random Search and Grid Search.

Thus, the use of TPE within the study allowed us to select the optimal set of hyperparameters for each clustering model, maximizing the values of the selected metrics. At the same time, the target function was further modified to impose constraints on the resulting clustering models during optimization. First, separately for HDBSCAN, a constraint on the fraction of noise points was imposed - no more than 35%. Second, for all algorithms, a test for statistical significance of the resulting clustering was added ($p < 0.005$).

For the spectral and agglomerative clustering methods, a label reassignment test was conducted using Silhouette Score as the clustering quality metric. Here, the null hypothesis $H0$ states that there is no internal clustering structure, and any observed score of the metric is not different from the scores obtained by randomly reassigning the labels of the clusters (while preserving the noise assignments). By repeatedly reassigning labels and re-calculating the metric, a null distribution is formed. The P-value is determined by the fraction of permutation estimates that are the same or higher, than the observed estimate. A significant result supports the alternative hypothesis that the clustering structure is non-random and statistically significant.

For HDBSCAN, a data-level statistical significance test was used with DBCV as a validation metric. The null hypothesis $H0$ states that the clustering structure revealed by HDBSCAN is no better than the structure derived from the uniformly distributed data. In other words, the observed clusters are neither denser nor more separable than the clusters found in the random data. To test this assertion, null data sets were generated from a uniform sample within the observed data boundaries, and each was clustered using HDBSCAN to calculate the corresponding DBCV value. A right-sided test of the observed DBCV value from the original clustering was then performed using the generated empirical distribution. A low p-value would indicate that the observed clustering pattern was statistically significant, thereby rejecting H0 in favor of the alternative hypothesis that the observed clustering quality exceeds what would be expected by chance.

It is worth noting that the optimization was performed using both metrics simultaneously, in other words, the optimization was multi-objective. Since in such a case there can be many optimal combinations of hyperparameters, each set of which corresponds to one of the compromise solutions (Pareto front), for simplicity in such a case two sets of hyperparameters were selected, each of which would achieve the maximum value of one of the metrics. At the same time, the value of the second metric was not taken into account in this approach.

### 2.8 Clustering and Testing the Stability of Clusters

Determining the optimal parameters that maximize the selected metrics allows us to perform clustering using the methods described above more efficiently. But before giving an example of clustering algorithms' performance after all the preparatory works, it is also worth describing the last step of validation of clustering results, applied to the results of the model with the best hyperparameters. The last test is the validation of clusters for stability.

The clustering robustness test evaluates the robustness and reproducibility of the clustering solution by generating bootstrap samples. For each of several bootstrap iterations, a random subsample of data is clustered using the same algorithm, and the cluster labels are aggregated for each data point present in that iteration. Each pair of clustered subsamples is then compared using Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) [31, 32].

ARI and AMI are widely used normalized metrics for comparing clustering results that are invariant to label permutation. Both metrics range from 0 (no agreement) to 1 (complete agreement) and make allowance for randomness, which is very important when assessing stability. High ARI or AMI values indicate that cluster distributions are replicated across subsamples, indicating that the underlying cluster structure is reliable. Stability is important for research because it is evidence that the clusters found are not artifacts of random sampling or parameter selection, but reflect stable patterns in the data.

## 3 Example of Analysis

Let us consider the clustering results and some summary tables describing the distribution of patients within clusters based on clinical diagnoses. We will not consider all the experiments, since describing the results of each of them would be too cumbersome. We will give two examples. The first is on a subsample of EEG, and the second is with a subsample of EEG + EEG_sumup + MRI.

### 3.1 Example 1. Clustering of a Subsample of EEG

PCA is used to first reduce the sample dimensionality to 50 principal components, preserving in this case 97% of the variance in the data. Next, the MLE-estimator is used to calculate the internal dimensionality, which determines that it is equal to 6. The hyperparameters of the clustering models are then optimized and the clustering with the best set of hyperparameters is applied on the data. The clustering is done in 6-dimensional space. After that, to get a visual representation of the result, two graphs of 3D projection of the data are plotted with point markup according to the clustering results. The projection in both cases is generated using UMAP. The first plot shows the mapping of the six-dimensional space into a three-dimensional space, i.e. UMAP, when building the low-level vector representation (embedding), optimizes the already optimized data structure obtained after the first application of UMAP. This sequential mapping of the data may lead to some artifacts in the 3D case, but better demonstrates the structure of the

space in which the clustering was performed. The second graph will be a direct mapping of the 50-dimensional space obtained immediately after the application of PCA into the three-dimensional space, labeling the points with the same cluster labels. This graph aims more at showing the clustering results in a space closer to the initial data structure. Considering that PCA was used to remove noise in the data and remove low-information features while preserving most of the variance, displaying this particular representation of the data seems logical. In addition to the visual representation, cluster stability analysis (which was not performed during the optimization phase) is performed, key metrics are calculated, and two summary tables containing the distribution of patients into clusters based on clinical diagnoses are displayed. The first table shows the proportion of patients of each diagnosis in each individual cluster, while the second table shows the distribution of patients of each diagnosis across clusters. An example visualization of the clustering of the selected subsample, as well as the metrics and (merged) supporting tables, are presented in Figs. 2–3 and Tables 3, 4 and 5. The order of diagnoses in the column of patient distribution across clusters is the same for all of the following tables: Alzheimer's disease, dementia, depression, vascular dementia, control group and Parkinson's disease, if presented.

In addition to the clustering visualization and summary table, the most informative features for each cluster were identified, and the distributions of data values for the features within each individual cluster were visualized to see how the values of the features forming differ from cluster to cluster. To assess the importance of the traits for each cluster, Mutual Information (MI) was calculated between each feature and the binary indicator of membership in each individual cluster. Mutual Information is well suited for this task, as it captures both linear and non-linear dependencies without assuming any particular distribution of the data [33]. To account for the differences in uncertainty inherent in clusters, MI estimates were normalized by the entropy of the cluster's binary label. This normalization brings MI to a range [0, 1], which facilitates interpretation by showing the fraction of the maximum possible information contained in each feature [34]. A graph with the most informative features, defined in the above-described way, for each cluster for the two algorithms is shown in Fig. 4.

### 3.2 Example 2. Clustering of Intersection of All Results

Lastly, we examine the results of analyzing the subsample with the smallest number of patients and the largest initial feature space. To avoid cluttering of graphs and tables, let us consider only the most visually interesting clustering and distribution of patients into clusters, as well as the obtained metrics for all three algorithms. We will omit the list of the most informative features of each cluster and compare it with clustering of other algorithms.

The visualization shows the clustering results using the spectral clustering algorithm. As can be seen in Fig. 5, the algorithm was able to identify three clusters in the data. And on the left plot of the Fig. 5 it is noticeable that each of the clusters is mostly located in one of the corners of the figure formed by all points. The zero cluster, judging by the same visualization, is better separated from the other two clusters. If we look on the right plot of the same Figure, we can conclude that all clusters are convex or almost convex. With the exception of cluster 1, which is slightly bulging toward cluster zero, no

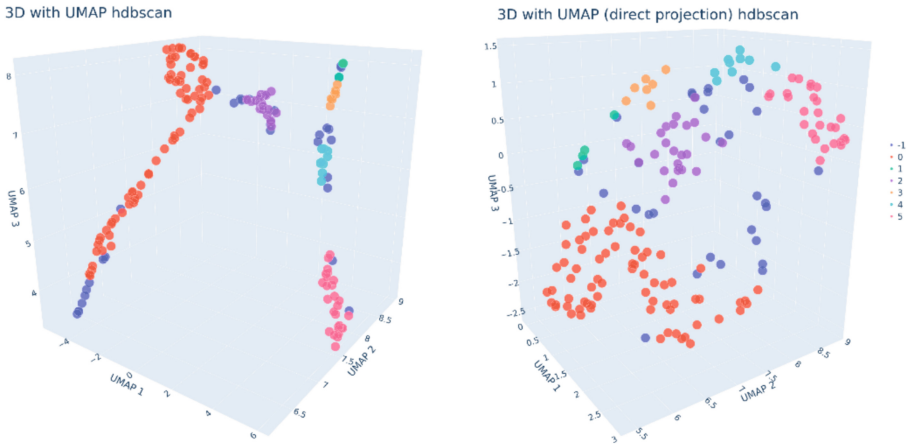


**Fig. 2.** Visualization of clustering results using HDBSCAN

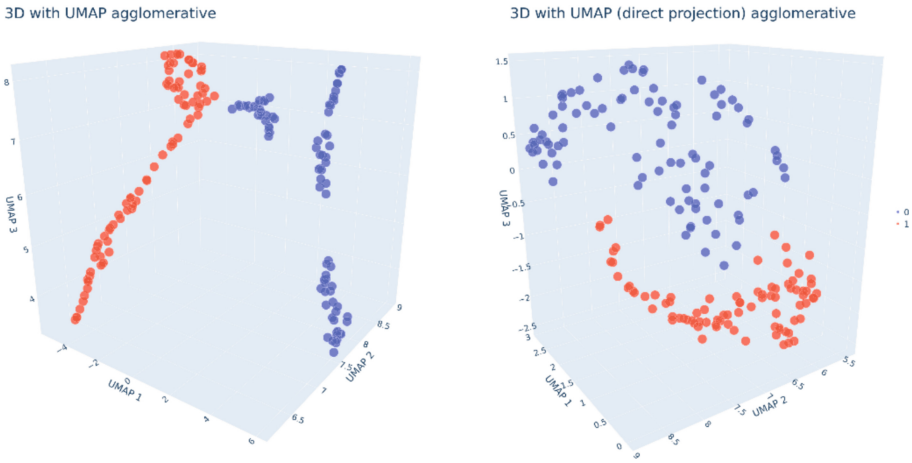


**Fig. 3.** Visualization of clustering results using Agglomerative Clustering

other "intersections" are visually apparent. The assumption that the clusters are convex is confirmed by the relatively high value of the Silhouette Score metric. It is noteworthy that in the resulting clustering we managed to identify a cluster that included 60% of patients with clinical diagnosis of Alzheimer's disease. The sizes of the clusters were almost identical (Figs. 6, 7 and Tables 6, 7).

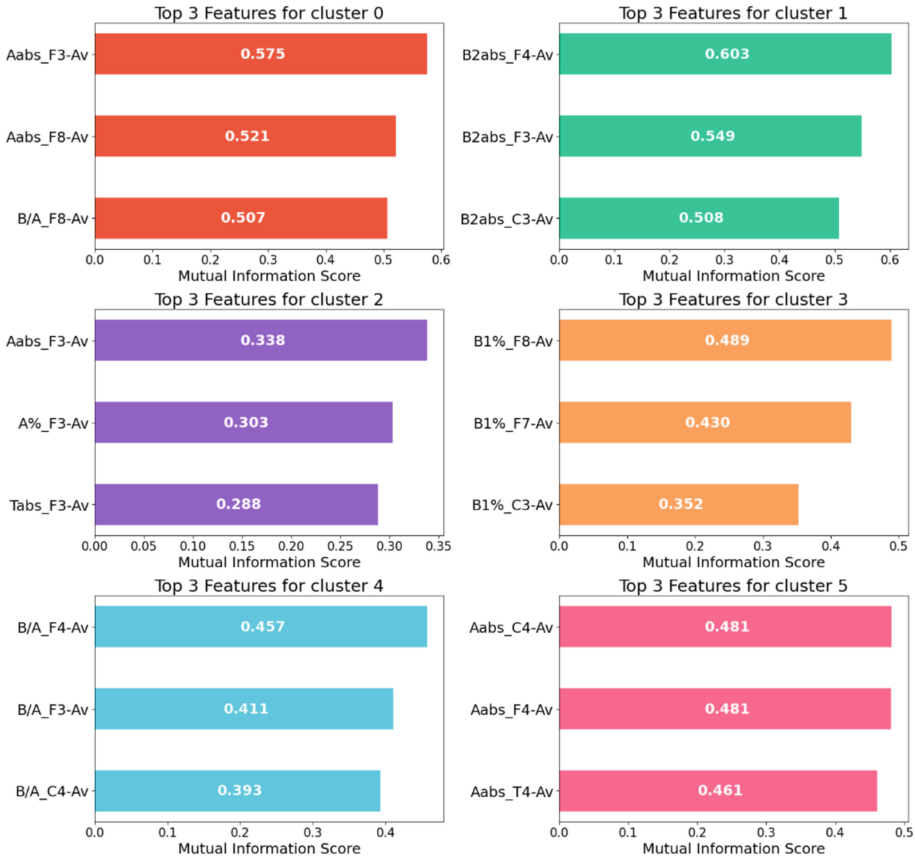


**Fig. 4.** Most informative features of HDBSCAN clusters

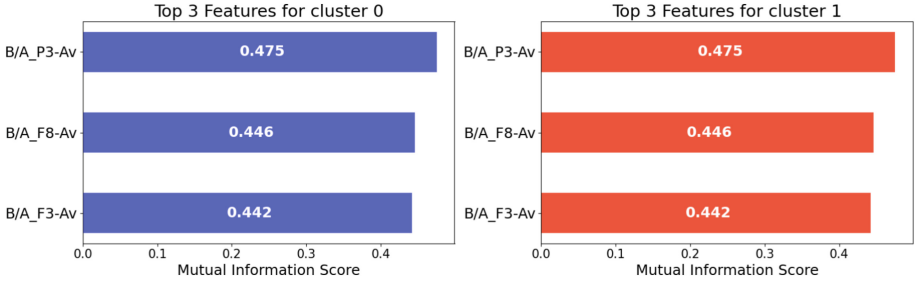


**Fig. 5.** Most informative features of agglomerative clustering clusters

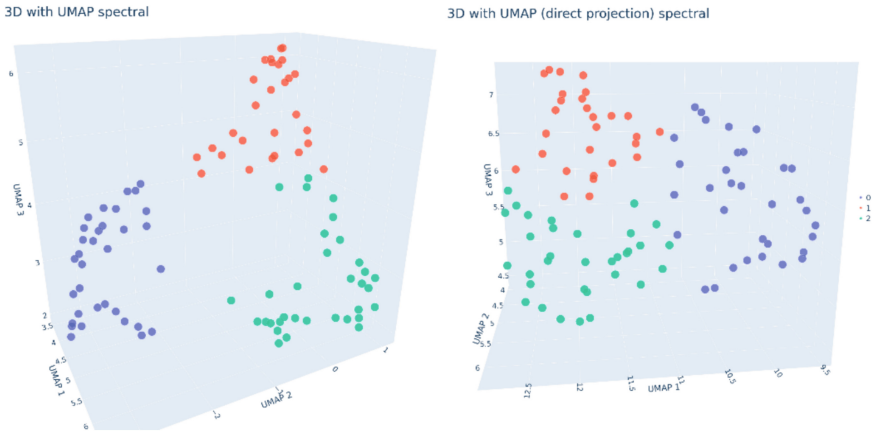


**Fig. 6.** Visualization of clustering results using Spectral Clustering

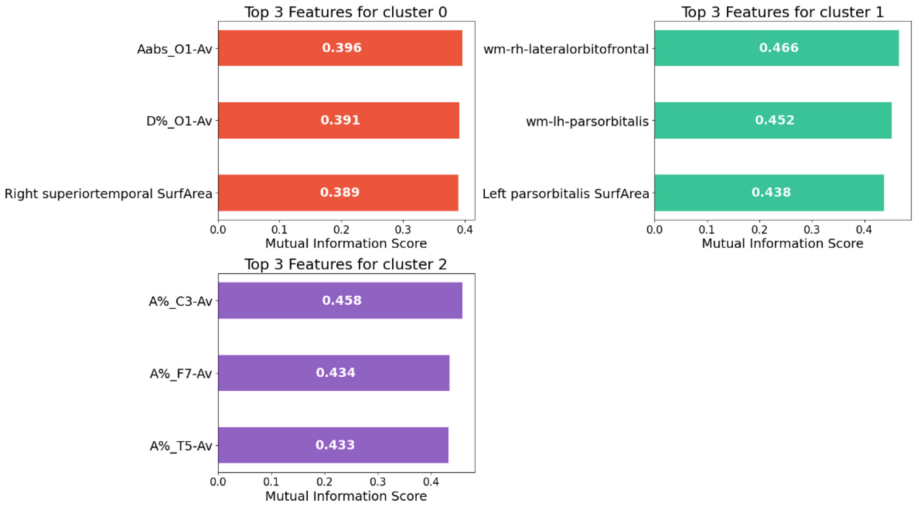


**Fig. 7.** Most informative features of spectral clustering clusters

**Table 6.** Distribution of patients by cluster using Spectral Clustering

| Cluster Label | No. of cases | No. of Clinician Diagnoses in clusters | Clinician Diagnoses Across Clusters (%) | Clinician Diagnoses Within Clusters (%) |
|---|---|---|---|---|
| 0 | 33 | 15/3/6/5/4 | 60.0/33.3/28.6/29.4/18.2 | 45.5/9.1/18.2/15.2/12.1 |
| 1 | 28 | 4/3/7/6/8 | 16.0/33.3/33.3/35.3/36.4 | 14.3/10.7/25.0/21.4/28.6 |
| 2 | 33 | 6/3/8/6/10 | 24.0/33.3/38.1/35.3/45.5 | 18.2/9.1/24.2/18.2/30.3 |

**Table 7.** Percentage labeled, metrics and cluster stability scores (EEG + EEG_sumup + MRI)

| Algorithm | Labeled (%) | DBCV | Silhouette | ARI | AMI |
|---|---|---|---|---|---|
| HDBSCAN | 70.21 | 0.25 | 0.25 | 0.66 | 0.72 |
| Agglomerative | 100.00 | 0.22 | 0.34 | 0.51 | 0.51 |
| Spectral | 100.00 | -0.06 | 0.40 | 0.85 | 0.85 |

## 4 Conclusion

Analyzing medical data, especially in the context of neurodegenerative diseases such as Alzheimer's disease, is challenging with limited sample size and high dimensionality. This paper discusses a methodological approach to tackle this problem, focusing on data preprocessing, dimensionality reduction and clustering techniques. The clustering results, although not interpreted in terms of subject matter in this study, highlight the potential for generating new hypotheses about disease biomarkers. By using PCA and UMAP for dimensionality reduction and optimizing the clustering algorithms, the proposed approach effectively handles data sparsity while preserving meaningful structures in the data. These methods have the potential to identify novel biomarkers for the diagnosis of Alzheimer's disease without resorting to invasive procedures. Continued research on larger, more diverse datasets and additional biomarkers will be critical to improve these methods and make them effective clinical tools for early detection and prevention of severe Alzheimer's disease.

It is important to acknowledge, however, that the proposed methodology is not intended as a universal solution. The findings are based on a single, institution-specific dataset. As a result, direct generalization to other datasets – even within the same domain – should be approached with caution. Nevertheless, the broader concept of structured, sequential data processing, with an emphasis on intermediate validation and data-driven decision-making, offers a flexible framework. With appropriate adaptation, this approach may be applicable to similar medical research efforts and even extend to other domains where complex, high-dimensional datasets are encountered.

## References

1. Li, X., Feng, X., Sun, X., Hou, N., Han, F., Liu, Y.: Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2019. Front. Aging Neurosci. **14**, 937486 (2022)
2. GBD 2019 Dementia Forecasting Collaborators: Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health **7**, e105–e125 (2022)
3. Alzheimer's disease facts and figures: Alzheimers Dement. **20**(5), 3708–3821 (2024)
4. Alzheimer's Disease International: World Alzheimer's Report 2015: The Global Impact of Dementia. Alzheimer's Disease International (2015), https://www.alzint.org/resource/world-alzheimer-report-2015/, last accessed 2025/03/16
5. Hardy-Sosa, A., León-Arcia, K., Llibre-Guerra, J.J., et al.: Diagnostic accuracy of blood-based biomarker panels: a systematic review. Front. Aging Neurosci. **14**, 683689 (2022)

6. Brand, A.L., Lawler, P.E., Bollinger, J.G., et al.: The performance of plasma amyloid beta measurements in identifying amyloid plaques in Alzheimer's disease: a literature review. Alz Res Therapy **14**, 195 (2022)
7. Barthélemy, N.R., Salvadó, G., Schindler, S.E., et al.: Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. Nat. Med. **30**, 1085–1095 (2024)
8. Park, S.A., Jang, Y.J., Kim, M.K., et al.: Promising blood biomarkers for clinical use in Alzheimer's disease: a focused update. J. Clin. Neurol. **18**(4), 401–409 (2022)
9. Krix, S., Wilczynski, E., Falgàs, N., et al.: Towards early diagnosis of Alzheimer's disease: advances in immune-related blood biomarkers and computational approaches. Front. Immunol. **15**, 1343900 (2024)
10. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
11. van der Ploeg, T., Austin, P.C., Steyerberg, E.W.: Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med. Res. Methodol. **14**, 137 (2014)
12. Silvey, S., Liu, J.: Sample size requirements for popular classification algorithms in tabular clinical data: an empirical study. J. Med. Internet Res. **26** (2024)
13. Andreev, E., Ananieva, N., Zalutskaya, N., Beltceva, I.A., Neznanov, N.: Use of the voxel-based morphometry method in the diagnosis of dementia of the Alzheimer type. V.M. BEKHTEREV REVIEW OF PSYCHIATRY AND MEDICAL PSYCHOLOGY (4), 66–72 (2017)
14. Lemoine, B., Rayburn, S., Benton, R.: Data fusion and feature selection for Alzheimer's diagnosis. In: Overgaard, N.C., Smedegaard, A.Ø., Rendsvig, R.K. (eds.) Medical Information Extraction and Mining, LNCS, vol. 6232, pp. 320–327. Springer, Heidelberg (2010)
15. Korkhov, V., et al.: Data storage, processing and analysis system to support brain research. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2018, LNCS, vol 1 10943, pp. 78–90. Springer, Cham (2018)
16. Ros, F., Riad, R.: Feature and Dimensionality Reduction for Clustering with Deep Learning, 1st edn. Springer, Cham (2024)
17. Levina, E., Bickel, P.J.: Maximum Likelihood estimation of intrinsic dimension. In: Proceedings of the 18th International Conference on Neural Information Processing Systems, pp. 777–784. MIT Press, Cambridge (2004)
18. McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 (2018)
19. Chari, T., Pachter, L.: The Specious Art of Single-Cell Genomics. bioRxiv (2022)
20. Castellano-Escuder, P., Zachman, D.K., Han, K., Hirschey, M.D.: Interpretable multi-omics integration with UMAP embeddings and density-based clustering. bioRxiv (2024)
21. Busch, E.L., et al.: Multi-view manifold learning of human brain state trajectories. bioRxiv (2022)
22. Diaz-Papkovich, A., Anderson-Trocmé, L., Gravel, S.: A review of UMAP in population genetics. J. Hum. Genet. **66**, 85–91 (2021)
23. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. LNCS, vol. 7819. Springer, Heidelberg (2013)
24. McInnes, L., Healy, J.: Accelerated Hierarchical Density Clustering. arXiv:1705.07321 (2017)
25. Wierzchoń, S., Kłopotek, M.: Modern Algorithms of Cluster Analysis, 1st edn. Springer, Cham (2018)

26. Moulavi, D., Jaskowiak, P.A., Campello, R.J.G.B., Zimek, A., Sander, J.: Density-based clustering validation. In: Proceedings of the 2014 SIAM International Conference on Data Mining (SDM), pp. 839–847. SIAM, Philadelphia (2014)
27. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(1), 53–65 (1987)
28. Bergstra, J., Bardenet, R., Kégl, B., Bengio, Y.: Algorithms for hyper-parameter optimization. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 24, pp. 2546–2554. Curran Associates Inc, Red Hook (2011)
29. Watanabe, S.: Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. arXiv:2304.11127 (2023)
30. Turner, R., et al.: Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. arXiv:2104.10201 (2021)
31. Romano, S., Vinh, N.X., Bailey, J., Verspoor, K.: Adjusting for Chance Clustering Comparison Measures. arXiv:1512.01286 (2015)
32. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
33. La, V., Lee, S., Park, Y.T., d'Auriol, B.: A novel feature selection method based on normalized mutual information. Appl. Intell. **37**(2), 219–237 (2012)
34. Nagel, D., Diez, G., Stock, G.: Accurate estimation of the normalized mutual information of multidimensional data. J. Chem. Phys. **161**(5), 054107 (2024)