

## RESEARCH ARTICLE

# Multiple Imputation for Robust Cluster Analysis to Address Missingness in Medical Data

ARNOLD A. HARDER<sup>1</sup>, GAYLA R. OLBRICHT<sup>1,2</sup>, (Member, IEEE), GODWIN EKUMA<sup>3</sup>, DANIEL B. HIER<sup>1,2</sup>, (Senior Member, IEEE), AND TAYO OBAFEMI-AJAYI<sup>1,2,4</sup>, (Member, IEEE)

<sup>1</sup>Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, USA

<sup>2</sup>Applied Computational Intelligence Laboratory, Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA

<sup>3</sup>Department of Computer Science, Missouri State University, Springfield, MO 65897, USA

<sup>4</sup>Engineering Program, Missouri State University, Springfield, MO 65897, USA

Corresponding author: Tayo Obafemi-Ajayi (tayoobafemijayi@missouristate.edu)


**ABSTRACT** Cluster analysis has been applied to a wide range of problems as an exploratory tool to enhance knowledge discovery. Clustering aids disease subtyping, i.e. identifying homogeneous patient subgroups, in medical data. Missing data is a common problem in medical research and could bias clustering results if not properly handled. Yet, multiple imputation has been under-utilized to address missingness, when clustering medical data. Its limited integration in clustering of medical data, despite the known advantages and benefits of multiple imputation, could be attributed to many factors. This includes methodological complexity, difficulties in pooling results to obtain a consensus clustering, uncertainty regarding quality metrics, and a lack of accepted pipelines. A few studies have examined the feasibility of implementing multiple imputation for cluster analysis on simulated/small datasets. While these studies have begun to address how to pool imputed values and quantify uncertainty in clustering due to imputation, a need remains for a complete framework that integrates MI in the clustering of complex medical data and sophisticated cluster algorithms. We propose a cluster analysis framework that mitigates bias and addresses these limitations. It includes methods to pool multiple imputed datasets, create a consensus cluster solution by ensemble methods, and select an optimal number of clusters based on validity indices. It also estimates uncertainty about cluster membership attributable to the imputation and identifies features that characterize the derived clusters. The utility of this framework is illustrated by its application to a traumatic brain injury dataset with missing data. Our analysis revealed six multifaceted clusters that differed with respect to Glasgow Coma Score (GCS), mechanism of injury, sociodemographics, vitals, lab values, and radiological presentation. The most severe cluster consisted of single, relatively young patients injured by motor accident, with higher GCS severity scores. Comparative analysis with the `miclust` R package, along with statistical validation of cluster characterization, demonstrates its robust performance.

**INDEX TERMS** Multiple data imputation, clustering, ensemble learning, canonical discriminant analysis, mixture models, traumatic brain injury, missingness.

## I. INTRODUCTION

The identification of meaningful subgroups of patients by unsupervised machine learning (ML) methods is a key component of the precision medicine initiative [1], [2]. Various diseases (cancers, neurological disorders, genetic

disorders, autoimmune disorders, etc.) exhibit high heterogeneity in their clinical presentation, trajectory, and outcome. Identifying clinically meaningful patient subgroups enables individualized care. The explosion in the variety and volume of medical data has created opportunities for data-driven ML methods to discover disease subtypes with implications for precision medicine [1], [3]. Clustering, an unsupervised ML technique, have been successfully applied to varied disease

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin .

datasets to yield important insights that could guide treatment or prognosis [4], [5], [6], [7]. Nevertheless, medical data with its complex structure, disparate data types, and missingness poses challenges to the application of cluster analysis to precision medicine. Missing data is a pervasive problem in healthcare [8]. Even carefully conducted prospective studies often have missing data on baseline characteristics or outcome measures [9]. While the absence of some recorded data may be unavoidable in clinical studies, the performance of ML algorithms suffers from bias when data is incomplete [10]. Furthermore, missingness poses a significant challenge when clustering medical data since these algorithms require a complete matrix of input features.

Various methods have been proposed to handle missingness [11], [12], [13]. The most common approach is complete-case analysis (CCA) in which any individual with missing data on any of the predictor or outcome variables is deleted from the analysis [8], [9], [13]. A variation on this method is to delete features that exceed a specified threshold of missingness. Although this approach creates a fully observed dataset, it is problematic in that it decreases either the number of cases or the number of features available, potentially resulting in the loss of pertinent information. Furthermore, when data are missing on multiple features, a substantial proportion of the sample may be excluded from analysis, leading to a loss of precision and statistical power [14], [15]. Case deletion can introduce selection bias if there is a systematic difference between patients with and without missing values [9].

An alternative to CCA is single imputation which replaces each missing value with a single value. Let  $\mathcal{D}$  denotes an input dataset consisting of  $p$  features for  $n$  samples with missing values i.e.  $\mathcal{D} = \{\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_p\}$  where  $|\mathbf{F}_j| = n$ , a simple and commonly used single imputation technique is to replace each missing data point in  $\mathcal{D}$  with its mean for normally distributed continuous features, its median for non-normal continuous features, and its mode for categorical features [9]. This approach is also problematic as it can lead to biased parameter estimates, an altered distribution of the features, or a disruption in the relationship among the features [12]. Other advanced methods have been proposed within the conditional estimation framework for single imputation [13], [16]. For example, conditional estimates of missing values based on values of other observed features in the data (e.g. by regression) can yield unbiased estimates under certain assumptions [9]. Another advanced approach to single imputation is the *IterativeImputer* method [16] (available in Scikit-learn package), which models features with missing values as a function of other features and uses the predicted estimates as imputed value. The *IterativeImputer* repeats this procedure up to a default maximum number of iterations and uses the average of the predicted values as the final prediction. The main disadvantage of single imputation is that the uncertainty of the imputed values is not incorporated in the final analysis [12], [14]. The treatment of estimates of

missing data values as though they are precise, rather than hypothetical estimates, ignores the uncertainty in the imputed values and may underestimate standard errors or overestimate the significance of model results [9], [12], [14].

Multiple imputation (MI) is recognized as a robust approach to missingness that accounts for uncertainty in the imputed values, addresses the issue of underestimated standard errors, and yields less biased estimates of missing values [12], [13], [14], [15]. MI methods have been used in predictive analysis (or supervised learning models) which entails building statistical models that estimate an outcome from a set of inputted features [12]. The three main steps of MI are imputation, analysis, and pooling [16]. MI replaces missing values in the dataset  $m$  times with plausible data points, resulting in  $m$  imputed datasets  $\tilde{\mathcal{D}}^{(1)} \dots \tilde{\mathcal{D}}^{(m)}$ . These values are drawn from a distribution specifically modeled for each missing entry. The analysis is conducted on each  $\tilde{\mathcal{D}}^{(h)}$  (where  $h \in \{1 \dots m\}$ ), and the results are pooled to obtain a final imputation that accounts for variation among the estimated values [9], [14]. Despite the importance of handling missing data appropriately and the demonstrated value of MI in prediction models [14], [17], [18], [19], MI has not been widely applied to unsupervised ML models that are tackling large complex datasets such as occur in healthcare [8].

A few studies have examined the feasibility of implementing MI for cluster analysis on simulated or small datasets [11], [20], [21], [22], [23]. Clustering is a multidimensional optimization problem and is considered exploratory data analysis since cases are generally unlabeled, unlike predictive models, which have a known outcome label. This poses a key challenge for MI as clustering does not require the estimation of specific model parameters, which makes the pooling step less clear. Prior feasibility studies [11], [20], [21], [22], [23] have emphasized that when MI is applied to clustering, the pooling of multiple imputations is challenging. Another problem is to quantify the uncertainty in clustering that occurs due to variance in imputed values. While these studies have begun to address how to pool imputed values and how to quantify uncertainty in clustering due to imputation, a need remains for a complete framework that integrates MI in the clustering of medical and healthcare data, which is typically more complex than the data considered in some of these earlier studies. It is important to address MI within the context of key data preprocessing and data curation steps, as well as in the context of more sophisticated clustering methods beyond  $k$ -means. Furthermore, the downstream effects of MI after clustering need to be assessed. After clustering, influential (discriminating) features for cluster formation need to be identified. The clusters formed must be characterized and interpreted. Post-cluster analysis is critical to making clustering useful for precision medicine. Single imputation methods largely influences data preprocessing prior to clustering, however multiple imputation impacts on all three stages of the clustering (data preprocessing,

clustering, and post-clustering analysis) (Fig. 1). This work focuses on a careful examination of all steps especially the post-clustering, in contrast to prior feasibility studies.

Our goal is to provide an explainable framework for the application of MI to the clustering of complex medical data that will support fairness and unbiased in the identification of disease subtypes, even in the presence of missing data [24]. We utilize a rigorous clustering method, ensemble clustering (also known as consensus clustering), which is an effective means to aggregate a collection of dissimilar clusterings to yield a more robust solution [25], [26]. It is particularly useful when dealing with complex datasets where different individual algorithms may excel in capturing different patterns or structures. The contributions of the work are as follows:

- We propose a framework that combines MI with ensemble clustering so as to ensure robustness at each stage, from data curation, to clustering, to cluster validation, to downstream statistical analysis, and finally to cluster characterization.
- Ensure robustness of cluster analysis by accounting for uncertainty due to the missingness.
- Empirically evaluate this framework in the context of traumatic brain injury (TBI), a common neurological disorder that poses a substantial public health burden [24].

## II. BACKGROUND

To provide a context for the proposed clustering framework, we briefly review basic underlying concepts and terminologies related to MI, clustering, and validation metrics. Each of the three main steps for MI (imputation, analysis, and pooling) are described in context of clustering.

### A. MECHANISMS OF MISSINGNESS

Prior to applying any imputation technique, it is important to consider the underlying mechanisms driving the missingness. The pattern and mechanism of missing data are important determinants as to which technique is utilized [27]. There are broadly three main mechanisms discussed in literature: missing completely at random (MCAR), Missing at Random (MAR), and missing not at random (MNAR) [12], [13], [14]. MCAR occurs when the probability of a data point being missing for a feature does not depend on values of the observed or unobserved data. Under MCAR, there are no systematic differences between missing and observed values. MAR implies the probability of a data point being missing can be explained by information contained in the observed data. MAR is a more realistic and broader assumption that also includes MCAR situations. MNAR assumes the probability of a data point being missing depends on information not available in observed data. MNAR data are the most challenging to address and require more complex methods as the systematic differences between the missing

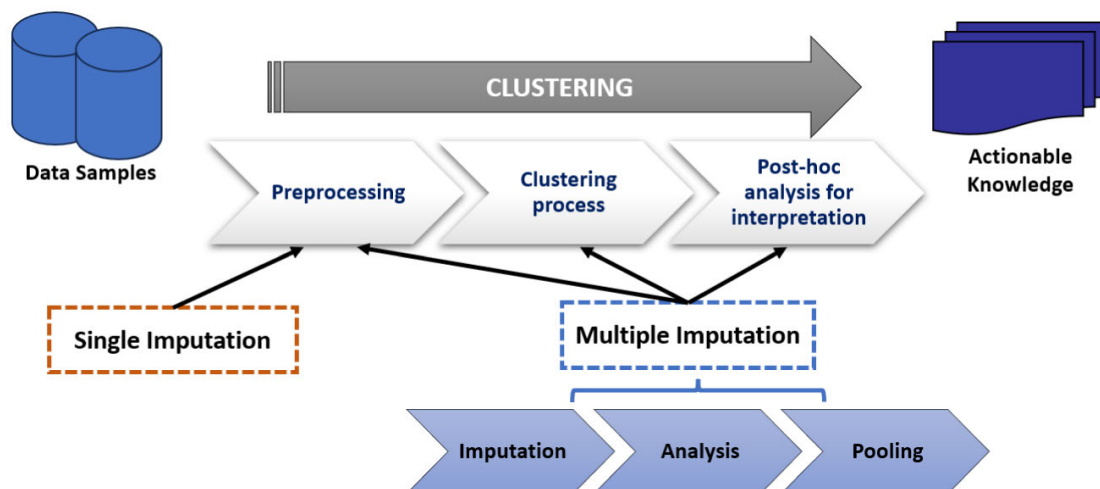
and observed values still remain even when the observed values have been taken into account [12], [14].

Although MAR is more realistic than MCAR for most datasets, it is challenging to know when the MNAR assumption is more appropriate since there are no statistical tests to determine whether the data are MAR or MNAR [12]. In predictive models, a sensitivity analysis [28] can be performed to assess how deviations from this assumption affect the results. However, to the best of our knowledge, no such methods are available for cluster analysis and this is an open problem. It is standard practice when using MI to assume MAR, as we do in this work.

### B. MULTIPLE IMPUTATION METHODS

When conducting MI, a key challenge for any type of downstream analysis is selecting an appropriate imputation method. Such method should i) account for the process that created the missingness; (ii) preserve the relationship among features; and (iii) account for the uncertainty about these relationships [16]. A widely used approach that addresses these requirements is multivariate imputation by chained equations (MICE), which utilizes a fully conditional specification framework (FCS). In FCS, a multivariate imputation model is specified for each incomplete feature through a set of conditional densities. Missing data are imputed one feature at a time by iterating over the conditional densities [16].

There are different definitions of multivariate imputation exist in the literature [13], [29]. The terms univariate and multivariate refer to the number of features in an analysis. Univariate implies one feature while multivariate, more than one feature. This terminology applies to imputation in two ways. It can refer to the number of features with missing values. It can also describe the type of model (or method) applied to generate the imputed values. A univariate imputation method, such as replacement by the mean, estimates the missing data points for a feature  $F_j \in \mathcal{D}$  by using observed values available for a single feature. In contrast, a multivariate imputation model, such as predictive mean matching [12], generates missing values for the feature  $F_j$  using observed values for  $F_j$  as well as other features in  $\mathcal{D}$ . Note that multivariate imputation is not the same as multiple imputation. Likewise, univariate imputation is not the same as single imputation. Multivariate imputation can be applied in a single imputation context where a single estimate is obtained per missing data point, resulting in a single imputed dataset  $\tilde{\mathcal{D}}$ . When applying multivariate imputation in the setting of multiple imputation, multiple values are estimated for each missing data point, resulting in multiple imputed datasets  $[\tilde{\mathcal{D}}^{(1)} \dots \tilde{\mathcal{D}}^{(m)}]$ . The number of imputed datasets ( $m$ ) required depends on factors such as the percentage of missingness and analysis goals. It should be carefully chosen to attain a low and stable between-imputation variance [30]. According to [12], a range of  $m = [5, 20]$  is sufficient for point estimation for a moderate amount of missingness. Using  $m > 20$  does not usually provide enough of a relative improvement



**FIGURE 1.** Challenge of integrating the 3 key components of multiple imputation into cluster analysis in contrast to single imputation.

in accuracy compared to increased computational cost. In this work,  $m$  is set to 15.

Fully conditional specification could be used to conduct multiple imputation, as it specifies a multivariate imputation model on a feature-by-feature basis by a set of conditional densities, one for each incomplete feature [16]. In selecting an appropriate multivariate imputation method, it is important to take into account the feature datatype (e.g., continuous, categorical, ordinal) in  $\mathcal{D}$ , as well as the nature of relationship among the features. There are some methods such as EMII algorithm [31] suited for specific data applications such as gene expression. Predictive mean matching (PMM) and random forest (RF) are commonly used multivariate imputation methods suited for mixed datatypes (a characteristic of complex medical data). PMM is an imputation method that preserves non-linear relationships between features [16]. RF is a ML-based approach for obtaining imputed values based on sampling values and combining predictions derived from multiple regression trees [32]. These approaches keep imputed values within biologically reasonable intervals. RF performs favorably against classic imputation methods, in that it is a better general function approximator and it allows for non-linearities and more complex interdependencies [33], [34].

When multivariate imputation is combined with multiple imputation, it is helpful include as many related features as are available in the dataset. Some features may be useful for the imputation model, even if these features are not used for the cluster analysis. Using every bit of available information yields multiple imputations that have minimal bias and maximal certainty [16].

### C. MI ANALYSIS STEP FOR CLUSTERING

A key aspect of MI is that the analysis is conducted on each of the  $m$  imputed datasets, and intermediate results are

pooled together to yield a final result [16]. This analysis step has historically involved predictive modeling with a known outcome where parameter inference is a key goal. However, cluster analysis does not lie within this modeling scope since it is an unsupervised (exploratory) method [11]. The goal of cluster analysis, especially in context of medical data, involves more than simply estimating the number of clusters (subgroups) using a specified clustering algorithm. It also includes determining final cluster assignment of each individual in the dataset as well as a holistic characterization of the subgroups in terms of membership, discriminating features (or biomarkers), statistical validation, and clinical relevance/explainability. These highlight several steps in the clustering framework that need to be addressed in terms of pooling information and evaluation of uncertainty across the multiple imputations. Although these steps are straightforward to implement with a single dataset, when dealing with  $m$  imputed datasets, the analysis step has to ensure that the variation among estimates of the missing values is addressed.

#### 1) DATA PRE-PROCESSING

Data pre-processing (or data curation) ensures that the data utilized for clustering is of optimal quality. Key data curation tasks include addressing missing data, eliminating redundant features, removing outliers, and feature normalization [35]. When incorporating MI, correlation and/or outlier detection techniques [35] are conducted on each of the  $m$  datasets  $[\tilde{\mathcal{D}}^{(1)} \dots \tilde{\mathcal{D}}^{(m)}]$ . For MI to work properly, the only difference in the datasets should be in the imputed values. All features, samples, and complete data points should remain the same across datasets. Correlation analysis can lead to removal of a feature from  $\tilde{\mathcal{D}}^{(h)}$  while outlier detection can lead to the removal of a sample. For the set of features and samples to be the same across all datasets, a redundant feature (or outlier



sample) should be eliminated only if it is flagged in at least  $p\%$  of the  $m$  datasets, where  $p\%$  is determined by the domain application.

To ensure proper normalization of all data points in each  $\tilde{\mathcal{D}}^{(h)}$ , it has to be conducted globally across all  $m$  datasets for each feature. Specifically for continuous and ordinal features, the minimum and maximum values for determining the normalized value between 0 and 1 for each instance are selected per feature across all  $m$  datasets. This standardization  $f_s$ , ensures that the variation across datasets is in the imputed values, not in the observed values. Thus  $f_s([\tilde{\mathcal{D}}^{(1)} \dots \tilde{\mathcal{D}}^{(m)}]) \rightarrow [\tilde{\mathcal{D}}^{(1)} \dots \tilde{\mathcal{D}}^{(m)}]$ .

## 2) ENSEMBLE CLUSTERING ALGORITHM AND VALIDATION METRICS

In clustering, different sets of clusters  $C_l = \{c_1, \dots, c_k\}$  can be obtained by multiple clustering algorithms or by varying a given parameter for the same algorithm [36]. Ensemble clustering (or consensus clustering) is an advanced clustering approach that aggregates a collection of dissimilar clusterings to yield a more robust solution [25], [26]. It allows us to leverage the strength of diverse algorithms, as they can vary significantly in performance and outcome. A key benefit is the increased stability of clusters and the greater likelihood of obtaining clustering solutions with low sensitivity to noise, outliers, or sampling variation.

The premise of ensemble clustering is to apply a set of clustering algorithms to an input data matrix such that each algorithm yields its own clustering result  $C_l$ . These results are pooled using a consensus decision metric to determine the final clustering solution  $C$ . Consensus decision metrics allow for the fusion of the output partitions obtained from various clustering algorithms in the ensemble into a final partition. Multiple metrics are available that include the mixture model, graph closure, and majority voting [25].

When applying ensemble clustering, multiple solutions can be obtained depending on the consensus decision metric or the tuning parameters. How does one identify the optimal solution that translates to a ‘meaningful’ configuration for a given domain application [36]? Cluster validation indices estimate how well a given clustering configuration aligns with the structure of the underlying data [37]. The validation process is key to cluster analysis [38], given that multiple configurations are obtained by varying parameters and/or performing multiple iterations. Several validation indices [39] are available for use. Each index views the task of determining the optimal clustering configuration from a different perspective. To leverage the strengths of multiple indices, we utilize an ensemble validation model [6], which provides ranked order of clustering partitions so that the user can select the most optimal  $r$  configurations or results.

When applying MI to clustering, it is important to pool the results obtained per imputed dataset so that final clustering solution  $C = \{c_1, \dots, c_k\}$  aligns across all  $m$  datasets.

In the last phase of MI (see section II-D), we outline steps for pooling and aligning the final cluster partitions and membership across the imputed datasets.

## 3) CLUSTER CHARACTERIZATION

A key aspect of using clustering in precision medicine is to provide an explanation for the results [40]. An initial explanatory step is to characterize the clusters by discriminating features and to quantify feature differences between clusters. Integration of statistical analysis at every level of the model interpretation allows for the quantification of the differences between clusters and ensures that a robust description of the clinical characteristics of each cluster is made. Clinical relevance, also referred to as *usefulness* [6], [40]), addresses the question “Do the derived clusters have clinically relevant predictive power?” For example, do the “severe” cases have a worse outcome or leave the injured brain more susceptible to future damage or progression? To assess clinical relevance, varied outcome measures (separate from the input features) that evaluate prognosis and/or recovery trajectories are usually selected by domain experts. The key is to select ones pertinent to evaluating the clinical research objectives.

Both clinical interpretation and usefulness are important components of explainability of cluster analysis. In the context of MI, only the interpretation phase is directly impacted since it involves analyzing the input features, which need to be summarized across the  $m$  datasets. In contrast, evaluation of clinical relevance relies on the final cluster membership and involves analysis of data obtained from varied outcome measures outside of  $\mathcal{D}$ . Interpretation focuses on which features are important in separating the clusters and understanding how they differ statistically. There are multiple approaches to identifying these features [35]. One option is to perform canonical discriminant analysis (CDA) which evaluates the discriminative power of multiple linear combinations of the features and captures interactions [41]. SHAP (SHapley Additive exPlanation) values [42] can also be used to quantify the importance and influence of each feature  $\mathbf{F}_j \in \mathcal{D}$ , with respect to cluster membership, for further interpretation and explainability.

The above methods identify which features are useful in classifying an individual into one of the proposed clusters. However, they do not inform on how clusters differ by feature. To better understand the nature of differences among the clusters, statistical analyses are performed to identify which clusters differ significantly on each input feature. This can aid clinicians interpret the nature of the clusters based on the input features. The appropriate statistical methods to assess and quantify these pairwise differences depend on the datatype of each feature. Multiple testing corrections are required to control the false positive rate when many tests that are conducted. When conducting MI, feature importance evaluation and pairwise testing needs to be performed for

each of the  $m$  datasets. These results are pooled together as described in section II-D.

#### D. POOLING RESULTS

The last phase of MI involves pooling information from the  $m$  analyses conducted to obtain a final result along with an assessment of the uncertainty due to variance in the imputed values. In predictive models, the standard is usually to apply Rubin's rules [12], [43] for pooling to obtain parameter estimates with standard errors that account for variation between and within imputed datasets. When applying MI to clustering, the pooling process is required at two stages: (1) to obtain a final cluster membership and (2) post-hoc analysis of features after the cluster membership has been determined. Since clustering is an unsupervised learning process with no *ground truth label*, the pooling step is no longer a direct application of Rubin's rules [11].

To pool the  $m$  different clustering results  $\{C_1, \dots, C_m\}$  together to obtain a final partition  $C$ , we leverage a similar solution utilized in consensus clustering, a known approach to aggregate a collection of dissimilar clusterings (see Section II-C2). The final consensus partition obtained is highly dependent on the consensus decision metric applied, which is non-trivial. Prior feasibility studies on MI in cluster analysis have utilized majority voting [20] or non-negative matrix factorization methods [11]. In this work, we use the same consensus decision metric applied in the ensemble clustering model for obtaining the initial  $m$  clustering results. Thus, to assign the final cluster membership per sample across the  $m$  imputed datasets, each  $C_h$  result is pooled together to yield  $C$  using this rigorous consensus decision metric.

When integrating the MI pooling step after a final clustering partition is obtained, the main challenge is that post-hoc analysis operates on both the set of imputed datasets  $[\tilde{D}^{(1)} \dots \tilde{D}^{(m)}]$ , and the single final clustering result  $C$ . For the methods used to determine feature importance (such as SHAP and CDA), the  $m$  set of results obtained from each type of analysis can be pooled in a similar manner to the correlation analysis (see Section II-C1). A feature is considered important if it is flagged as a discriminating feature in at least  $p\%$  of the  $m$  datasets. In each of these analyses, a value is generated per  $F_j$  that suggests whether it is important in distinguishing between clusters or not. For example, CDA produces a loading per  $F_j$  on multiple canonical variables. Pooled within-group correlations ( $-1 \leq \rho \leq 1$ ) are then calculated between the features and the standardized canonical scores. An absolute value closer to 1 indicates a stronger relationship which implies the feature is more important in distinguishing the clusters. Any feature in which the correlation ( $\rho$ ) exceeds a certain threshold  $t$  on any canonical variable is deemed important.

Summarizing these values across the  $m$  datasets to obtain a global one is non-trivial. There are currently no known methods that offer a way to appropriately combine the

loadings or the pooled within-group correlations. Taking the mean of the values across datasets for CDA is not a viable approach since the canonical variables may not be optimally aligned (e.g., the signs of the loadings may differ). Generating a global value across the  $m$  imputed datasets remains an open problem. In this work, we generate a binary value for each feature within each of the  $m$  datasets by comparing the values to the threshold  $t$  and labeling it as important (1) or not (0). Results are then pooled across datasets by declaring a feature important if it is flagged in at least  $p\%$  of the datasets.

The statistical analyses to identify differences on input features between clusters in the final result  $C$  are one aspect where Rubin's rules for predictive analysis can be adapted for pooling. In predictive models, there is a quantity of interest ( $Q$ ) that is being estimated from the data. When applying MI, an estimate for  $Q$  is obtained for each dataset to yield  $m$  estimates  $\{\hat{Q}_1, \dots, \hat{Q}_m\}$ . These estimates are pooled together using Rubin's rules [12], [43] to obtain a final estimate ( $\tilde{Q}$ ) and its standard error. The standard error calculation incorporates the uncertainty in the estimate of  $Q$  that arises due to the variation in the imputed values across the  $m$  datasets. These rules can be applied to the statistical validation tests for post-hoc cluster analysis, which assess whether there is an association between the labels from the final clustering result  $C$  (predictor variable) and each of the features  $F_j$  utilized in the clustering (response variable). After the analyses for each imputed dataset  $\tilde{D}^{(h)}$  is conducted, the final estimate is given by  $\tilde{Q} = \frac{1}{m} \sum_{h=1}^m \hat{Q}_h$ .

The total variance of  $\tilde{Q}$ , denoted by  $\tilde{T}$ , is a combination of the within-imputation variance ( $\tilde{U}$ ) and between-imputation variance ( $\tilde{B}$ ).  $\tilde{U}$  is computed as,  $\tilde{U} = \frac{1}{m} \sum_{h=1}^m \hat{U}_h$ , where  $\hat{U}_h$  is the variance for each  $\tilde{D}^{(h)}$ , and  $\tilde{B}$  is given by,

$$\tilde{B} = \frac{1}{(m-1)} \sum_{h=1}^m (\hat{Q}_h - \tilde{Q})^2. \quad (1)$$

Hence, the total variance ( $\tilde{T}$ ) is derived as:

$$\tilde{T} = \tilde{U} + (1 + \frac{1}{m})\tilde{B}. \quad (2)$$

This calculation of total variance used in the post-hoc testing provides a measure of uncertainty that accounts for variation associated with missingness.

### III. METHODOLOGY

We propose a framework for integrating MI into cluster analysis that addresses the varied layers of complexity identified in Section II. This framework consists of three phases: input data preprocessing, ensemble clustering and validation, and cluster characterization and uncertainty analysis, as illustrated in Fig. 2. Each phase is described briefly below.

In the data preprocessing phase, we initially selected features to be employed for clustering (in consultation with a domain expert) and evaluated the missingness. Multiple imputation is then performed in the mice 2.9 package

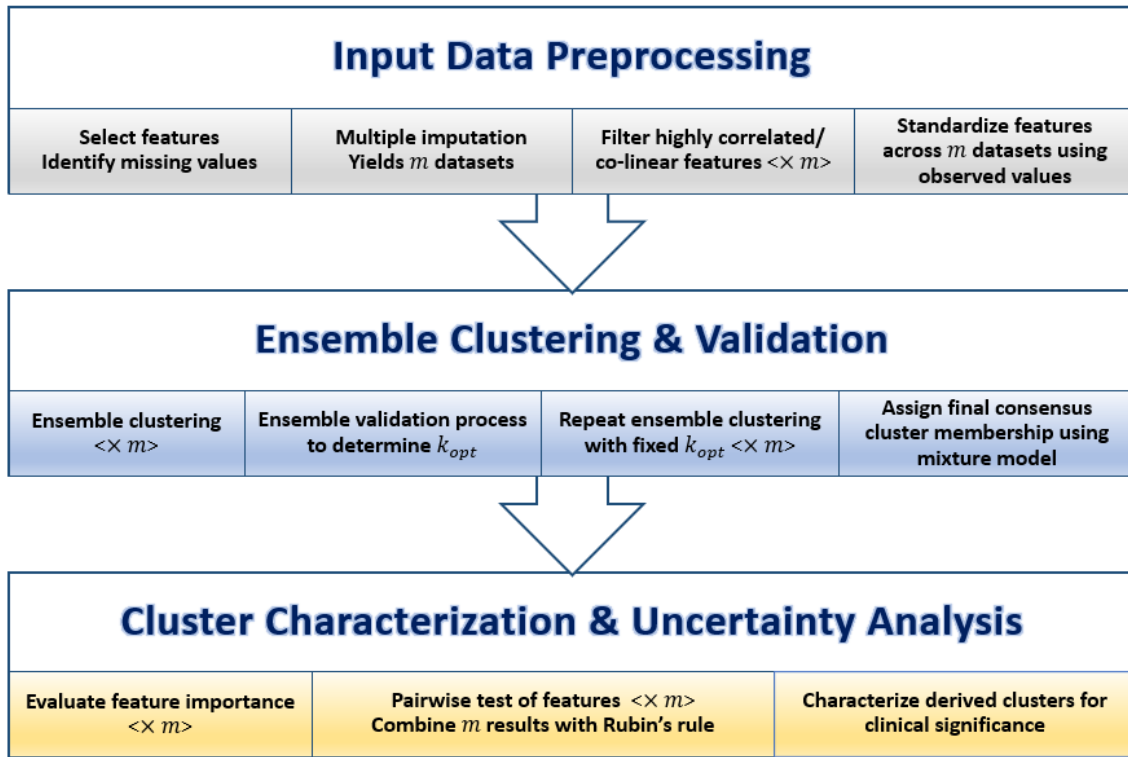


FIGURE 2. Overview of multiple imputation framework for cluster analysis.

available in R [16], assuming a MAR missing mechanism, to yield  $m = 15$  datasets. For implementation purposes with the mice package, PMM was utilized for features with only one missing observation and RF was used for features with more than one missing observation. For each of the  $m$  datasets, we conducted a pairwise correlation analysis using both Pearson and Spearman rank correlations. Features that exceeded the domain guided threshold [6] in at least  $p = 80\%$  of the datasets were dropped. The data was standardized, as described in Section II-C1, using only the observed values per feature across all  $m$  datasets cohesively.

The input data preprocessing phase yielded  $m$  standardized data matrices  $[\tilde{D}^{(1)} \dots \tilde{D}^{(m)}]$  for subsequent cluster analysis. The ensemble cluster analysis, as outlined in Algorithm 1, produced a final clustering solution ( $C_{k_{opt}}$ ) with  $k_{opt}$  clusters such that the final cluster assignment for each sample  $i$  is same across all the  $[\tilde{D}^{(1)} \dots \tilde{D}^{(m)}]$  matrices. The ensemble clustering model,  $\Pi$ , consists of  $v$  base algorithms with the mixture model consensus decision function,  $\Psi$ , where  $v$  is the number of individual clustering algorithms in the model. In this work,  $v=4$  ( $k$ -means, spectral, Gaussian mixture, and agglomerative clustering with Ward's linkage), and  $k$  is varied from  $a=2$  to  $b=10$ . (Note that for the mixture model consensus metric,  $k$  has to be specified *a priori*.) The ensemble model is based on previous work (E2 model/MM consensus) [6]. Since non-deterministic algorithms were applied, the seed was fixed to ensure that the difference in results across

datasets is due to the variation in the imputed values, not the random aspect of the algorithm.

The final clustering solution  $C_{k_{opt}}$  is obtained in an iterative manner (see Algorithm 1). The ensemble cluster validation model combines 7 different cluster performance indices [38] to rank the initial set of clustering results obtained per  $\tilde{D}^{(h)}$ . The global optimal value for the number of clusters,  $k_{opt}$ , is the most frequently occurring  $k$  among the top  $r = 3$  optimal clustering configurations across the  $m$  datasets. Using this value of  $k_{opt}$  for  $\Psi$ , another iteration of ensemble clustering is performed per  $\tilde{D}^{(h)}$ . During the second rerun, there is no need for the cluster validation model as only one clustering solution is generated per  $\tilde{D}^{(h)}$  since  $k_{opt}$  is fixed. However, the cluster assignment of each sample  $i$  can vary across the data matrices. To ensure a common cluster assignment per sample  $i$  across all data matrices, the  $m$  clustering configurations  $[C_{1,k_{opt}}, \dots, C_{m,k_{opt}}]$  are pooled together using the mixture model consensus decision metric to obtain the final configuration  $C_{k_{opt}}$ .

For the cluster characterization and uncertainty analysis phase, we applied both CDA and SHAP to evaluate the importance of the input features and identify the ones that contribute the most to the determination of cluster membership assignment. To identify the informative features, we pooled the results obtained across the  $m$  imputed datasets by marking a feature as important if it was flagged in at least  $p = 80\%$  of the datasets.

**Algorithm 1** Ensemble Clustering With Multiple Imputation

---

▷ Notations  
 $[\tilde{D}^{(1)} \dots \tilde{D}^{(m)}]$ :  $m$  standardized imputed datasets.  
 $\Pi$ : ensemble clustering model made up of  $\nu$  base algorithms.  
 $\Psi_{[a:b]}$ : mixture model consensus function with  $k$  varied from a minimum value ( $a$ ) to maximum value ( $b$ ).  
 $freq(k, r)$ : frequency of number of clusters  $k$  among top  $r$  ranked clustering results.

▷ Initial ensemble clustering runs  
**for**  $h := 1$  to  $m$  **do**  
  Run  $\Pi$  on  $\tilde{D}^{(h)} \rightarrow \tau$  base clustering solutions  
  Apply  $\Psi_{[a:b]}$  on  $\tau$  clusterings  $\rightarrow$   
   $[C_{h,a}, \dots, C_{h,k}, \dots, C_{h,b}]$  ensemble clustering outcomes  
  Rank  $[C_{h,a} \dots C_{h,b}]$  in order of most optimal solution using ensemble validation model  
**end**  
▷ Select optimal  $k$  value  
 $k_{opt} = \max_{k, \forall h} freq(k, r)$  on **ranked**  $[C_{h,a} \dots C_{h,b}]$

▷ Rerun ensemble clustering  
**for**  $h := 1$  to  $m$  **do**  
  Run  $\Pi$  on  $\tilde{D}^{(h)} \rightarrow \tau$  clustering solutions  
  Apply  $\Psi_{[k_{opt}]}$  on  $\tau$  base clustering solutions  $\rightarrow$   
   $[C_{h,k_{opt}}]$   
**end**  
▷ Pool results for final cluster assignment  
Apply  $\Psi_{[k_{opt}]}$  on  $[C_{1,k_{opt}}, \dots, C_{m,k_{opt}}] \rightarrow [C_{k_{opt}}]$

---

The CDA was implemented in JMP. A feature was flagged as discriminating if  $|\rho| > 0.20$  on any of the significant ( $p$ -value  $< 0.05$ ) canonical variables and this criteria is satisfied in at least 80% of the imputed datasets. Note that CDA has an underlying assumption that the features are continuous [44], [45], [46]. However, in practice, many datasets have a mixture of continuous and categorical variables such as in this work. We had four data types: continuous, ordinal, categorical, and binary. Ordinal features were treated as continuous in the CDA. For binary or categorical features, we utilized indicator variable encoding (i.e., for a binary feature, a “1” is assigned to one of the categories and a “0” for the other category). For a categorical feature with “ $l$ ” number of levels, a total of “ $l - 1$ ” indicator variables were used to represent the categories. This approach provided numerical values for the features and enabled the analysis to be conducted. Since the features were broken down into multiple variables, it did not provide

an overall loading for the feature and made interpretation more challenging. Further research is needed to determine the optimal way to handle categorical features in CDA.

SHAP values [42] were computed using the random forest model. Unlike CDA, SHAP does not have a threshold value for importance. Hence, the final feature importance values were averaged across the  $m$  imputed datasets for each  $F_j$ . With respect to categorical features (A4, A6), similar to CDA, we first applied one-hot encoding and then, averaged across all levels of encoding to obtain a final SHAP value.

The Tukey procedure based on ANOVA was utilized for continuous and ordinal scale features for pairwise tests, while a  $\chi^2$  test was conducted for nominal scale features. For each feature, a test was performed on each of the  $m$  datasets and results were pooled using Rubin’s rules to obtain the appropriate test statistic and p-value that accounts for both the within and between imputation variance, as described in Section II-D. Rubin’s rules were performed to properly incorporate the uncertainty that arises from the multiple imputed datasets into the testing procedures, as quantified by the standard errors. The Tukey method was implemented using the R packages *mitml* and *multcomp*, while the  $\chi^2$  test utilized the *miceadds* package [47], [48], [49]. A false discovery rate (FDR) adjustment was subsequently performed on the p-values obtained to control the expected proportion of false discoveries at 5% across all of the features tested.

The domain expert utilized the results of these analyses along with varied visualizations of the clusters (Uniform Manifold Approximation and Projection (UMAP) [50], CDA canonical plots, SHAPley bee-swarm plots [42], and heat maps) to derive a robust interpretation of the clinical significance of the cluster analysis.

#### IV. APPLICATION OF FRAMEWORK TO A TBI DATASET WITH MISSING DATA

To evaluate the effectiveness of the MI clustering framework, we applied it to a TBI sample drawn from the Citicoline Brain Injury Treatment Trial (COBRIT) [51] (which was used in prior work [6]). TBI has been identified as an ideal candidate for precision medicine data analysis, given its heterogeneity and complexity. Better tools are needed to characterize TBI severity subtypes beyond clinical classification systems such as the Glasgow Coma Scale (GCS) or cranial computer tomography (CT) metrics, such as Marshall or Rotterdam scores. For example, GCS classifies the severity of TBI as mild, moderate, or severe. However, it does not capture the pathoanatomical features or pathophysiology in individual patients and is confounded by factors such as endotracheal intubation, use of drugs, alcohol, and/or medications [24].

To demonstrate the robustness of our framework, we carried out comparisons on three fronts as follows. First, where a different multiple imputation framework (Miclust [20]) is applied to cluster analysis with a basic clustering algorithm. Miclust (package available in R) performs



$k$ -means clustering and uses a relabeling and voting process to combine the partitions obtained from the  $m$  imputed datasets. Second, the case where ensemble clustering is applied but without multiple imputation i.e. previous work by Yeboah et al. [6]. Third, a case in which cluster analysis is applied using partitioning around medoids (PAM) algorithm with a different imputation approach (see Section V) [52].

### A. DATA

The COBRIT study was a phase 3, double-blind, randomized clinical trial conducted over 4 years (2007 to 2011) to investigate the effectiveness of citicoline compared to placebo for TBI in 1213 subjects (ages 18 to 70 years) [51]. COBRIT data is available through the Federal Interagency Traumatic Brain Injury Research (FITBIR) [53] data repository. All participants had non-penetrating head injuries with varying levels of severity, as quantified by the GCS score, and required inpatient hospitalization. CT scan findings included intraparenchymal hemorrhages (10 mm or greater total diameter), acute extra-axial hematomas (epidural or subdural thickness of 5 mm or greater), subarachnoid hemorrhage (visible on at least 2 contiguous 5-mm slices or at least 3 contiguous 3-mm slices), intraventricular hemorrhage (present on 2 slices), and midline shift (5 mm or greater). Demographics and details on the injury were collected at baseline. Metabolic, liver, and hematologic values, vital signs, and other selected blood laboratory values were collected at multiple time points. The study found no benefit for functional or cognitive status in the active drug group.

The input features selected for cluster analysis in this paper are similar to those utilized in prior work [6]. Yeboah et al. [6] investigated a set of 32 features consisting of baseline measurements on demographics, details of injury, CT scan findings, metabolic, liver, and hematologic results obtained from blood samples, GCS scores, and vital signs. As a result of missing data, especially among the CT scan features, the sample size was reduced to 859 patients. Given the application of MI in this work to address missingness, we performed the cluster analysis on the full study sample of 1213 subjects and added the features of alcohol blood level, high and low diastolic blood pressure, and CT lesion high mixed density that had been excluded previously for missingness above 0.1%. Based on domain expert guidance, we excluded the hypertonic saline total volume as a feature not of clinical interest. For this study, we included the additional demographic features of gender and marital status. The correlation filter analysis identified a set of features that were excluded due to high correlation (collinearity). This included CT epidural lesion anatomic site, CT subdural lesion anatomic site, and CT intraparenchymal lesion anatomic site. Table 1 shows the set of 36 features, with the level of missingness, used in clustering the 1213 TBI patients.

**TABLE 1. Description of input features and evaluation of significance based on  $k6$  clustering output.**

Features	Tag	Data type	Missing	Important
<b>Pre-injury factors/demographic</b>				
Age	A1	Continuous	0.0%	Yes <sup>c,s</sup>
Weight	A2	Continuous	0.3%	
Sex	A3	Binary	0.0%	
Marital status	A4	Categorical	0.1%	Yes <sup>c,s</sup>
<b>Injury Related factors</b>				
Hours between injury & scan	A5	Continuous	0.0%	Yes <sup>s</sup>
Mechanism of Injury	A6	Categorical	0.0%	Yes <sup>c,s</sup>
GCS total	A7	Ordinal	0.0%	Yes <sup>c</sup>
Alcohol blood level	A8	Continuous	13.55%	
<b>Radiology/Imaging</b>				
CT Epidural lesion volume	A10	Ordinal	26.8%	
CT Midline shift category	A11	Ordinal	0.4%	Yes <sup>s</sup>
CT Subarachnoid hemorrhage type	A12	Ordinal	0.4%	Yes <sup>s</sup>
Intraventricular hemorrhage	A13	Binary	0.4%	Yes <sup>s</sup>
CT Subdural lesion volume	A14	Ordinal	26.8%	
CT intraparenchymal lesion volume	A16	Ordinal	26.8%	Yes <sup>s</sup>
CT Mesencephalic cisterns type	A18	Ordinal	0.3%	Yes <sup>s</sup>
CT Lesion high mixed density	A19	Binary	0.0%	
Hydrocephalus present	A20	Binary	0.3%	
<b>Clinical Factors</b>				
Heart rate (highest)	A21	Continuous	0.3%	Yes <sup>c</sup>
Heart rate (lowest)	A22	Continuous	0.3%	
Diastolic blood pressure (highest)	A23	Continuous	1.1%	
Diastolic blood pressure (lowest)	A24	Continuous	1.0%	
Systolic blood pressure (highest)	A25	Continuous	0.1%	
Systolic blood pressure (lowest)	A26	Continuous	0.1%	
Prothrombin INR (highest)	A27	Continuous	2.3%	
Prothrombin INR (lowest)	A28	Continuous	2.3%	Yes <sup>c</sup>
WBC count (highest)	A29	Continuous	0.5%	
<b>Laboratory Test Results</b>				
Hematocrit level (highest)	A30	Continuous	0.5%	
Hematocrit level (lowest)	A31	Continuous	0.5%	Yes <sup>c</sup>
Glucose level (highest)	A32	Continuous	0.6%	
Glucose level (lowest)	A33	Continuous	0.6%	
Sodium level (highest)	A34	Continuous	0.3%	
Sodium level (lowest)	A35	Continuous	0.3%	
Temperature (highest)	A36	Continuous	0.7%	
Oxygen Saturation (lowest)	A37	Continuous	0.5%	
Platelet count (lowest)	A38	Continuous	3.2%	
Hemoglobin count (lowest)	A39	Continuous	3.1%	

GCS: Glasgow Coma Scale; CT: Computed Tomography; INR: International Normalized Ratio. WBC: White blood cell; Important features are those identified, by either CDA (<sup>c</sup>) or SHAP (<sup>s</sup>), as discriminative in cluster membership assignment for the  $k6$  result.

### B. ENSEMBLE CLUSTERING & VALIDATION RESULTS

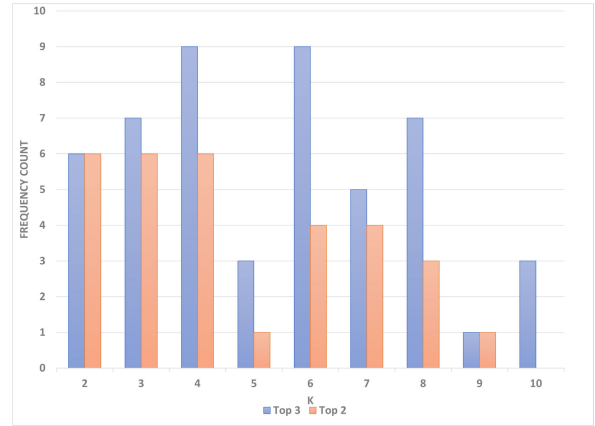
The outcome of initial clustering and validation for the 15 datasets is illustrated in Table 2. The value of  $k_{opt}$  (see Fig. 3) was determined to be either 4 or 6, based on the most frequent appearing value of  $k$  in the top 2 and top 3 highest ranked results. The next iteration of the ensemble clustering was conducted with  $k$  fixed at both 4 and 6 for  $k_{opt}$  for the mixture model consensus across all 15 datasets. Heat maps of normalized mutual information (NMI) (see Fig. 4) demonstrate the relative variation in the results among the 15 imputed datasets after using both fixed  $k_{opt}$  values for ensemble clustering. We used NMI to quantify the amount of uncertainty in cluster membership due to the data imputation since the differences across the 15 datasets is attributable to

**TABLE 2. Cluster validation model ranking outcome across all datasets.**

$k$ clusters	Weighted Score	Sil	Db	Xb	Dunn	CH	$\mathcal{I}$	SDbw
k3	27	5	2	5	5	5	5	0
k7	17	0	4	3	2	2	2	4
k6	16	2	0	2	4	4	3	1
k10	16	0	3	0	4	3	1	5
<hr/>								
k2	26	5	4	5	5	3	4	0
k6	18	3	1	0	4	5	2	3
k3	17	4	2	3	3	2	3	0
k8	17	0	5	4	2	1	0	5
<hr/>								
k2	22	5	5	5	0	2	5	0
k8	16	4	0	0	3	4	0	5
k4	13	3	1	3	0	3	3	0
k6	13	1	4	2	2	0	2	2
k7	13	0	2	4	4	1	1	1
<hr/>								
k4	27	4	5	5	3	5	5	0
k3	18	5	4	4	1	1	3	0
k8	17	0	2	0	5	4	1	5
<hr/>								
k4	25	4	5	3	3	5	5	0
k7	15	0	4	1	4	2	0	4
k5	14	2	3	4	0	1	3	1
<hr/>								
k2	25	5	5	5	0	5	5	0
k4	22	4	4	4	2	4	4	0
k10	14	0	3	0	5	1	0	5
<hr/>								
k2	23	5	4	5	0	4	5	0
k7	22	3	5	4	4	2	1	3
k6	13	0	0	0	5	3	3	2
<hr/>								
k2	25	5	5	5	0	5	5	0
k8	19	0	4	2	4	3	1	5
k4	18	2	2	3	3	4	4	0
<hr/>								
k3	30	5	5	5	5	5	5	0
k6	17	0	4	4	3	2	1	3
k9	17	1	2	2	4	3	0	5
k4	14	4	0	1	0	4	4	1
<hr/>								
k3	23	5	2	3	3	5	5	0
k7	20	0	5	4	5	2	1	3
k6	18	1	4	5	2	1	3	2
<hr/>								
k8	23	4	5	2	4	3	1	4
k3	21	5	3	3	0	5	5	0
k5	18	1	4	5	0	4	4	0
<hr/>								
k3	22	5	1	1	5	5	5	0
k4	19	2	4	5	0	4	4	0
k6	17	1	5	4	2	3	1	1
<hr/>								
k6	26	4	5	5	1	5	3	3
k4	24	5	2	4	5	4	4	0
k8	17	0	3	2	4	3	0	5
<hr/>								
k5	22	1	5	5	3	2	4	2
k4	21	5	1	0	4	5	5	1
k10	17	0	2	2	5	3	0	5
<hr/>								
k2	26	5	5	5	1	5	5	0
k6	26	4	4	3	3	4	4	4
k8	13	0	3	1	5	1	0	3

Sil: Silhouette; Db: Davies-Bouldin; Xb: Xie-Beni; CH: Calinski-Harabasz; S\_Dbw: Scatt + Dens\_bw.

the imputed values. As shown in Fig. 4, there is a significant degree of variation in cluster assignment across datasets based on NMI. The  $k6$  results appear to be more consistent. The final pooled result, obtained from the consensus cluster membership assignment using the mixture model, for both  $k_{opt}$  values is illustrated in Fig. 5 by UMAP on dataset #15. Even though the final pooled cluster labels are the same for all 15 datasets, since each dataset differs on the imputed values, one dataset (#15) is utilized to visualize all the results.



**FIGURE 3. Summary of the number of clusters  $k$  value in the highly ranked (top 2 vs. top 3) cluster validation results across all datasets. The most frequently occurring  $k$  results were  $k=4$  and 6.**

	set 1	set 2	set 3	set 4	set 5	set 6	set 7	set 8	set 9	set 10	set 11	set 12	set 13	set 14	set 15
set 1	1.00	0.83	0.51	0.73	0.67	0.84	0.76	0.65	0.43	0.62	0.51	0.72	0.70	0.53	0.66
set 2	0.83	1.00	0.56	0.72	0.67	0.81	0.61	0.67	0.52	0.61	0.57	0.72	0.71	0.65	0.63
set 3	0.51	0.56	1.00	0.57	0.64	0.47	0.65	0.56	0.45	0.54	0.50	0.57	0.49	0.55	0.60
set 4	0.73	0.72	0.57	1.00	0.81	0.74	0.76	0.66	0.55	0.73	0.63	0.95	0.62	0.64	0.57
set 5	0.67	0.67	0.64	0.81	1.00	0.65	0.80	0.74	0.69	0.64	0.78	0.80	0.60	0.80	0.55
set 6	0.84	0.81	0.47	0.74	0.65	1.00	0.61	0.56	0.40	0.65	0.46	0.73	0.72	0.52	0.63
set 7	0.76	0.61	0.65	0.76	0.80	0.61	1.00	0.81	0.57	0.62	0.64	0.76	0.52	0.65	0.62
set 8	0.65	0.67	0.56	0.66	0.74	0.56	0.81	1.00	0.62	0.56	0.65	0.65	0.57	0.83	0.56
set 9	0.43	0.52	0.45	0.55	0.69	0.40	0.57	0.62	1.00	0.60	0.73	0.55	0.48	0.63	0.46
set 10	0.62	0.61	0.54	0.73	0.64	0.65	0.62	0.56	0.60	1.00	0.78	0.73	0.45	0.60	0.58
set 11	0.51	0.57	0.50	0.63	0.78	0.46	0.64	0.65	0.73	0.78	1.00	0.64	0.40	0.71	0.44
set 12	0.72	0.72	0.57	0.95	0.80	0.73	0.76	0.65	0.55	0.73	0.64	1.00	0.60	0.64	0.57
set 13	0.70	0.71	0.49	0.62	0.60	0.72	0.52	0.57	0.48	0.45	0.40	0.60	1.00	0.46	0.53
set 14	0.53	0.65	0.55	0.64	0.80	0.52	0.65	0.83	0.63	0.60	0.71	0.64	0.46	1.00	0.53
set 15	0.66	0.63	0.60	0.57	0.55	0.63	0.62	0.56	0.46	0.58	0.44	0.57	0.53	0.53	1.00

(a)  $k$  is fixed at 4.

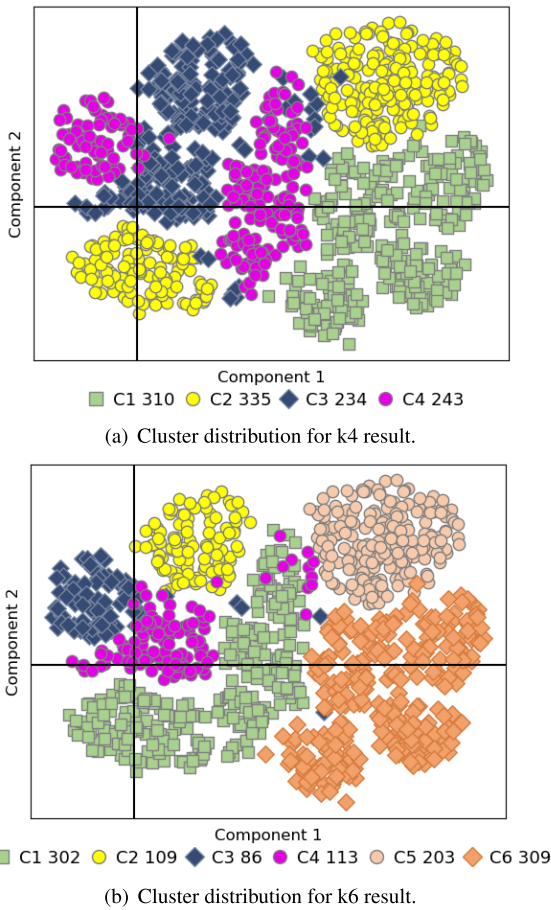
	set 1	set 2	set 3	set 4	set 5	set 6	set 7	set 8	set 9	set 10	set 11	set 12	set 13	set 14	set 15
set 1	1.0	0.79	0.83	0.78	0.76	0.70	0.79	0.77	0.78	0.69	0.77	0.67	0.76	0.79	0.77
set 2	0.79	1.0	0.78	0.76	0.75	0.79	0.77	0.71	0.87	0.63	0.89	0.72	0.78	0.75	0.84
set 3	0.83	0.78	1.0	0.82	0.83	0.73	0.85	0.84	0.86	0.75	0.86	0.77	0.83	0.86	0.87
set 4	0.78	0.76	0.82	1.0	0.79	0.73	0.79	0.78	0.78	0.67	0.77	0.66	0.75	0.88	0.77
set 5	0.76	0.75	0.83	0.79	1.0	0.74	0.80	0.84	0.82	0.67	0.81	0.78	0.78	0.80	0.80
set 6	0.70	0.79	0.73	0.73	0.74	1.0	0.73	0.66	0.79	0.74	0.79	0.67	0.75	0.72	0.79
set 7	0.79	0.77	0.85	0.79	0.80	0.73	1.0	0.78	0.81	0.70	0.82	0.70	0.78	0.80	0.80
set 8	0.77	0.71	0.84	0.78	0.84	0.66	0.78	1.0	0.78	0.73	0.78	0.83	0.78	0.83	0.83
set 9	0.78	0.87	0.86	0.78	0.82	0.79	0.81	0.78	1.0	0.69	0.94	0.80	0.85	0.82	0.92
set 10	0.69	0.63	0.75	0.67	0.67	0.74	0.70	0.73	0.69	1.0	0.68	0.66	0.72	0.71	0.69
set 11	0.77	0.89	0.86	0.77	0.81	0.79	0.82	0.78	0.94	0.68	1.0	0.79	0.87	0.82	0.92
set 12	0.67	0.72	0.77	0.66	0.78	0.67	0.70	0.83	0.80	0.66	0.79	1.0	0.77	0.72	0.82
set 13	0.76	0.78	0.83	0.75	0.78	0.75	0.78	0.78	0.85	0.72	0.87	0.77	1.0	0.81	0.86
set 14	0.79	0.75	0.86	0.88	0.80	0.72	0.80	0.83	0.82	0.71	0.82	0.72	0.81	1.0	0.83
set 15	0.77	0.84	0.87	0.77	0.80	0.79	0.80	0.83	0.92	0.69	0.92	0.82	0.86	0.83	1.00

(b)  $k$  is fixed at 6.

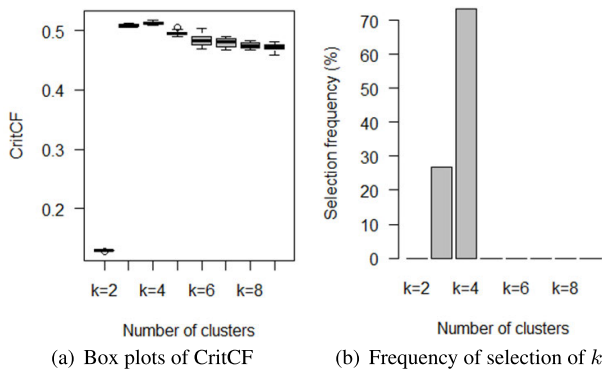
**FIGURE 4. Heatmap to illustrate variance in clustering assignment across imputed datasets using normalized mutual information (NMI) after fixing  $k$  at 4 and 6, prior to pooling across all for final cluster membership.**

**C. COMPARATIVE ANALYSIS USING MICLUST**

The outcome of the cluster analysis using Miclust is shown in Fig. 6. Based on the initial  $k$ -means results obtained per dataset, Miclust determines the optimal value of  $k$  using the CritCF criterion [54]. The CritCF [54] provides a ranking of partitions in feature subspaces of different cardinalities. It simultaneously searches for both relevant feature subspaces and optimal partitions. Higher values for CritCF are preferred. The box plots in Fig. 6(a) illustrate the distribution of CritCF values across the 15 imputed datasets for  $k$  varied from 2 to 9 (the maximum possible  $k$  in Miclust). The highest median CritCF value is observed in the  $k = 4$  results, which aligns with the results in Fig. 6(b) that show

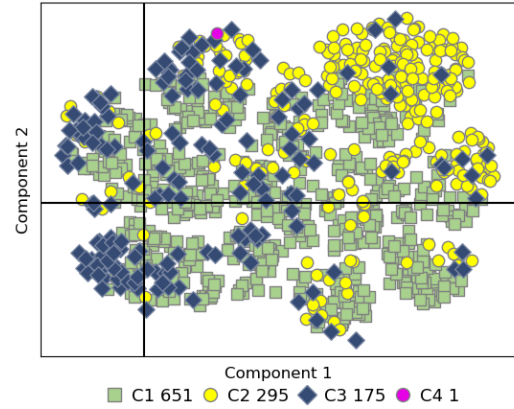


**FIGURE 5.** Visualization of clusters by UMAP after pooled membership across all datasets. Even though the final pooled cluster labels are same regardless of dataset, since each dataset differs on the imputed values, results are displayed for dataset #15.



**FIGURE 6.** MiClust results: (a) Show box plots of the between-imputation distribution of CritCF by the number of clusters ( $k$ ). From (b), the most frequently occurring optimal result based on the CritCF criterion was  $k=4$ .

$k = 4$  was selected in around 70% of the datasets. MiClust subsequently refits the clustering across all 15 datasets using  $k = 4$ , aligns the partitions, and finally assigns each sample a cluster membership based on majority voting. The final MiClust result is illustrated in Fig. 7 with UMAP on dataset #15. Note that one of the clusters is comprised of only one



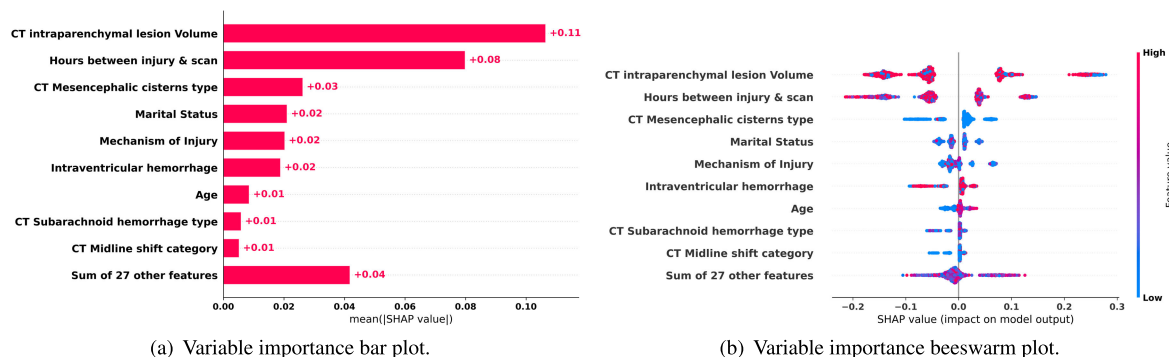
**FIGURE 7.** Visualization of clusters by UMAP for  $k4$  using MiClust package. Even though the final pooled cluster labels are same regardless of dataset, since each dataset differs on the imputed values, results are displayed for dataset #15.

individual, which could be an outlier. There is also much more overlap between the MiClust clusters compared to our results (Fig. 5).

#### D. OUTCOME OF CLUSTER CHARACTERIZATION & UNCERTAINTY ANALYSIS

The outcome of the feature importance evaluation, based on CDA, for the  $k4$  and  $k6$  clustering outputs as well as MiClust are presented in Table 3. All canonical variables were statistically significant at ( $p < 0.05$ ) across all 15 imputed datasets for all of three clustering results. CDA identified 4 discriminating features for the  $k4$  output (A1, A4, A6, and A21) while the  $k6$  results selects these in addition to three other non-demographic features (A7, A28, and A31). For the MiClust result, presence of the singleton cluster appears to inflate the contribution of the CDA correlations. Almost all features (27 out of 36) were deemed important in discriminating between the clusters. The domain expert (D.H) determined the optimal clustering result based on a holistic approach that took into account the cluster composition of each result by means and distributions, UMAP visualizations, CDA outcomes, SHAP plots, and NMI heatmaps. The  $k6$  cluster derivation was chosen as more clinically meaningful than the  $k4$ . All the subsequent results for cluster characterization and uncertainty analysis are presented for  $k_{opt} = 6$ .

Figure 8 displays the SHAP results obtained when a random forest model is fit on the  $k6$  clustering output using bar plots (Fig. 8(a)) and beeswarm plots (Fig. 8(b)). The height of each vertical bar in the bar plot indicates the magnitude of the feature’s importance while the beeswarm plot shows the SHAP value distribution. It denotes the importance by color. Red indicates a positive correlation, while blue indicates a negative correlation on the predicted value. The intensity of the color reflects the level of impact a specific feature has on the prediction. Both red and blue are important, but they have opposite effects on the output.



**FIGURE 8.** *k6* SHAP results using dataset #15: (a) shows the mean absolute value of the SHAP values for each feature. (b) shows the distribution of the impacts each feature has on the model output. The color represents the feature value (red high, blue low).

**TABLE 3.** Canonical discriminant analysis (CDA) results for top two optimal clustering outputs from our framework and Miclust.

Clustering Output	Cluster Sizes	Features important for distinguishing clusters
<i>k4</i>	C1: 310	Can 1: A1, A4, A6, Can 2: A4, A6 Can 3: A4, A6, A21
	C2: 335	
	C3: 234	
	C4: 243	
<i>k6</i>	C1: 302	Can 1: A1, A4, A6, Can 2: A1, A4, A6, A7, A21, A28, A31 Can 3: A4, A6 Can 4: A6 Can 5: A4, A6, A7, A21
	C2: 109	
	C3: 86	
	C4: 113	
	C5: 203	
	C6: 309	
Miclust <i>k4</i>	C1: 651	Can 1: A27, A28 Can 2: A6, A7, A11, A13, A14, A18, A21, A23, A24, A25, A26, A29, A30, A31, A32, A33, A34, A36, A39 Can 3: A1, A2, A4, A6, A16, A22, A23, A25, A32, A33, A34, A35
	C2: 295	
	C3: 175	
	C4: 1	

All listed features have  $|\rho| > 0.20$  across at least 80% of the datasets. For the Miclust result, presence of the singleton cluster appears to inflate the contribution of the CDA correlations.

The SHAP results suggest that some of the radiology/imaging features (A11, A12, A13, A16, and A18) were important in addition to the demographic and injury related ones (A1, A4, A5, and A6). Both SHAP and CDA correlations agreed on age, mechanism of injury, and marital status as contributing the most to determining cluster membership, though they differed on the other selected features.

Figure 9 provides the pairwise comparison results for the *k6* clustering output based on Tukey and  $\chi^2$  tests. The p-values for each pair of clusters is provided above the diagonal (grey boxes). The orange boxes below the diagonal visually denote pairs of clusters that exhibited significant differences ( $p < 0.05$ ). These results were obtained using Rubin’s rules that account for the uncertainty in the imputed values (see Section II-D). A total of 26 features showed statistical significance on the pairwise tests for at least one pair. Fig. 9 displays the results for the 20 features that exhibited significance for at least three cluster pairs. The

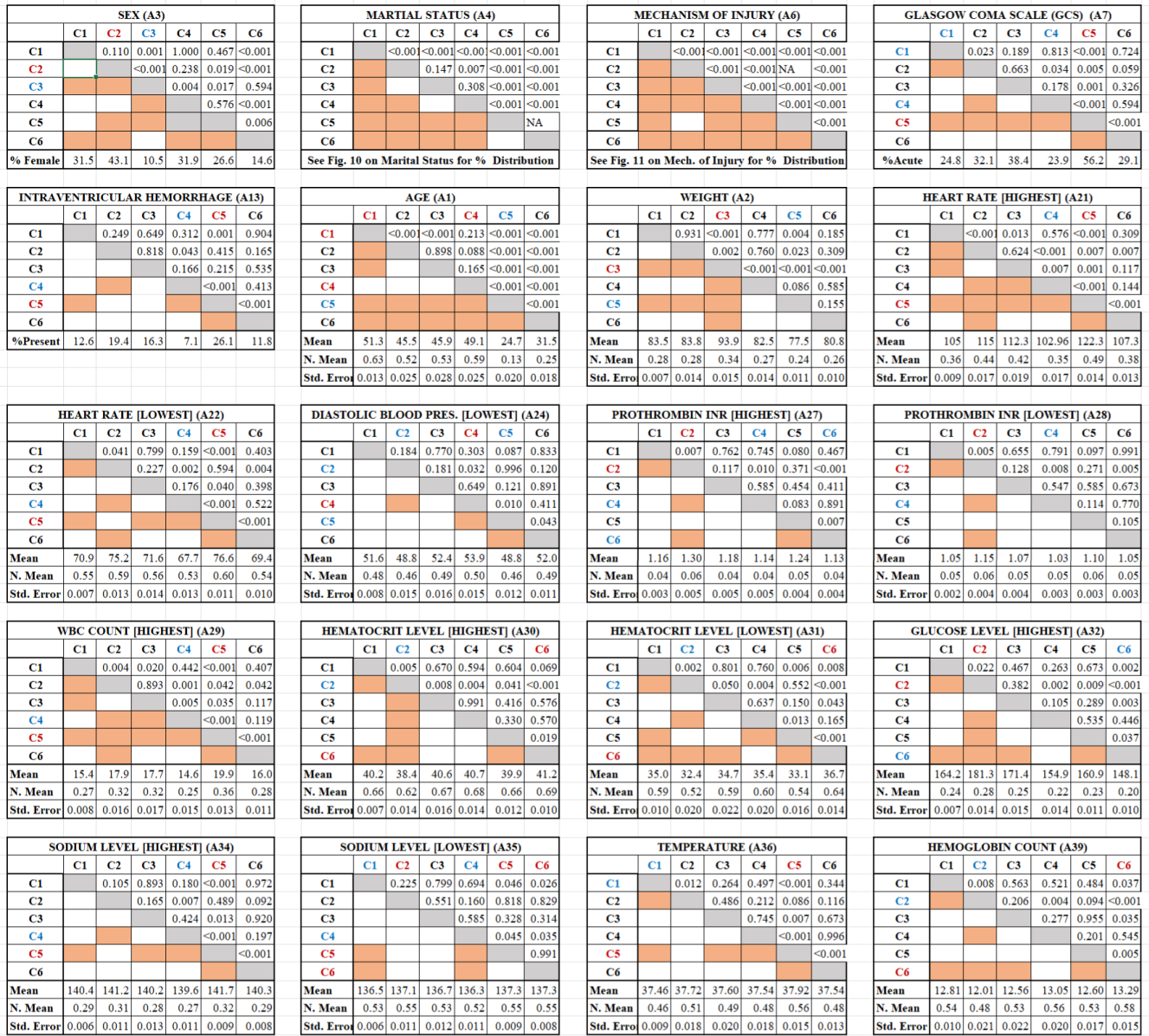
means on both the original and normalized scale per cluster are listed at the bottom of each sub-figure for continuous features, along with the standard errors for the normalized means. For ordinal and binary features (A3, A7, A13), we compare the clusters by the percentages of the most relevant level/category. The cluster(s) with the highest mean (or percentage) is highlighted in red, while the cluster(s) with the lowest mean (or percentage), in blue. The standard errors, obtained via Rubin’s rules, provide a measure of uncertainty in how much each cluster mean varies (incorporating both within and between imputation variation). From Fig. 9, these standard errors are similar for each cluster on all the features. More importantly, the p-values derived are based on a test statistic that incorporates this type of standard error (of the difference in two means). Hence, the significance results account for the uncertainty in the imputed values. Figures 10 and 11 illustrate the frequency distributions for categorical features (A4, A6) for each category by cluster.

The CDA, SHAP, and pairwise results provide different perspectives of feature importance. Pairwise tests are conducted on a feature-by-feature basis, which does not address any correlation that exists between features. In contrast, CDA and SHAP are multivariate methods that account for such relationships. However, only the pairwise tests provide a way to utilize Rubin’s rules to quantify and incorporate the imputation uncertainty into the results. We take a holistic approach by having the domain expert review the results of all three methods to select key features to characterize the clusters (as summarized in Table 4 and Fig. 12), but highlight that this is an area where further research is needed.

### V. DISCUSSION

Missing data remains a challenge to the usability and reliability of large biomedical datasets. This is true regardless of whether the source of data is carefully conducted clinical trials or electronic health records [8], [9], [56], [57]. Moreover, since clustering methods rely on complete case records, missing data poses obstacles to applying clustering to these datasets. We have presented a framework to address missingness in biomedical data by multiple imputation before





**FIGURE 9.** Input features by cluster that exhibited significance on pairwise tests using the multiple testing criterion, false discovery rate (FDR). P-values from Rubin's rules are shown on the upper diagonal. Orange shading in lower diagonal indicates the cluster pair is significantly different at  $p \leq 0.05$  by FDR. For continuous variables, the mean, normalized mean, and standard error (obtained via Rubin's rules) for each cluster are reported.

clustering. This framework has allowed us to cluster a large traumatic brain injury dataset from a clinical trial without deleting features or case records due to missingness. Though the empirical analysis in this paper is focused on TBI data, the framework generalizes to other types of large medical datasets, with MAR/MCAR missingness, for other heterogeneous diseases (such as cancer, stroke, autism spectrum disorders, etc.) for which cluster analysis is the end goal. We demonstrate that multiple imputation can be applied to address missingness in a manner that allows the successful application of clustering and offers a realistic estimate of the uncertainty introduced by the imputation.

The main premise of MI is that when there are missing data points present, there is an inherent uncertainty in trying to replace those values. Attempts to predict a single most accurate value for a missing data point (e.g., single imputation) would be treating the value as known, which can negatively impact the ability to make valid statistical inferences. MI addresses the uncertainty in the unknown values by generating multiple imputed values, applying an analysis to each imputed dataset, and then pooling them together in a way that allows valid inferences to be made. Though the imputed values will vary, by generating multiple clustering results that account for these fluctuations in the

TABLE 4. Overall cluster characterization narrative based on key Features.

Cluster (n)	C1 (302)	C2 (109)	C3 (86)	C4 (113)	C5 (203)	C6 (309)
<b>Demographics</b>	Older Divorced/Married	All Married Female↑↑	Married Male ↑↑ Higher weight	Older Married	Younger↓↓ All Single  Lower weight	Younger↓ All Single Male↑
<b>GCS Severity Mech. of Injury Hrs. btw Injury &amp; Scan</b>	Lower ↓ Fall from stationary object	All Motor Accident	Mostly Motorcycle Shorter	Lower↓↓	Higher All Motor Accident	Longer
<b>CT Presentation*</b>	Less ↓↓ Midline shift Abnormal Mesencephalic cistern Intraventricular blood Intraparenchymal lesion Subarachnoid hemorrhage	Less	Less ↓  More More	Less Less	More↑ More More  Less	More ↑↑ More
<b>Vitals</b>	Higher diastolic BP <sup>h</sup>	Lower diastolic BP <sup>l</sup>		Lower heart rate <sup>h,l</sup>	Higher temperature Lower diastolic BP <sup>h,l</sup> Higher heart rate <sup>h,l</sup>	
<b>Laboratory</b>		Higher Glucose Lower hematocrit <sup>h,l</sup>		Lower WBC	Higher WBC	Lower Glucose Higher hematocrit <sup>h,l</sup>

<sup>h,l</sup> denotes highest or lowest readings for vitals and laboratory values.  
 \* CT Presentation features based on percentage present, regardless of severity of presentation.

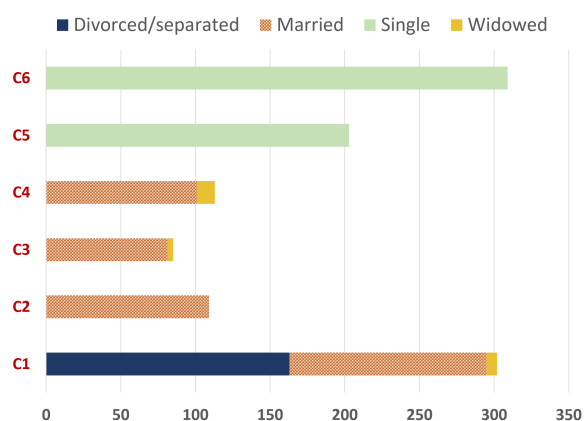


FIGURE 10. Distribution of marital status across clusters.

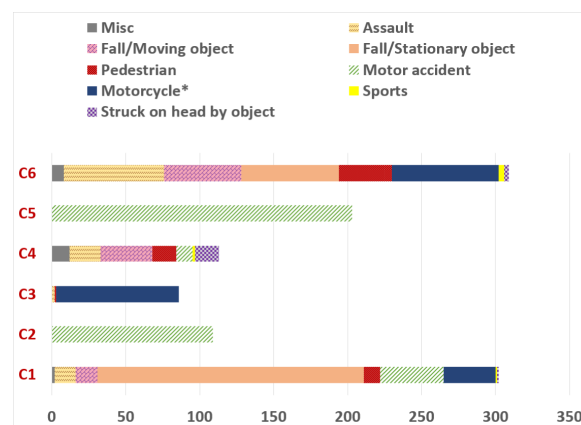
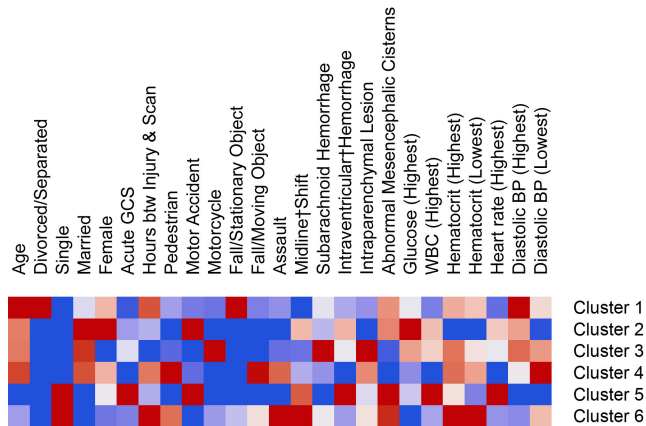


FIGURE 11. Distribution of clusters by Mechanism of Injury (MOI). Note, Motorcycle\* category includes accidents from ATV and golf carts as well.

imputed values and appropriately pooling them together, improves the overall robustness of the results. Some key aspects that distinguish our work from prior research on MI in clustering include: a start-to-finish framework for pre-processing, clustering, and post-hoc analysis; utilization of more robust clustering approach with ensemble clustering that utilizes multiple algorithms and decision metrics; mixture model to obtain consensus (pooled) cluster result; post-hoc analysis for cluster characterization with pairwise comparisons that employ Rubin’s rule to quantify uncertainty.

We applied this framework to ensemble clustering of 1213 TBI subjects in the COBRIT trial [51]. The trial enrolled subjects with either mild complicated TBI (n = 807) or moderate-to-severe TBI (n= 406). Subjects had received either citicoline or placebo randomized 1:1 in a double-blinded design. The study was halted after enrolling

1213 subjects due to the lack of a treatment effect. Enrolled subjects were heterogeneous in age (18 to 70 years), sex, marital status, and mechanism of injury. Subjects with mild TBI (concussions) or severe TBI (fixed dilated pupils, etc.) were excluded. Based on a consensus validation method that utilized multiple cluster quality metrics (Table 2 and Fig. 3) combined with domain expert opinion, we chose a cluster solution of k6. Cluster membership for the k6 solution was stable across all 15 missing data imputations (Fig. 4). Examination of the UMAP (Fig. 5) showed robust clusters, although some overlap was observed. The sub optimal result of the comparison method (Miclust) might be due to difference in algorithm: *k*-means vs. ensemble clustering with multiple algorithms (include *k*-means); utilization of a single validation metric (CritCF) vs. multiple decision



**FIGURE 12.** Heat map of features by cluster for k6. Features are organized similar to Table 4 demographics, GCS, mechanism of injury, CT scan, laboratory, vital signs. The color scale is cool/warm from Orange [55] normalized to the interval [0,1]. For continuous variables, the color scale reflects the cluster mean while for categorical variables, the proportion for each cluster.

metrics combined in an ensemble manner; and sensitivity to presence of outliers. Mclust clustering is based on  $k$ -means on each imputed dataset. The optimal value of  $k$  is selected based on the value of  $k$  that maximizes the CritCF criteria in most datasets. This approach differs from ours, which utilizes more sophisticated ensemble clustering and validation approach that utilize multiple algorithms and validation metrics to select  $k_{opt}$ . We were limited in comparisons to other approaches [11], [21], [22], [23] beyond mclust due to limited availability of implementable packages.

In contrast to prior work of conducting clustering on the same dataset without MI [6], here we obtain more complex clusters rather than simple clusters primarily driven by mechanism of injury. Definitive characterization of the six clusters based on ensemble clustering of the 1213 subjects in this dataset is more complex. We specifically looked at whether age, sex, initial TBI severity (GCS), mechanism of injury, medical, laboratory, and radiological activity were features that predicted cluster membership. Figures 9, 11, and 10 demonstrate the differences in features by cluster. Figures 11 and 10 emphasize the importance of the mechanism of injury and marital status in determining cluster membership. The mean values of continuous features and proportions of categorical features are converted to heat maps for side-by-side visual comparisons to aid in characterizing the clusters (Fig. 12). These results are converted to narrative tables emphasizing how clusters differ by feature (Table 4).

Folweiler et al. [52] conducted cluster analysis on the same 1213 COBRIT patient sample. They selected a set of features that exhibited  $<10\%$  missingness (MAR or MCAR) and applied multiple imputation with  $m = 5$ . They performed feature selection, using a generalized low-rank model, individually, on each of the 5 imputed datasets, and obtained a set of 6 features (platelet count, hematocrit, hemoglobin,

prothrombin time (PT), PT international normalized ratio (INR), and blood glucose level) that intersected all datasets. Cluster analysis was then carried out on a single data set pooled by averaging the  $m$  point estimates for each missing value on these features only. Averaging the values prior to clustering overlooks the notion that these imputed values are not deterministic and exhibit a degree of uncertainty. Thus, similar to single imputation, the clustering results and analysis does not address variation or uncertainty due to imputation. Clustering of these six features across the 1213 samples yielded three clusters. In contrast, our approach integrates MI in every aspect of the cluster analysis to yield a more robust performance. Four of their six clustering input features were also determined by our method to differ significantly across all the six clusters identified.

Our work places a key emphasis on cluster characterization, which is directly influenced by data imputation techniques. This can be subsequently extended for clinical relevance examination based on outcome measures, as conducted in prior work [6]. We observe that the C5 group appear to be the most severe group whose all had a motor accident (either as a driver or passenger), are all single and the youngest. Their vitals and lab values are extreme as well as their GCS scores. In contrast, the C2 group were all involved in a motor accident as well but are all married and have the highest percentage of females. They did have the least presence of intraparenchymal lesion. It would be interesting to compare the clinical outcome trajectory of both groups.

A strength of this study was that no subjects were lost, and no features were excluded due to missing data. Multiple imputation allowed us to use NMI as an estimate of the uncertainty in clustering introduced by imputed values. The MI framework supported the use of ensemble methods for clustering and ensemble methods for optimizing the number of clusters. Furthermore, the MI framework supported methods to find the discriminating features (SHAP and CDA) and methods to characterize and interpret the derived clusters. A primary limitation of this study is the increased time complexity of MI compared to single imputation. As illustrated in Fig. 2, after obtaining  $m$  imputed datasets, all subsequent analyses are  $\times m$ , until the application of Rubin's rule when the results are all pooled together. However, this is a linear effect, which is worthwhile given the resulting increase in robustness of the results. Furthermore, MI depends on the assumption that missing data are missing at random (MAR), which cannot always be verified. Although the COBRIT trial showed no treatment effect, we did not consider treatment by citicoline as an input feature. Although the primary outcome was 90-day survival and status, we did not characterize derived clusters by their outcome. This posed difficulties in generating cluster solutions that are *meaningful, interesting, and biologically plausible* [58] reflect the realities of the underlying data rather than any limitations of the methods used to impute missing data. Even the best imputation and clustering algorithms are limited by the quality and structure of the underlying data.

## VI. CONCLUSION

The paper presents a start to finish package for integration of multiple imputation in cluster analysis of complex biomedical datasets which mandate post-hoc steps to obtain robust and meaningful characterization of the clustering results. We demonstrate the utility of our framework by integrating the ensemble (consensus) clustering method. This allows us to leverage the strength of diverse algorithms. The goal of this work is not to introduce another clustering algorithm but rather to present a method for integration of multiple imputation in complex cluster analysis. Hence, the clustering module of the pipeline can be replaced with another algorithm of choice.

This paper demonstrates the utility of our framework to address missingness in a large traumatic brain injury dataset that utilizes multiple imputation prior to ensemble clustering. The pipeline sequentially identified features with missing values, used multiple imputation to create fifteen datasets, filtered out co-linear features, standardized features, and performed ensemble clustering on imputed datasets. Derived clusters in the consensus solution were characterized with heat maps, tables of feature means and proportions, statistical analysis of pairwise differences with Rubin's rules, and an overall narrative table. This approach allowed for an estimation of uncertainty based on both NMI among imputed datasets and standard error of the cluster means.

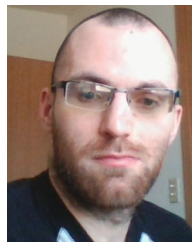
Comparisons with three different approaches on the same dataset demonstrate a more robust performance. The analysis revealed six multifaceted clusters that differed with respect to GCS, mechanism of injury, sociodemographics, vitals, lab values, and radiological presentation. The most severe cluster consists of single, relatively young patients that were injured by motor accident, and had higher GCS severity scores. This methodology for integrating multiple imputation in cluster analysis to tackle the missingness issue in a robust manner is generalizable. It can be applied to other medical datasets exhibit significant heterogeneity with comparable levels of missing data. Future work will focus on dealing with implications of datasets for which MNAR mechanism of missingness is more appropriate (such as developing a sensitivity analysis for clustering), and improving robustness of canonical discriminant analysis for all data types as well as exploring alternative pooling strategies that utilize Rubin's rules.

## REFERENCES

- [1] S. Saria, A. Butte, and A. Sheikh, "Better medicine through machine learning: What's real, and what's artificial?" *PLOS Med.*, vol. 15, no. 12, Dec. 2018, Art. no. e1002721.
- [2] J. Wilkinson, K. F. Arnold, E. J. Murray, M. van Smeden, K. Carr, R. Sippy, M. de Kamps, A. Beam, S. Konigorski, C. Lippert, M. S. Gilthorpe, and P. W. G. Tennant, "Time to reality check the promises of machine learning-powered precision medicine," *Lancet Digit. Health*, vol. 2, no. 12, pp. e677–e680, Dec. 2020.
- [3] T. P. Quinn, S. Jacobs, M. Senadeera, V. Le, and S. Coghlan, "The three ghosts of medical AI: Can the black-box present deliver?" *Artif. Intell. Med.*, vol. 124, Feb. 2022, Art. no. 102158.
- [4] A. Moya, E. Pretel, E. Navarro, and J. Jaén, "A systematic literature review of clustering techniques for patients with traumatic brain injury," *Artif. Intell. Rev.*, vol. 56, no. S1, pp. 351–419, Oct. 2023.
- [5] C. A. Åkerlund, A. Holst, N. Stocchetti, E. W. Steyerberg, D. K. Menon, A. Ercole, and D. W. Nelson, "Clustering identifies endotypes of traumatic brain injury in an intensive care cohort: A CENTER-TBI study," *Crit. Care*, vol. 26, no. 1, pp. 1–15, 2022.
- [6] D. Yeboah, L. Steinmeister, D. B. Hier, B. Hadi, D. C. Wunsch, G. R. Olbricht, and T. Obafemi-Ajayi, "An explainable and statistically validated ensemble clustering model applied to the identification of traumatic brain injury subgroups," *IEEE Access*, vol. 8, pp. 180690–180705, 2020.
- [7] M.-A. Schulz, M. Chapman-Rounds, M. Verma, D. Bzdok, and K. Georgatzis, "Inferring disease subtypes from clusters in explanation space," *Sci. Rep.*, vol. 10, no. 1, p. 12900, Jul. 2020.
- [8] S. Nijman, A. Leeuwenberg, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, and T. Debray, "Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review," *J. Clin. Epidemiol.*, vol. 142, pp. 218–229, Feb. 2022.
- [9] B. Y. Gravesteijn, C. A. Sewalt, E. Venema, D. Nieboer, E. W. Steyerberg, and C.-T. Collaborators, "Missing data in prediction research: A five-step approach for multiple imputation, illustrated in the CENTER-TBI study," *J. Neurotrauma*, vol. 38, no. 13, pp. 1842–1857, 2021.
- [10] Y. Luo, "Evaluating the state of the art in missing data imputation for clinical data," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab489.
- [11] V. Audigier and N. Niang, "Clustering with missing data: which equivalent for Rubin's rules?" *Adv. Data Anal. Classification*, pp. 1–35, 2022.
- [12] S. Van Buuren, *Flexible Imputation Missing Data*, 2nd ed. London, U.K.: Chapman & Hall, 2018.
- [13] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative robust semi-supervised missing data imputation," *IEEE Access*, vol. 8, pp. 90555–90569, 2020.
- [14] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls," *Brit. Med. J.*, vol. 338, Sep. 2009, Art. no. b2393.
- [15] L. Marston, J. R. Carpenter, K. R. Walters, R. W. Morris, I. Nazareth, and I. Petersen, "Issues in multiple imputation of missing data for large general practice clinical databases," *Pharmacoepidemiol. Drug Saf.*, vol. 19, no. 6, pp. 618–626, Jun. 2010.
- [16] S. V. Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.
- [17] J. L. Schafer, "Multiple imputation in multivariate problems when the imputation and analysis models differ," *Statistica Neerlandica*, vol. 57, no. 1, pp. 19–35, Feb. 2003.
- [18] P. Zhang, "Multiple imputation: Theory and method," *Int. Stat. Rev.*, vol. 71, no. 3, pp. 581–592, Dec. 2003.
- [19] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2019, vol. 793.
- [20] X. Basagaña, J. Barrera-Gómez, M. Benet, J. M. Antó, and J. Garcia-Aymerich, "A framework for multiple imputation in cluster analysis," *Amer. J. Epidemiol.*, vol. 177, no. 7, pp. 718–725, Apr. 2013.
- [21] L. Bruckers, G. Molenberghs, and P. Dendale, "Clustering multiply imputed multivariate high-dimensional longitudinal profiles," *Biometrical J.*, vol. 59, no. 5, pp. 998–1015, Sep. 2017.
- [22] L. Fauchoux, M. Resche-Rigon, E. Curis, V. Soumelis, and S. Chevret, "Clustering with missing and left-censored data: A simulation study comparing multiple-imputation-based procedures," *Biometrical J.*, vol. 63, no. 2, pp. 372–393, Feb. 2021.
- [23] R. Aschenbruck, G. Szeppannek, and A. F. X. Wilhelm, "Imputation strategies for clustering mixed-type data with missing values," *J. Classification*, vol. 40, no. 1, pp. 2–24, Apr. 2023.
- [24] A. I. Maas, D. K. Menon, G. T. Manley, M. Abrams, C. Åkerlund, N. Andelic, M. Aries, T. Bashford, M. J. Bell, and Y. G. Bodien, "Traumatic brain injury: Progress and challenges in prevention, clinical care, and research," *Lancet Neurol.*, vol. 21, no. 11, pp. 1004–1060, 2022.
- [25] A. Topchy, A. K. Jain, and W. Punch, "A mixture model for clustering ensembles," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2004, pp. 379–390.



- [26] D. Greene, A. Tsymbal, N. Bolshakova, and P. Cunningham, "Ensemble clustering in medical diagnostics," in *Proc. 17th IEEE Symp. Comput.-Based Med. Syst.*, Jun. 2004, pp. 576–581.
- [27] R. J. Little, J. R. Carpenter, and K. J. Lee, "A comparison of three popular methods for handling missing data: Complete-case analysis, inverse probability weighting, and multiple imputation," *Sociol. Methods Res.*, 2022, Art. no. 00491241221113873.
- [28] K. J. Lee, J. B. Carlin, J. A. Simpson, and M. Moreno-Betancur, "Assumptions and analysis planning in studies with missing data in multiple variables: Moving beyond the MCAR/MAR/MNAR classification," *Int. J. Epidemiol.*, vol. 52, no. 4, pp. 1268–1275, Aug. 2023.
- [29] (2018). *Imputation of Missing Values*. Accessed: Jan. 1, 2023. [Online]. Available: <https://scikit-learn.org/stable/modules/impute.html#iterative-imputer>
- [30] E. Casiraghi, D. Malchiodi, G. Trucco, M. Frasca, L. Cappelletti, T. Fontana, A. A. Esposito, E. Avola, A. Jachetti, J. Reese, A. Rizzi, P. N. Robinson, and G. Valentini, "Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments," *IEEE Access*, vol. 8, pp. 196299–196325, 2020.
- [31] O. M. Elzeki, M. F. Alrahmawy, and S. Elmougy, "A new hybrid genetic and information gain algorithm for imputing missing values in cancer genes datasets," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 12, pp. 20–33, Dec. 2019.
- [32] E. Slade and M. G. Naylor, "A fair comparison of tree-based and parametric methods in multiple imputation by chained equations," *Statist. Med.*, vol. 39, no. 8, pp. 1156–1166, Apr. 2020.
- [33] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study," *Amer. J. Epidemiol.*, vol. 179, no. 6, pp. 764–774, Mar. 2014.
- [34] B. Ramosaj and M. Pauly, "Predicting missing values: A comparative study on non-parametric approaches for imputation," *Comput. Statist.*, vol. 34, no. 4, pp. 1741–1764, Dec. 2019.
- [35] K. Al-Jabery, T. Obafemi-Ajayi, G. Olbricht, and D. Wunsch, *Computational Learning Approaches to Data Analytics in Biomedical Applications*. New York, NY, USA: Academic, 2019.
- [36] T. Nguyen, J. Viehman, D. Yeboah, G. R. Olbricht, and T. Obafemi-Ajayi, "Statistical comparative analysis and evaluation of validation indices for clustering optimization," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 3081–3090.
- [37] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, Dec. 2001.
- [38] T. Nguyen, K. Nowell, K. E. Bodner, and T. Obafemi-Ajayi, "Ensemble validation paradigm for intelligent data analysis in autism spectrum disorders," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, May 2018, pp. 1–8.
- [39] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, Jun. 2013.
- [40] C. Combi, B. Amico, R. Bellazzi, A. Holzinger, J. H. Moore, M. Zitnik, and J. H. Holmes, "A manifesto on explainability for artificial intelligence in medicine," *Artif. Intell. Med.*, vol. 133, Nov. 2022, Art. no. 102423.
- [41] Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu, *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*, vol. 760. Hoboken, NJ, USA: Wiley, 2011.
- [42] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020.
- [43] C. D. Newgard and J. S. Haukoos, "Advanced statistics: Missing data in clinical research—Part 2: Multiple imputation," *Academic Emergency Med.*, vol. 14, no. 7, pp. 669–678, May 2007.
- [44] R. Gittins, *Canonical Analysis: A Review With Applications in Ecology*. Cham, Switzerland: Springer, 1985.
- [45] *The Candisc Procedure: Computational Details*. Accessed: Nov. 11, 2022. [Online]. Available: [https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug\\_candisc\\_sect012.htm](https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_candisc_sect012.htm)
- [46] M. Friendly and J. Fox. (2021). *Candisc: Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis*. R Package Version 0.8-6. [Online]. Available: <https://cran.r-project.org/web/packages/candisc/candisc.pdf>
- [47] S. Grund, A. Robitzsch, and O. Luedtke. (2021). *MITML: Tools for Multiple Imputation Multilevel Modeling*. R Package Version 0.4-3. [Online]. Available: <https://CRAN.R-project.org/package=mitml>
- [48] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical J.*, vol. 50, no. 3, pp. 346–363, Jun. 2008.
- [49] A. Robitzsch and S. Grund. (2022). *miceadds: Some Additional Multiple Imputation Functions, Especially for 'MICE'*. R Package Version 3.15-21. [Online]. Available: <https://CRAN.R-project.org/package=miceadds>
- [50] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell RNA-seq data," *Nature Rev. Genet.*, vol. 20, no. 5, pp. 273–282, May 2019.
- [51] R. D. Zafonte, E. Bagiella, B. M. Ansel, T. A. Novack, W. T. Friedewald, D. C. Hesdorffer, S. D. Timmons, J. Jallo, H. Eisenberg, and T. Hart, "Effect of citicoline on functional and cognitive status among patients with traumatic brain injury: Citicoline brain injury treatment trial (COBRIT)," *J. Amer. Med. Assoc.*, vol. 308, no. 19, pp. 1993–2000, 2012.
- [52] K. A. Folweiler, D. K. Sandsmark, R. Diaz-Arrastia, A. S. Cohen, and A. J. Masino, "Unsupervised machine learning reveals novel traumatic brain injury patient phenotypes with distinct acute injury profiles and long-term outcomes," *J. Neurotrauma*, vol. 37, no. 12, pp. 1431–1444, Jun. 2020.
- [53] National Institute of Health. (Jun. 2019). *Federal Interagency Traumatic Brain Injury Research (FITBIR)*. U.S. Dept. Health & Human Services. [Online]. Available: <https://fitbir.nih.gov/>
- [54] M. Breaban and H. Luchian, "A unifying criterion for unsupervised clustering and feature selection," *Pattern Recognit.*, vol. 44, no. 4, pp. 854–865, Apr. 2011.
- [55] J. Demars and B. Zupan, "Orange: Data mining fruitful and fun—A historical perspective," *Informatica*, vol. 37, no. 1, p. 55, 2013.
- [56] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, and S. A. Murphy, "The prevention and treatment of missing data in clinical trials," *New England J. Med.*, vol. 367, no. 14, pp. 1355–1360, 2012.
- [57] B. K. Beaulieu-Jones, D. R. Lavage, J. W. Snyder, J. H. Moore, S. A. Pendergrass, and C. R. Bauer, "Characterizing and managing missing structured data in electronic health records: Data analysis," *JMIR Med. Informat.*, vol. 6, no. 1, p. e11, Feb. 2018.
- [58] D. L. Weed and S. D. Hursting, "Biologic plausibility in causal inference: Current method and practice," *Amer. J. Epidemiol.*, vol. 147, no. 5, pp. 415–425, Mar. 1998.



**ARNOLD A. HARDER** received the B.S. degree in mathematics from UAA and the M.S. degree in applied mathematics from Missouri University of Science and Technology (MST), where he is currently pursuing the Ph.D. degree with the Department of Mathematics and Statistics, under the advisement of Dr. Gayla R. Olbricht. His research interests include statistical analysis of missing data, genomic data, and cluster analysis applications.



**GAYLA R. OLBRICHT** (Member, IEEE) received the B.S. degree in mathematics from Missouri State University and the M.S. degree in applied statistics and the Ph.D. degree in statistics from Purdue University. She is currently an Associate Professor of statistics with the Department of Mathematics and Statistics, Missouri University of Science and Technology. Her research interests include statistical modeling of biological data, specializing in statistical genomics, and statistical analysis in clustering applications for biomedical data.



**DANIEL B. HIER** (Senior Member, IEEE) is currently an Adjunct Professor of electrical and computer engineering with Missouri University of Science and Technology and a Professor Emeritus of neurology and rehabilitation with the University of Illinois at Chicago (UIC). Previously, he was a Physician Executive with Cerner Corporation, Kansas, MO, USA, and the Head of Neurology and Rehabilitation with UIC. He is board-certified in both neurology and medical informatics.



**GODWIN EKUMA** received the B.S. degree in computer science from the Federal University of Technology, Owerri, Nigeria, and the master's degree in computer science from Missouri State University, under the advisement of Dr. Tayo Obafemi-Ajayi. His research interests include deep learning and biomedical informatics.



**TAYO OBAFEMI-AJAYI** (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in computer science from Illinois Institute of Technology. She is currently an Associate Professor of electrical engineering with the Engineering Program, Missouri State University (MSU). She is also the Director of the Computational Learning Systems Laboratory, MSU. Her research interests include machine learning, data mining, and biomedical informatics.

...