Human Brain Mapping

WILEY

REVIEW ARTICLE OPEN ACCESS

A Practical Guide to Identifying Robust Clusters in Neuroimaging Data

Johan Nakuci[1,2] | Dobromir Rahnev[2]

[1]U.S. Army DEVCOM Army Research Laboratory, Aberdeen, Maryland, USA | [2]School of Psychology, Georgia Institute of Technology, Atlanta, Georgia, USA

Correspondence: Johan Nakuci (jnakuci@gmail.com)

Received: 25 February 2025 | **Revised:** 13 August 2025 | **Accepted:** 16 August 2025

Funding: This work was supported by Army Research Laboratory, W911SR-15-2-0001.

Keywords: clustering reliability | consensus-based clustering | hierarchical clustering | K-means | modularity-maximization | SVM

ABSTRACT

Clustering algorithms are essential tools in data-driven research, enabling the discovery of hidden structures in complex datasets. In neuroimaging, data-driven research and clustering have been instrumental in identifying and unraveling hidden relationships. However, there are concerns associated with exploratory techniques in that they can provide erroneous results unless properly verified. Here we address this issue by examining three widely used approaches: K-means, community detection via modularity maximization, and hierarchical clustering. We first highlight their methodologies, applications, and limitations. We then discuss the critical steps for rigorous validation strategies. We further show how to apply these steps using both synthetic and real data, and provide code to facilitate their application. By contextualizing clustering within robust methodological frameworks, we demonstrate the potential of clustering-based analyses to reveal meaningful patterns and provide practical guidelines for their application in neuroscience and related fields. Clustering, when appropriately applied, is a powerful and indispensable computational method.

1 | Introduction

Data-driven research has become an integral part of modern neuroscience, offering powerful tools to uncover patterns and relationships in complex datasets. Methods such as dimensionality reduction (Van Der Maaten et al. 2009), machine learning (Bzdok et al. 2017), and clustering (Altman and Krzywinski 2017) provide the means to identify structure within high-dimensional data without relying on pre-specified hypotheses. Dimensionality reduction techniques, such as Principal Component Analysis (Lever et al. 2017) or t-distributed Stochastic Neighbor Embedding (van der Maaten and Hinton 2008), enable researchers to distill large datasets into low-dimensional representations, revealing dominant trends. Supervised machine learning methods allow for predictive modeling, linking inputs to outcomes, while unsupervised techniques like clustering focus on discovering hidden structures in the data. Collectively, these computational techniques have revolutionized the study of the brain.

Among the different techniques used in data-driven research, clustering has unraveled the foundational architecture of brain organization (Yeo et al. 2011), object representation (Kamitani and Tong 2005), and diagnostic boundaries (Drysdale et al. 2017). There are at least three reasons why clustering analyses are important for neuroimaging research. The main reason to perform a clustering analysis is because one wants to investigate if there are underlying subgroups in the data. A second reason to cluster the data is to determine if there is an undesired factor that strongly influences the data. If such a factor exists, clustering could help with identifying it and potentially suggesting ways of mitigating its influence. A third reason to cluster is to reduce the dimensionality of the data. For instance, activity across neurons

that exhibit similar patterns over time can be averaged together (Stringer et al. 2025). The benefit of using clustering for dimensionality reduction is that it might increase the signal-to-noise ratio while preserving the variance in the data. However, despite the many benefits of clustering, the interpretability and reliability of this analysis technique remain a critical concern.

Indeed, a key challenge associated with clustering is that it can lead to erroneous results unless properly verified. Said another way, "clustering finds patterns in data—whether they are there or not" (Altman and Krzywinski 2017). In other words, a challenge associated with clustering techniques is that even when no true clusters exist in the data, these algorithms will still partition the data, potentially creating an illusion of meaningful groupings.

The concern that clustering algorithms will find clusters raises an important question: how can we trust that the identified clusters are meaningful and not artifacts of the algorithm? Here we propose a framework that one can use to increase the confidence in the clustering results. The framework highlights: (1) consensus-based partitioning as a means to increase the confidence that the identified clusters reflect stable partitions and not simply random partitions of the data; (2) classifier-based corroboration as a means to quantitatively assess the separability of identified clusters; and (3) comparison against noise, experimental/task conditions, and demographics to determine if the identified clusters do not simply reflect differences in potential confounds. We aim to provide a framework for leveraging clustering methods effectively in neuroscience research that emphasizes rigorous methodology and context-driven interpretation.

In Section 2, we provide a brief overview of commonly used clustering methods in neuroscience. In Section 3, we then focus on the critical analyses for identifying robust clustering results. In Section 4, we apply these methods in simulated and real data. Overall, we argue that by leveraging these strategies, researchers can better ensure that clustering-based insights are not merely artifacts of the algorithm but instead represent stable, reproducible, and robust patterns in the data.

2 | Clustering Algorithms

Clustering algorithms play a crucial role in analyzing complex datasets by identifying patterns and grouping data points based on their similarities. These methods vary in their underlying principles and applications, offering unique advantages and challenges. Partitioning methods, such as K-means, focus on dividing data into distinct groups with predefined criteria, making them efficient and straightforward for large-scale problems. In contrast, network-based methods, such as community detection through modularity maximization, are specifically designed to uncover densely connected subgroups in graph-like structures. Hierarchical clustering provides an alternative approach by constructing a nested hierarchy of clusters, revealing relationships between data points across multiple levels of granularity. In this section, we provide a brief exploration of these three widely used clustering algorithms—K-means, modularity-maximization, and hierarchical clustering—highlighting their methodologies, strengths, and limitations (Figure 1). (For detailed description on other types of clustering methods see Jaeger and Banks 2023 and Saxena et al. 2017).

2.1 | K-Means Clustering

K-means clustering is one of the simplest and most widely used clustering methods (Altman and Krzywinski 2017). It partitions

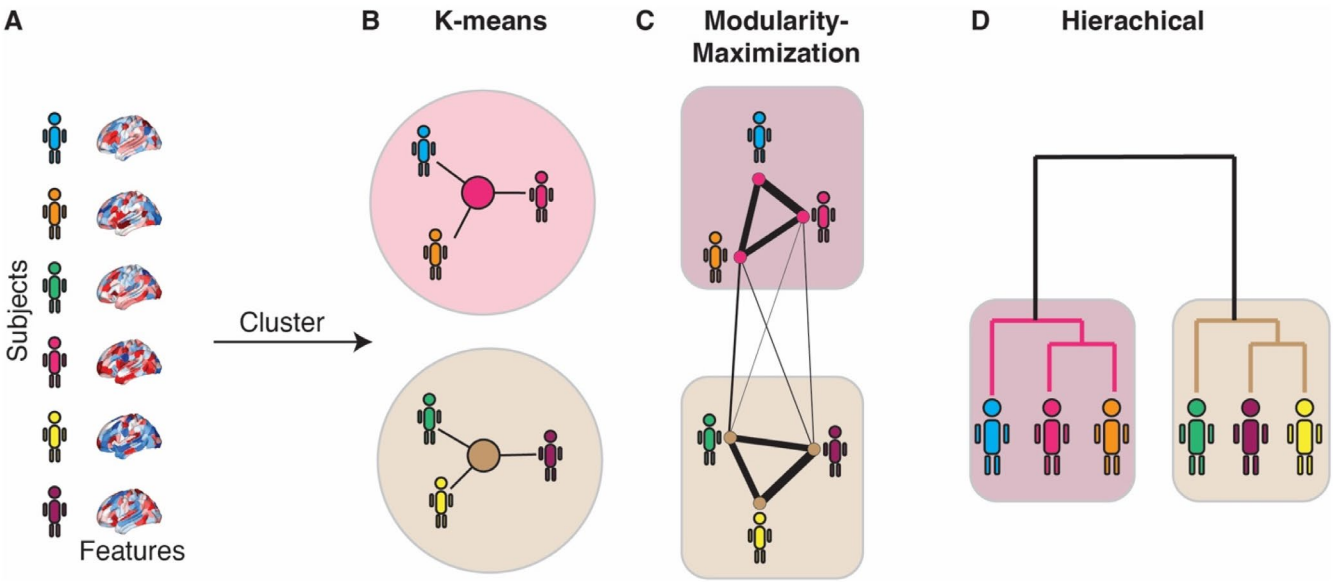

FIGURE 1 | Commonly used clustering algorithms. (A) Brain features are used to cluster the six individuals. (B) K-means clustering: A centroid-based method that partitions the data into a predefined number of clusters, assigning each individual to the nearest cluster center. (C) Modularity-maximization: A graph-based clustering approach that detects communities within a network by optimizing modularity, a measure of within-cluster connectivity. (D) Hierarchical clustering: A method that builds a tree-like structure of nested clusters, allowing for flexible cluster selection at different levels of granularity. Note that clustering analyses are not limited to between subjects and can also be conducted within a subject. Brain features in the figure are only for visualization purposes and are not meant to be indicative of actual data.

a dataset into a predefined number of clusters by assigning each data point to the cluster whose centroid is nearest, iteratively updating centroids to minimize the within-cluster variance. Its simplicity and computational efficiency make K-means particularly attractive for handling large datasets. The algorithm can adapt to diverse data types through the modification of distance metrics, such as Euclidean, Manhattan, or cosine distance, making it suitable for a variety of applications. In neuroimaging, the seminal work by Yeo et al. (2011) used K-means clustering to identify the functional organization of the brain.

Despite its versatility, K-means faces certain limitations. Its reliance on the assumption of spherical clusters and its sensitivity to initialization can lead to suboptimal results. Furthermore, the need to specify the number of clusters (K) a priori is often challenging, especially for exploratory analyses. To address this, methods such as the silhouette score, gap statistic, and elbow method are commonly employed to determine the optimal K (Kodinariya and Makwana 2013). While K-means does not work well for non-spherical clusters or datasets with overlapping structures, its ability to provide interpretable results through cluster centroids ensures its continued relevance as a tool for analyzing structured datasets.

2.2 | Modularity-Maximization

Modularity-maximization focuses on identifying groups of nodes, or communities, within a network where connections are denser internally than externally (Newman 2004). Modularity-maximization evaluates the quality of a network's partition by comparing the observed strength of edges within communities to the expected density in a random network. Maximizing the modularity score provides a partition where communities are well-defined, and higher modularity scores indicate stronger community structures. Algorithms such as the Louvain method efficiently optimize modularity, making this technique scalable for networks of moderate size. (See Zamani Esfahlani et al. 2021 for a detailed review on the usage of modularity-maximization in neuroimaging).

Modularity-maximization has proven invaluable for analyzing systems ranging from social networks (Girvan and Newman 2002) to brain connectivity (Sporns and Betzel 2016). One advantage of modularity-maximization based clustering is that it does not require the number of clusters to be specified a priori. However, modularity-maximization is not without limitations. The resolution limit of modularity can obscure smaller communities in large networks, but various methods and solutions have been examined for parameter selection to mitigate the issue (Arenas et al. 2008; Pinto et al. 2024; Reichardt and Bornholdt 2006). Additionally, some algorithms, such as Louvain, are sensitive to initialization, which can affect the reproducibility of results. Despite these challenges, modularity-maximization remains a powerful tool for uncovering the structure of complex networks.

2.3 | Hierarchical Clustering

Hierarchical clustering provides a fundamentally different approach to clustering by organizing data into a nested hierarchy of clusters (Reynolds et al. 2006). This method generates a tree-like structure, known as a dendrogram, which captures relationships between data points at various levels of granularity. Hierarchical clustering can be performed in two primary ways: agglomerative, which begins with each data point as its own cluster and iteratively merges the closest clusters, and divisive, which starts with all data points in a single cluster and recursively splits them. In neuroscience, hierarchical clustering has shown how the propagation of activity proceeds in clustered neuronal ensembles during a seizure (Feldt Muldoon et al. 2013).

Similar to modularity-maximization, an advantage of hierarchical clustering is that it does not require the number of clusters to be specified a priori. Instead, the structure of the data determines the appropriate number of clusters. The flexibility of hierarchical clustering extends to the choice of linkage criteria, such as single linkage, complete linkage, average linkage, or Ward's method, which minimizes within-cluster variance (Reynolds et al. 2006). This adaptability makes hierarchical clustering particularly valuable for exploratory data analysis, as the dendrogram provides a rich visualization of the clustering process. However, hierarchical clustering also has limitations. Its computational complexity makes it unsuitable for very large datasets, and its sensitivity to noise and outliers can distort the hierarchical structure. Moreover, once a merge or split is made, it cannot be undone, which may lead to suboptimal clustering decisions. Despite these challenges, hierarchical clustering's ability to handle non-spherical clusters and provide hierarchical perspectives ensures its utility in uncovering complex patterns in data.

2.4 | Other Clustering Methods

In addition to K-means, modularity-maximization, and hierarchical clustering, there are other, less commonly used clustering methods such as spectral clustering, deep learning-based clustering, autoencoder-based clustering, and self-organizing maps. Spectral clustering is a graph-based method that performs dimensionality reduction before clustering, making it more robust to noise and non-linear relationships in the data (Von Luxburg 2007). Other approaches would be to use deep learning-based clustering (Zhou et al. 2024), autoencoder-based clustering (Lu and Li 2021) or self-organizing maps (Kohonen 2013) to cluster the data. These methods leverage neural networks to learn structures within the data and improve clustering performance on complex or high-dimensional data that might be missed by traditional methods. However, training these models often requires large amounts of data and therefore may not be suitable for typical neuroimaging datasets.

2.5 | Selecting a Clustering Technique

The multitude of clustering techniques raises the question of which one to use. Selecting the most appropriate clustering approach is challenging since the different clustering methods have not been extensively compared in the context of neuroimaging data, especially across different sample sizes, numbers of features for clustering, or the types of data. As a result, little is

known about when one method is appropriate versus another. Here, we focus on best practices for validating the results of any clustering method, which can ultimately also help with selecting the most appropriate clustering technique for a given analysis.

3 | Establishing the Robustness of Clustering Results

Ensuring the robustness of clustering results is critical for deriving meaningful and reproducible insights across scientific disciplines. Clustering methodologies are inherently sensitive to variations in data quality, parameter settings, and algorithmic choices, which can introduce instability in the identified cluster structures. Therefore, it is essential to implement validation strategies that assess the consistency and reliability of clustering solutions and to distinguish genuine patterns from noise-induced artifacts.

In recent years, multiple metrics have been developed to evaluate clustering results (Gan et al. 2020). These include measures such as root-mean-squared standard deviation, sum-of-the-squares, and normalized Hubert Γ statistic, which assess within-cluster homogeneity or between-cluster separation (Halkidi et al. 2002). For a detailed review on these and other metrics, see Gao et al. (2023) and Saxena et al. (2017).

These are useful metrics that can estimate the agreement between the clustering results and the underlying data (Gao et al. 2023). Moreover, these metrics can be used to determine the number of optimal clusters in the data that minimize the within-cluster variation, a critical challenge in all clustering analyses. However, these metrics tend to increase as the number of clusters increases and therefore can erroneously suggest that the data contains more clusters than are really present. Additionally, these methods provide limited information on the underlying factors driving the clustering. Despite these limitations, metrics such as root-mean-squared standard deviation are valuable for assessing the quality of a partition, but other methods are needed to assess the stability of the cluster composition or the underlying factors driving the separation.

Below we describe three essential elements that clustering analyses could use to evaluate the robustness of the results. These elements include consensus clustering, classifier-based corroboration, and comparative analyses against noise, experimental/task, and demographic factors. By systematically incorporating these techniques, researchers can increase confidence in the stability and interpretability of clustering results, ensuring that the identified patterns generalize across datasets and conditions.

3.1 | Consensus-Based Partitioning

One of the primary concerns in clustering analyses is the stability and reliability of the identified clusters (Altman and Krzywinski 2017). Clustering results can be sensitive to variations in initial conditions, parameter choices, and data noise, potentially leading to inconsistent and irreproducible findings. Consensus-based partitioning provides a methodology to ensure robustness by aggregating multiple clustering results into a single consensus solution (Strehl and Ghosh 2002; Topchy et al. 2005). This step is critical in scenarios where randomness, noise, and choice of parameters may compromise the integrity of the clustering process, ensuring that identified clusters reflect meaningful structures in the data rather than artifacts of specific algorithmic conditions. To perform consensus-based partitioning, one needs to perform the following steps.

3.1.1 | Step 1: Generate Multiple Clustering Solutions

To begin, multiple clustering solutions should be generated using the same algorithm but with variations in parameters, initial conditions, or data subsets (Figure 2A,B). In our analyses, we primarily employed modularity-maximization based clustering and iterated the clustering process 100 times. Multiple clustering solutions can be obtained by:

- Altering the initial conditions (e.g., different random seeds for K-means).
- Modifying parameter settings (e.g., different values of K in K-means or γ in modularity-maximization) (Jeub et al. 2018).
- Clustering data across different types of features, allowing for multimodal clustering–consensus partitioning (Nakuci et al. 2022).

3.1.2 | Step 2: Construct an Affinity Matrix

Once multiple clustering solutions have been obtained, an affinity matrix is constructed (Figure 2C). This matrix captures how frequently pairs of data points are assigned to the same cluster across iterations (Lancichinetti and Fortunato 2012). To ensure robustness, the matrix is thresholded by retaining only values that exceed what would be expected from a random affinity matrix (i.e., a null model estimated by permuting the partitions) as suggested by (Bassett et al. 2013).

3.1.3 | Step 3: Derive the Consensus Partition

The consensus-based partition is obtained by applying a secondary clustering procedure to the thresholded affinity matrix. This step aggregates consistent clustering assignments across iterations, ensuring stability and reliability. By following these steps, consensus-based partitioning mitigates the influence of randomness and parameter sensitivity, leading to more robust and interpretable clustering results. This approach has been shown to identify consistent and stable partitions critical for identifying meaningful structures rather than artifacts introduced by specific algorithmic conditions or data noise (Bassett et al. 2013; Kwak et al. 2009; Lancichinetti and Fortunato 2012).

In an application, we have utilized this approach to identify robust multiple patterns of brain activity across trials in EEG (Nakuci, Covey, et al. 2023; Nakuci, Samaha, and Rahnev 2023) and fMRI (Nakuci et al. 2025). Brain activity during a task is highly variable, but the prevailing assumption is that there exists only a single pattern of activity identified by averaging across trials and any variability simply reflects noise (Arieli et al. 1996;

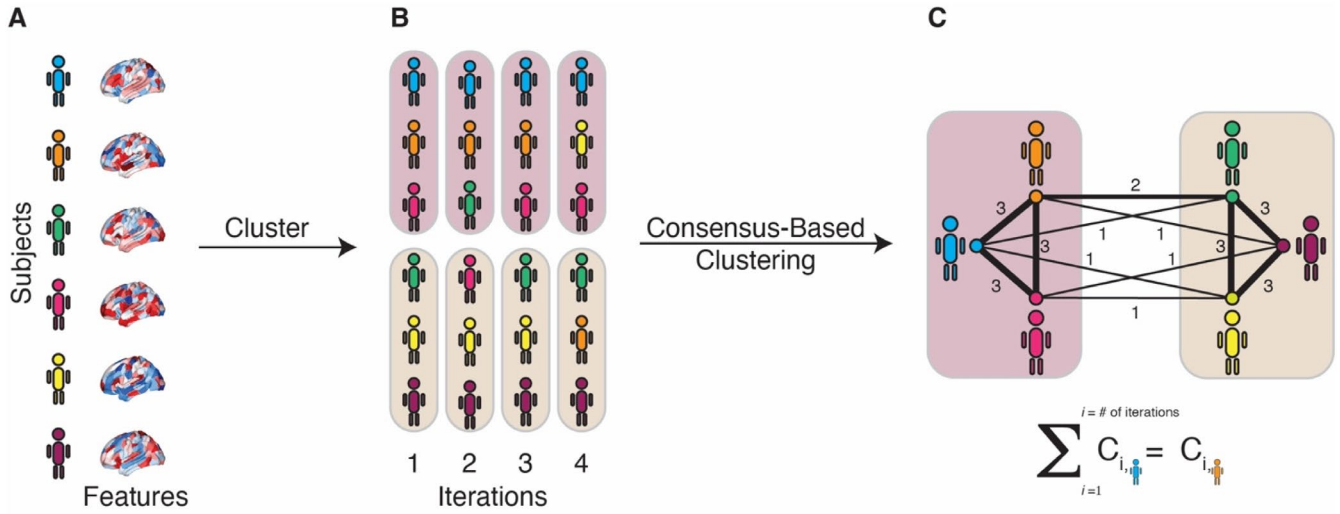

FIGURE 2 | Consensus-based partitioning. (A) Brain features are used to cluster the six individuals. (B) Example partitions of subjects across multiple iterations of the clustering. (C) Graphical visualization of the affinity matrix and consensus-based clustering. Consensus-based partitioning is derived from clustering the affinity matrix. The strength of connection between two subjects is estimated as the number of times two subjects are in the same cluster across iterations. The consensus matrix is then clustered to obtain the final partition of subjects into clusters. Note that brain features are only for visualization purposes and not meant to be indicative of actual data.

Goris et al. 2014). However, it is also possible that subsets of trials can produce meaningfully different patterns of activity that are not well captured by averaging brain activity across all trials. Therefore, we used clustering analyses to determine if the trial-to-trial variability was indicative of multiple patterns of brain activity associated with the completion of the same task. Specifically, we first applied modularity-maximization clustering to partition single-trial brain activity. The clustering process was iterated 100 times, and the results were aggregated into an affinity matrix, which was then clustered to derive the final partition. The analysis identified unique patterns of brain activity that were robust across subjects and datasets.

Moreover, consensus-based clustering can be applied to identify subtypes within a population by clustering across multiple features. For instance, we have investigated if the properties of large-scale structural brain networks derived from diffusion MRI in rats can identify subpopulations with differential propensities to develop post-traumatic epilepsy (Nakuci et al. 2022). Approximately 40% of rats that receive a traumatic brain injury using the lateral fluid-percussion injury model will develop spontaneous seizures, but biomarkers are crucially lacking (Smith et al. 2018). To address this issue, we first clustered animals based on 16 diverse brain network properties, then generated an affinity matrix based on how frequently two animals were in the same cluster in each of the features and followed by consensus-based clustering. This procedure identified a subpopulation of animals in which structural-functional modeling indicated increased propensities to synchronize brain activity relating to the heterogeneity in clinical outcomes.

One advantage of clustering across multiple features is that it eliminates the need for a priori feature selection. However, the clustering results obtained on individual features could be driven by noise. Consensus-based clustering addresses this concern by identifying consistent relationships across multiple features rather than relying on any single feature.

In summary, consensus-based clustering is an important tool for ensuring the reliability and stability of clustering solutions in complex datasets. By systematically aggregating multiple clustering results, it mitigates the impact of noise and randomness, leading to more robust and reproducible findings. This method is particularly valuable in neuroscience and other data-driven disciplines where identifying meaningful subpopulations is crucial for scientific discovery. As clustering techniques continue to evolve, consensus clustering will remain a cornerstone methodology for enhancing the validity of clustering-based analyses.

3.2 | Classifier-Based Corroboration

A classifier-based corroboration analysis serves as a complementary approach to consensus-based clustering by quantitatively assessing the separability of identified clusters. Machine learning classifiers, such as SVM, enable an empirical evaluation of whether the discovered clusters correspond to meaningful, distinguishable patterns rather than statistical noise. The classifier-based analysis is an important step for validating the clustering results because the classifier uses a different method for labeling the data compared to clustering. Specifically, the classifier-based analysis corroborates the clustering analysis because the clustering analysis implies that there is some signal in the data that is sufficiently strong to separate the data. The signal separating the data is learned by the classifier in a subset of the data and should correctly identify the cluster labels in the remaining data. A high classification accuracy would indicate that the identified clusters exhibit distinct and reproducible features. To perform classifier-based corroboration, one needs to perform the following steps.

3.2.1 | Step 1: Separate the Data Into Testing and Training Sets

For the classifier-based corroboration analysis, separate the data into a training and testing set obtained from consensus-based analysis (Figure 3A). In some cases, this might raise concerns pertaining to data leakage. To mitigate these concerns, prior to clustering, randomly split the data in half and cluster each half separately, followed by consensus-based partitioning. We note that, in our experience, we have obtained the same results whether the data was separated before (Nakuci et al. 2025) or after clustering (Nakuci, Covey, et al. 2023), indicating that data leakage is not a major factor.

3.2.2 | Step 2: Train Classifier

Train a classifier such as SVM on the training set and use it to predict the cluster labels on the testing set (Figure 3B,C). To assess the classifier's performance, compare the empirically derived cluster labels with those predicted by the SVM classifier (Figure 3D). This comparison helps evaluate the stability of clustering results and the extent to which the discovered structure is learnable by a supervised model. A crucial aspect of this analysis is determining what constitutes a "good" classification accuracy. In general, if the classification accuracy is significantly above chance (e.g., > 70% depending on the complexity of the data and number of clusters), this would indicate that the clustering solution is robust. In contrast, accuracy close to chance suggests that the clustering does not capture a clear separation in the data.

3.2.3 | Step 3: Compare to a Null Model

To establish a baseline for classifier performance, permute the cluster labels and re-run the analysis. This process should be repeated multiple times to obtain a distribution of accuracies under the null hypothesis. The number of iterations should be sufficiently large to ensure stability in the null distribution—typically at least 1000 iterations. This allows for statistical comparison, such as computing a *p*-value to assess whether the

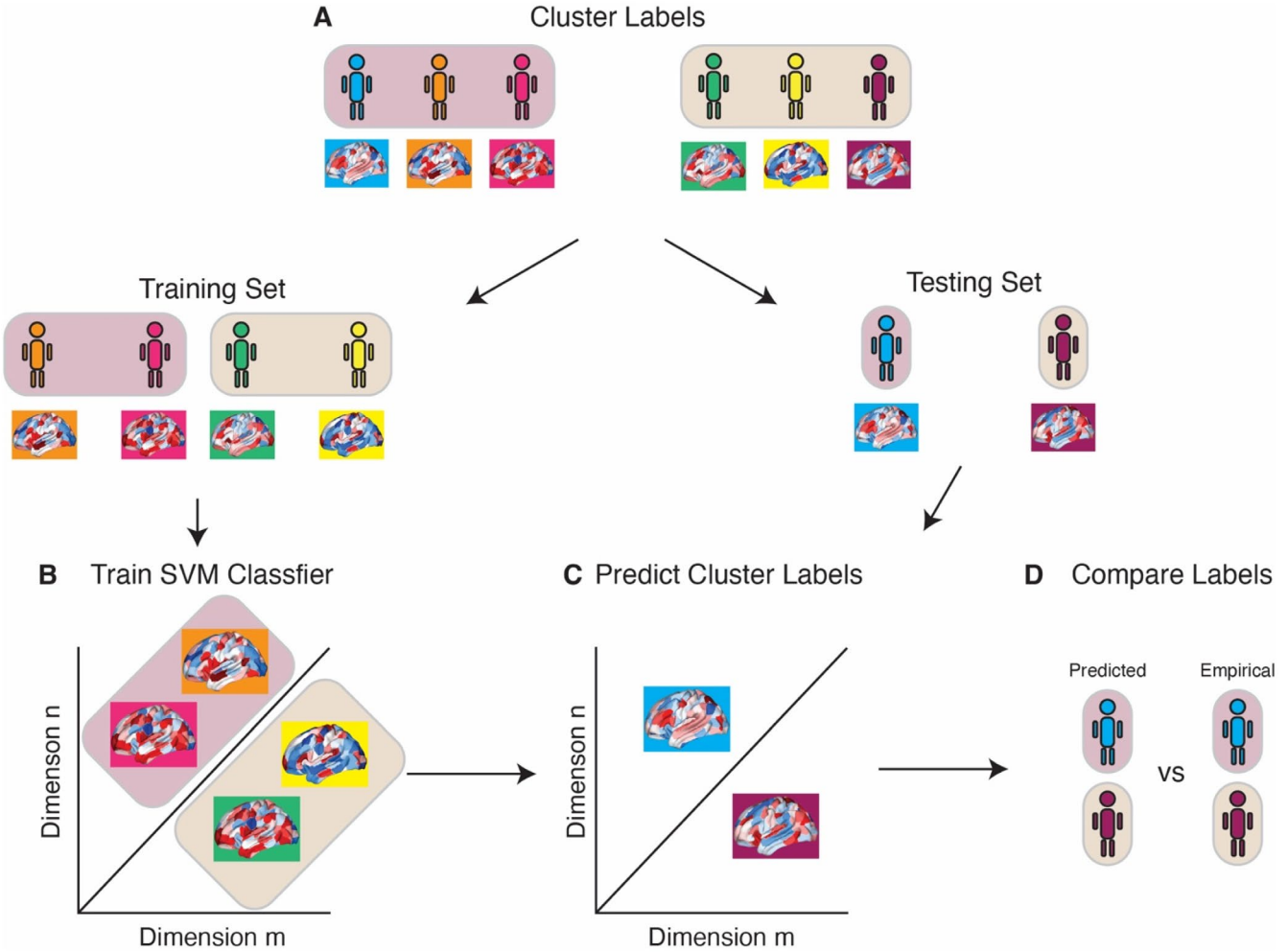

FIGURE 3 | Classifier-based corroboration. (A) Cluster labels are separated into training and testing sets. (B) An SVM classifier is trained on the training set and (C) is used to predict the cluster labels of the data in the testing set. The brain featuresed in for training and testing can represent activation patterns, network properties, and other properties. (D) Predicted and empirical labels are compared to estimate the accuracy. A classification accuracy above chance (e.g., > 70% depending on the complexity of the data and number of clusters) would indicate that the clustering solution is robust. In addition, a baseline for classifier performance can be estimated by permuting the cluster labels and re-running the analysis. Note that brain features are only for visualization purposes and not meant to be indicative of actual data.

observed classification accuracy significantly deviates from what would be expected by chance.

Beyond classification accuracy, classifier-based corroboration analyses can elucidate the key features contributing to cluster differentiation. In neuroimaging, this involves identifying brain regions or functional networks that predominantly influence classification performance, thereby providing biological relevance to the clustering results. By integrating machine learning classifiers into robustness assessments, researchers can enhance confidence in clustering-derived insights across various domains.

3.3 | Compare Against Noise, Experimental/Task Conditions and Demographics

The main reason to perform a clustering analysis is to determine if the data contains underlying hidden structure. Crucially, it is important to test if the clusters reflect factors inherent in the data. The specific factors will depend on the data, but in neuroimaging, common factors to test would be noise, experimental/task conditions, or demographics. By systematically comparing these potential confounds, clustering-based analyses can provide more reliable and interpretable insights into the organization of neural and behavioral data. The comparative analysis relies on three steps that are not necessarily sequential, and different steps may be applicable in different situations.

3.3.1 | Step 1: Test for Differences in Noise

Across empirical studies, data-driven clustering approaches are susceptible to noise confounds. Neuroimaging artifacts such as subject motion in fMRI studies may play a meaningful factor in differentiating data into clusters. Assessing robustness against such confounds is critical for ensuring valid conclusions. Standard strategies include computing noise-related metrics—such as frame displacement or signal-to-noise ratios—and testing their distribution across clusters (Power et al. 2012). (Figure 4A). A lack of significant associations between these factors and cluster assignments strengthens the reliability of the clustering results.

3.3.2 | Step 2: Test for Stability Across Experimental/Task Conditions

Another key consideration is the role of experimental/task conditions may play in shaping cluster assignments. Clustering solutions should generalize across varying experimental factors and task conditions. To test for stability across conditions: (1) plot

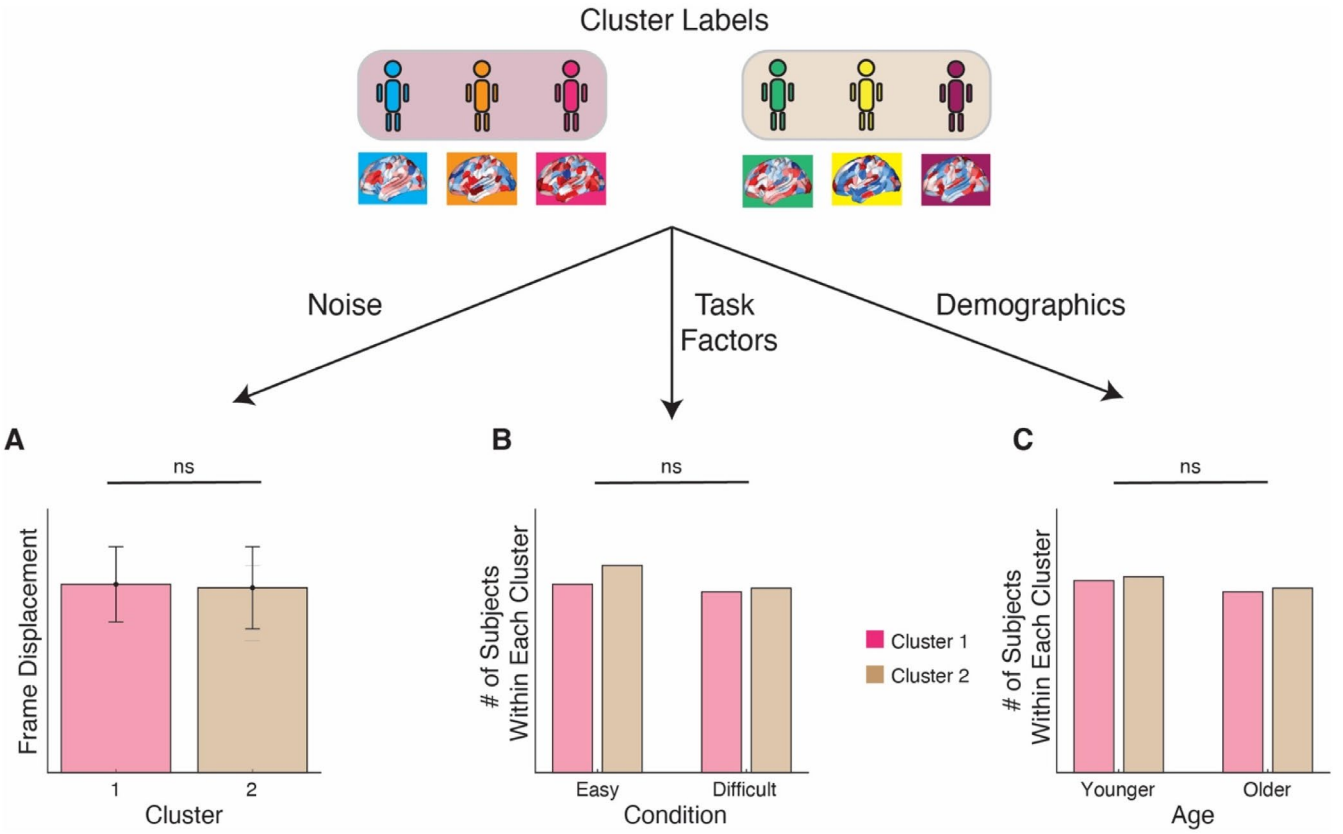

FIGURE 4 | Effects of noise, task conditions and demographics on clustering. In this example, clustering is done across subjects (i.e., each subject is assigned to a cluster) and the clustering is not driven by noise in the data, experimental conditions, or demographic factors. (A) Hypothetical frame displacement values between clusters. (B) Hypothetical separation of subjects based on task conditions within and between clusters. (C) Hypothetical separation of subjects based on demographic factors such as age. Note that the type of noise, task and demographic estimates will be specific to each dataset. This approach should be seen as a general framework for assessing the relationship between each of the factors and cluster composition. Significant differences would suggest the major source(s) of variance in the data and clusters.

the differences in the distribution of clusters among conditions, and/or (2) perform the clustering analysis separately on each experiment/task condition (Figure 4B). In our analyses focusing on single trial brain activity, to corroborate our main findings, we repeated the clustering analysis separately for each condition and found the same patterns within individual conditions, thus confirming that the task condition was not a factor driving the separation of trials (Nakuci, Covey, et al. 2023; Nakuci et al. 2025). These tests will aid in determining the extent to which clustering solutions are idiosyncratic to a specific dataset.

3.3.3 | Step 3: Test for Population Differences

To ensure that clustering solutions reflect meaningful neurobiological or behavioral distinctions rather than demographic differences between participants, it is essential to test for potential biases. Clusters should not merely segment participants based on demographic variables such as age, sex, education, or socioeconomic status (Figure 4C). Controlling for demographic effects in clustering models using covariate adjustment techniques (Kahan et al. 2014), such as propensity score matching (Caliendo and Kopeinig 2008; Haukoos and Lewis 2015) or statistical residualization, can further ensure that clusters are not simply reflecting known population differences.

Ultimately, if clusters reflect differences in noise, experimental/task conditions, or demographics, such a finding would not necessarily undermine the validity of the clustering results. Rather, it provides valuable insight into the factors driving variance in the data. By carefully evaluating these potential influences, researchers can refine their data analysis and clustering approaches so that the results are biologically or behaviorally meaningful rather than artifacts of methodological or demographic confounds.

4 | Working Examples

We next show how to apply these methods in simulated and real data. Additionally, we provide a toolbox for performing these analyses: https://github.com/jnakuci/Clustering_Toolbox.

4.1 | Application in Simulated Data

We first use simulated data, which allows us to know the ground truth. We simulated data consisting of 200 samples that belong to three clusters, with each sample containing 400 data points (Figure 5A). In the context of brain imaging, these data could be thought of as reflecting 400 ROI (data points) from 200 subjects (samples).

We first create the similarity matrix using Pearson correlation and apply modularity-maximization with a resolution parameter, $\gamma = 1$. Note that when using modularity-maximization with distance-based metrics (e.g., Euclidian distance), the distance values need to be converted to proximity values. We iterated the clustering 100 times for consensus-based partitioning, from which we created the affinity matrix (Figure 5B,C). The affinity matrix is then clustered to identify the stable partition of samples across iterations. In our example, the procedure identifies three clusters displayed on the re-oriented similarity matrix (Figure 5D).

To determine if the clusters correspond to meaningful, distinguishable patterns rather than statistical noise, an SVM classifier is used to corroborate the clustering. In our example, we used a 10-fold cross-validation procedure and found that the SVM classifier correctly predicted the cluster label in 99% of the samples (Figure 5E). However, given that this is simulated data, we cannot perform a rigorous comparison against noise, experimental/task conditions, or demographics. Nevertheless, since the data in our example are simulated to a priori contain three clusters, we can compare the empirical to the recovered cluster labels. Overall, the procedure was able to correctly label 197 of 200 samples into the appropriate clusters (Figure 5F). It should be noted that this is despite noise affecting the similarity between any two pairs (Figure 5G).

4.2 | Application in Real Data

We next show how we used the above steps in real data. Specifically, we use the data from a recent paper in which we clustered single-trial fMRI activation maps (Nakuci et al. 2025). For the analysis, the activation patterns from individual trials were concatenated across subjects, and a similarity matrix was estimated using Pearson correlation. The similarity matrix was clustered using modularity-maximization. The clustering was repeated 100× and followed by a consensus-based partition. The procedure identified three clusters, which we called subtypes (Figure 6A) and the corresponding average activation of trials in each subtype is shown in Figure 6B.

We then used SVM-based corroboration to test if the clusters correspond to meaningful, distinguishable patterns rather than statistical noise. For the analysis, we first split the data in half and performed the clustering on each half. We trained an SVM classifier on the two halves separately and then used the model to predict the subtype labels on the other half. The classifier correctly predicted the labels on 78.9% of trials (Figure 6C). Additionally, we determined how stable the clusters are when using different resolution (i.e., gamma) values in the clustering algorithm. The number of clusters increased as the gamma increased (Figure 6D, left), but, importantly, the core subtypes did not change (Figure 6D, right).

Lastly, we investigated if the clusters reflected differences in noise, experimental/task conditions, or demographics. We found that there were no significant differences between clusters in motion (i.e., frame displacement, FD; Figure 6E). Moreover, we observed that the clusters were equally distributed across the experimental conditions and experimental factors, including trial position, the time interval between successive trials, sex, and age (Figure 6F–J). Collectively, the different elements (consensus, corroboration, comparison to noise or demographics) provide support that the observed subtypes and activation patterns reflect meaningful variation in the data.

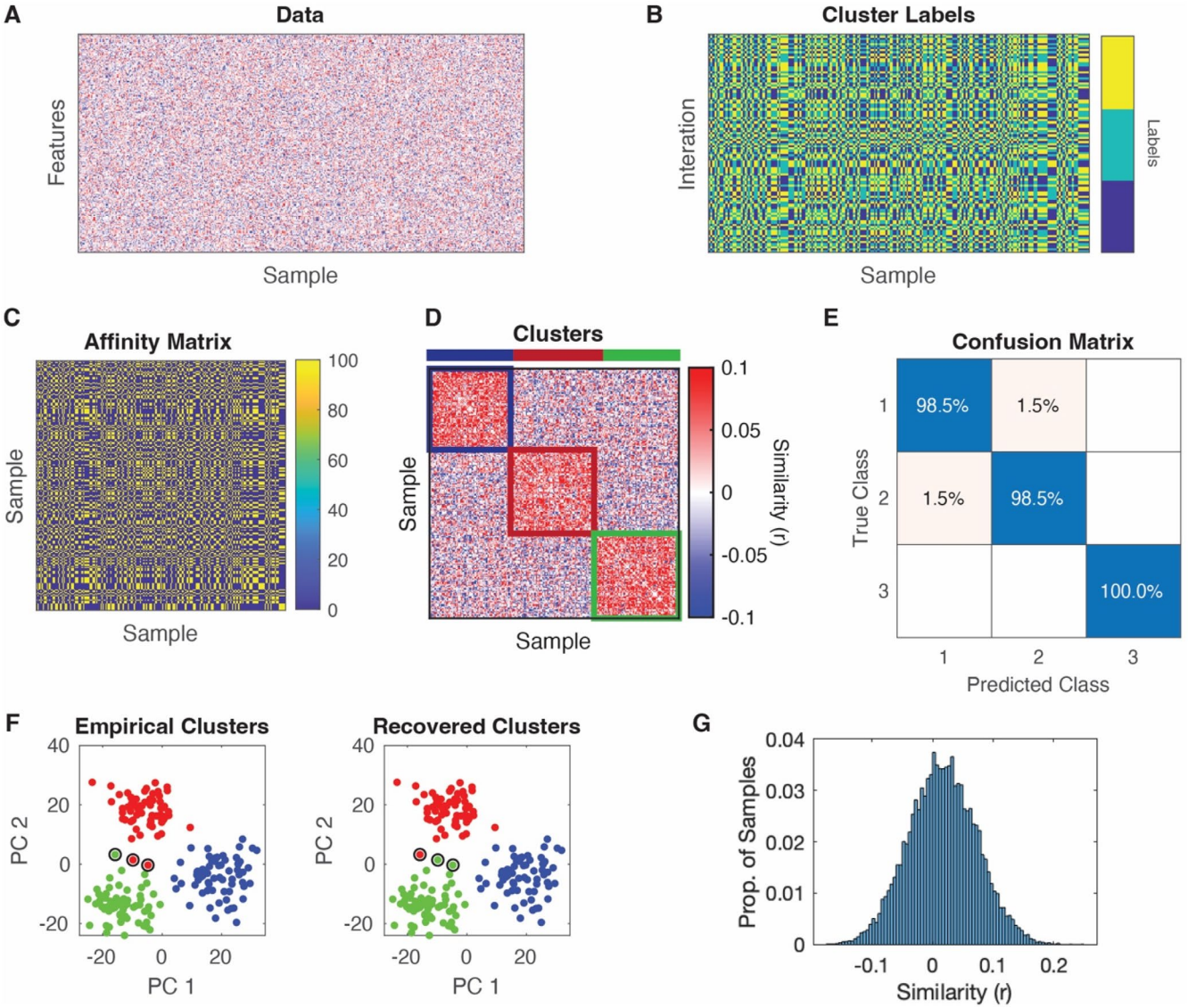

FIGURE 5 | A walk-through example with simulated data. (A) Simulated data containing 400 features from 200 samples. This can be thought of as ROI × Subject data. (B) The data is clustered using modularity-maximization with $\gamma = 1$. Specifically, the similarity matrix is estimated using Pearson correlation between samples and followed by clustering algorithm which is repeated 100×. (C) Consensus-based partition. Affinity matrix is estimated based on how often two samples are clustered together in panel B. The affinity matrix is clustered to obtain the final cluster labels. (D) Re-oriented similarity matrix based on the clustering partition. (E) Classifier-based corroboration. An SVM classifier is trained on a portion of the data and used to predict the labels in the remaining data. 10-fold cross-validation is used during training and testing. (F) Comparison of empirical and recovered cluster labels. Black circles highlight samples that were placed in the wrong cluster. (G) The distribution of the similarity values from panel D. Consensus-based partition and SVM-based corroboration are sensitive to the underlying partition despite the weak separability between individuals.

5 | Conclusion

Clustering methods serve as powerful tools for uncovering patterns in high-dimensional neuroscience data, yet their application requires careful validation to ensure meaningful and reproducible insights (Lange et al. 2004). A key concern is whether the identified clusters represent meaningful partitioning of the data. Here we provide a suite of methods that can increase confidence in the quality of the clustering results. Specifically, we highlight three critical elements: consensus-based clustering to find a stable partition, corroborating the partition using alternative methods such as classifier-based methods, and comparison against noise, experimental/task conditions, and demographics.

As the field of neuroscience increasingly relies on data-driven approaches, clustering will remain an essential technique for characterizing brain organization, connectivity, and function. However, its utility hinges on the implementation of robust validation procedures that guard against over interpretation and ensure reproducibility. By integrating consensus-based approaches, classification-based corroboration, and rigorous control analyses, researchers can harness the full potential of clustering to extract meaningful insights from complex neural data.

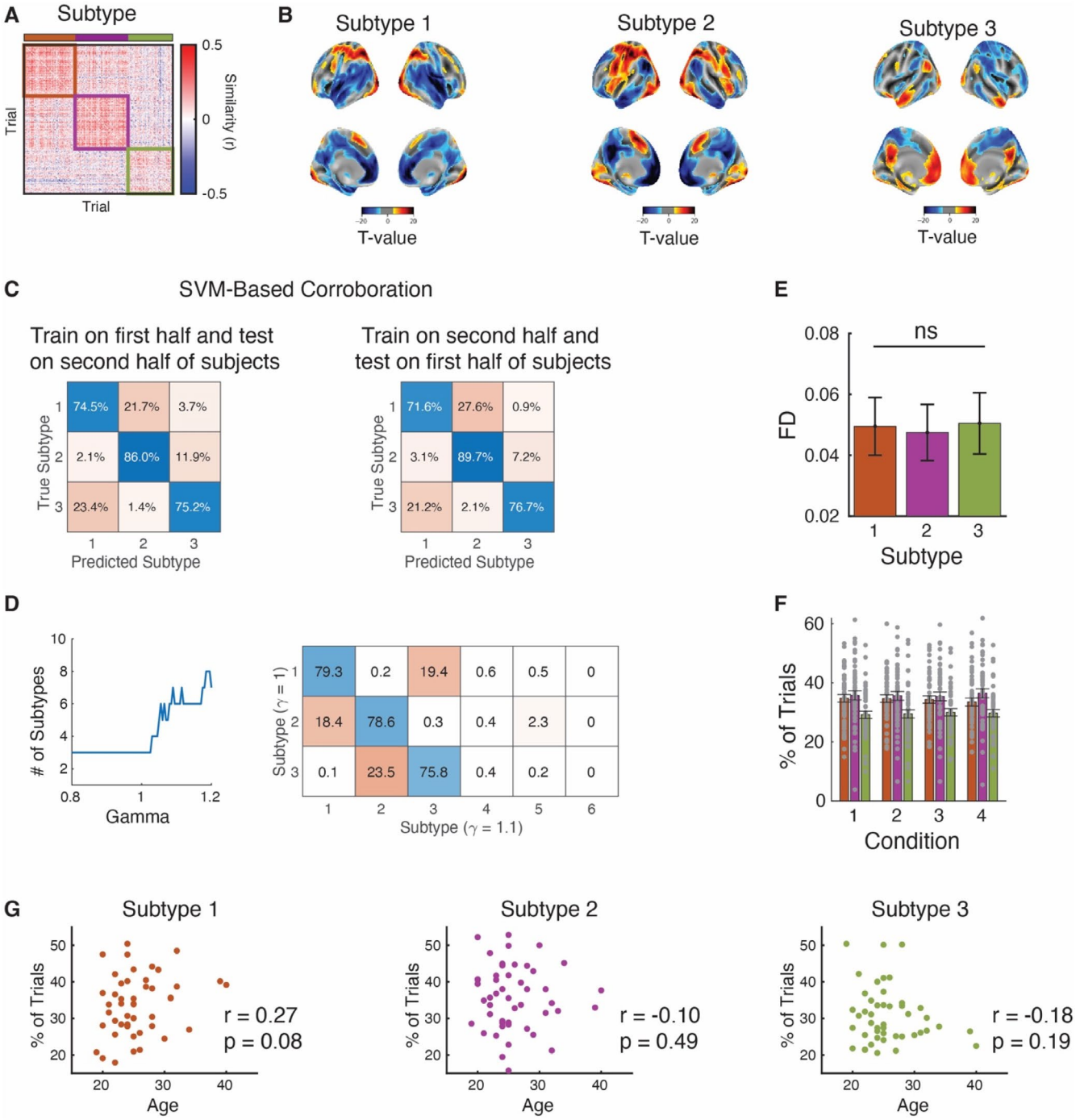

FIGURE 6 | A walk-through example with real data. (A) Modularity-maximization based clustering identified three subtypes of trials. (B) Activation maps for each subtype estimated by first averaging the trials for each subtype within a subject, followed by one-sided one-sample *t*-tests to identify regions in which brain activity increased or decreased in response to the task. (C) SVM-based corroboration. The SVM classifier correctly labeled on average 78.9% of trials across all tasks. (D) Sensitivity of clustering to resolution parameter. The number of clusters (subtypes) was stable over a range of resolution parameter (γ) values from 0.8 to 1.01 for each experiment (*left*). The increased number of subtypes as γ increases arises from separating a few trials from the main subtypes. To demonstrate this, we compared the clusters obtained with $\gamma = 1$ and $\gamma = 1.1$. As the figure demonstrates, there is a strong mapping between the first three subtypes obtained with $\gamma = 1$ and $\gamma = 1.1$ (*right*). Thus, higher γ values do not lead to qualitatively different subtypes. (E) The average Frame Displacement (FD) per subtype, demonstrating that the subtypes are not driven by subject motion. (F) The percent of trials classified as Subtype 1, 2, and 3 for each of the four stimulus conditions, demonstrating that the subtypes do not simply reflect the different experimental conditions. The dots represent individual subjects ($N_{subj} = 50$). (G) Correlation between age and percentage of trials classified as Subtype 1 (*left*), Subtype 2 (*middle*), and Subtype 3 (*right*). The analysis can help determine if the frequency with which each subtype occurs across subjects is driven by subject age. The figure is reprinted from Nakuci et al. (2025).

Author Contributions

Conceptualization: J.N., D.R. Methodology: J.N., D.R. Data Curation: J.N., D.R. Visualization: J.N., D.R. Funding acquisition: J.N., D.R. Writing – original draft: J.N., D.R. Writing – review and editing: J.N., D.R.

Acknowledgements

This research was supported by the U.S. Army DEVCOM Army Research Laboratory through army educational outreach program (W911SR-15-2-0001). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army DEVCOM Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

Altman, N., and M. Krzywinski. 2017. "Clustering." *Nature Methods* 14: 545–546.

Arenas, A., A. Fernandez, and S. Gomez. 2008. "Analysis of the Structure of Complex Networks at Different Resolution Levels." *New Journal of Physics* 10: 053039.

Arieli, A., A. Sterkin, A. Grinvald, and A. Aertsen. 1996. "Dynamics of Ongoing Activity: Explanation of the Large Variability in Evoked Cortical Responses." *Science* 273: 1868–1871.

Bassett, D. S., M. A. Porter, N. F. Wymbs, S. T. Grafton, J. M. Carlson, and P. J. Mucha. 2013. "Robust Detection of Dynamic Community Structure in Networks." *Chaos* 23: 013142.

Bzdok, D., M. Krzywinski, and N. Altman. 2017. "Machine Learning: A Primer." *Nature Methods* 14: 1119–1120.

Caliendo, M., and S. Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22: 31–72.

Drysdale, A. T., L. Grosenick, J. Downar, et al. 2017. "Resting-State Connectivity Biomarkers Define Neurophysiological Subtypes of Depression." *Nature Medicine* 23: 28–38.

Feldt Muldoon, S., I. Soltesz, and R. Cossart. 2013. "Spatially Clustered Neuronal Assemblies Comprise the Microstructure of Synchrony in Chronically Epileptic Networks." *Proceedings of the National Academy of Sciences* 110: 3567–3572.

Gan, G., C. Ma, and J. Wu. 2020. *Data Clustering: Theory, Algorithms, and Applications*. SIAM.

Gao, C. X., D. Dwyer, Y. Zhu, et al. 2023. "An Overview of Clustering Methods With Guidelines for Application in Mental Health Research." *Psychiatry Research* 327: 115265.

Girvan, M., and M. E. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences of the United States of America* 99: 7821–7826.

Goris, R. L. T., J. A. Movshon, and E. P. Simoncelli. 2014. "Partitioning Neuronal Variability." *Nature Neuroscience* 17: 858–865.

Halkidi, M., Y. Batistakis, and M. Vazirgiannis. 2002. "Clustering Validity Checking Methods: Part II." *ACM SIGMOD Record* 31: 19–27.

Haukoos, J. S., and R. J. Lewis. 2015. "The Propensity Score." *JAMA* 314: 1637–1638.

Jaeger, A., and D. Banks. 2023. "Cluster Analysis: A Modern Statistical Review." *Wiley Interdisciplinary Reviews: Computational Statistics* 15: e1597.

Jeub, L. G. S., O. Sporns, and S. Fortunato. 2018. "Multiresolution Consensus Clustering in Networks." *Scientific Reports* 8: 3259.

Kahan, B. C., V. Jairath, C. J. Doré, and T. P. Morris. 2014. "The Risks and Rewards of Covariate Adjustment in Randomized Trials: An Assessment of 12 Outcomes From 8 Studies." *Trials* 15: 139.

Kamitani, Y., and F. Tong. 2005. "Decoding the Visual and Subjective Contents of the Human Brain." *Nature Neuroscience* 8: 679–685.

Kodinariya, T. M., and P. R. Makwana. 2013. "Review on Determining Number of Cluster in K-Means Clustering." *International Journal* 1: 90–95.

Kohonen, T. 2013. "Essentials of the Self-Organizing Map." *Neural Networks* 37: 52–65.

Kwak, H., Y.-H. Eom, Y. Choi, H. Jeong, and S. Moon. 2009. "Consistent Community Identification in Complex Networks." arXiv preprint arXiv:09101508.

Lancichinetti, A., and S. Fortunato. 2012. "Consensus Clustering in Complex Networks." *Scientific Reports* 2: 336.

Lange, T., V. Roth, M. L. Braun, and J. M. Buhmann. 2004. "Stability-Based Validation of Clustering Solutions." *Neural Computation* 16: 1299–1323.

Lever, J., M. Krzywinski, and N. Altman. 2017. "Principal Component Analysis." *Nature Methods* 14: 641–642.

Lu, S., and R. Li. 2021. *DAC–Deep Autoencoder-Based Clustering: A General Deep Learning Framework of Representation Learning*, 205–216. Springer.

Nakuci, J., T. J. Covey, J. L. Shucard, D. W. Shucard, and S. F. Muldoon. 2023. "Single Trial Variability in Neural Activity During a Working Memory Task Reveals Multiple Distinct Information Processing Sequences." *NeuroImage* 269: 119895.

Nakuci, J., M. McGuire, F. Schweser, D. Poulsen, and S. F. Muldoon. 2022. "Differential Patterns of Change in Brain Connectivity Resulting From Severe Traumatic Brain Injury." *Brain Connectivity* 12: 799–811.

Nakuci, J., J. Samaha, and D. Rahnev. 2023. "Brain Signatures Indexing Variation in Internal Processing During Perceptual Decision-Making." *iScience* 26: 107750.

Nakuci, J., J. Yeon, N. Haddara, J.-H. Kim, S.-P. Kim, and D. Rahnev. 2025. "Multiple Brain Activation Patterns for the Same Perceptual Decision-Making Task." *Nature Communications* 16: 1785.

Newman, M. E. J. 2004. "Fast Algorithm for Detecting Community Structure in Networks." *Physical Review E* 69: 066133.

Pinto, L. D., J. O. Garcia, and K. Bansal. 2024. "Optimizing Parameter Search for Community Detection in Time-Evolving Networks of Complex Systems." *Chaos* 34, no. 2: 023133.

Power, J. D., K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen. 2012. "Spurious but Systematic Correlations in Functional Connectivity MRI Networks Arise From Subject Motion." *NeuroImage* 59: 2142–2154.

Reichardt, J., and S. Bornholdt. 2006. "Statistical Mechanics of Community Detection." *Physical Review E* 74: 016110.

Reynolds, A. P., G. Richards, B. de la Iglesia, and V. J. Rayward-Smith. 2006. "Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms." *Journal of Mathematical Modelling and Algorithms* 5: 475–504.

Saxena, A., M. Prasad, A. Gupta, et al. 2017. "A Review of Clustering Techniques and Developments." *Neurocomputing* 267: 664–681.

Smith, D., T. Rau, A. Poulsen, et al. 2018. "Convulsive Seizures and EEG Spikes After Lateral Fluid-Percussion Injury in the Rat." *Epilepsy Research* 147: 87–94.

Sporns, O., and R. F. Betzel. 2016. "Modular Brain Networks." *Annual Review of Psychology* 67: 613–640.

Strehl, A., and J. Ghosh. 2002. "Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions." *Journal of Machine Learning Research* 3: 583–617.

Stringer, C., L. Zhong, A. Syeda, F. Du, M. Kesa, and M. Pachitariu. 2025. "Rastermap: A Discovery Method for Neural Population Recordings." *Nature Neuroscience* 28: 201–212.

Topchy, A., A. K. Jain, and W. Punch. 2005. "Clustering Ensembles: Models of Consensus and Weak Partitions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27: 1866–1881.

van der Maaten, L., and G. Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605.

Van Der Maaten, L., E. O. Postma, and H. J. Van Den Herik. 2009. "Dimensionality Reduction: A Comparative Review." *Journal of Machine Learning Research* 10: 13.

Von Luxburg, U. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17: 395–416.

Yeo, B. T. T., F. M. Krienen, J. Sepulcre, et al. 2011. "The Organization of the Human Cerebral Cortex Estimated by Intrinsic Functional Connectivity." *Journal of Neurophysiology* 106, no. 3: 1125–1165.

Zamani Esfahlani, F., Y. Jo, M. G. Puxeddu, et al. 2021. "Modularity Maximization as a Flexible and Generic Framework for Brain Network Exploratory Analysis." *NeuroImage* 244: 118607.

Zhou, S., H. Xu, Z. Zheng, et al. 2024. "A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions." *ACM Computing Surveys* 57: 1–38.