IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 12, 2020, accepted September 24, 2020, date of publication September 28, 2020,
date of current version October 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027453

An Explainable and Statistically Validated Ensemble Clustering Model Applied to the Identification of Traumatic Brain Injury Subgroups

DACOSTA YEBOAH[1], LOUIS STEINMEISTER[2,3], DANIEL B. HIER[3], BASSAM HADI[4], DONALD C. WUNSCH II[3], (Fellow, IEEE), GAYLA R. OLBRICHT[2,3], (Member, IEEE), AND TAYO OBAFEMI-AJAYI[3,5], (Member, IEEE)

[1]Department of Computer Science, Missouri State University, Springfield, MO 65897, USA
[2]Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO 65409, USA
[3]Applied Computational Intelligence Laboratory, Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA
[4]Mercy Hospital Neurosurgery, St. Louis, MO 63128, USA
[5]Engineering Program, Missouri State University, Springfield, MO 65897, USA

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

This work was supported in part by the Missouri University of Science and Technology Intelligent Systems Center, in part by the Mary K. Finley Missouri Endowment Fund, and in part by the Leonard Wood Institute in cooperation with the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-14-2-0034.

ABSTRACT We present a framework for an explainable and statistically validated ensemble clustering model applied to Traumatic Brain Injury (TBI). The objective of our analysis is to identify patient injury severity subgroups and key phenotypes that delineate these subgroups using varied clinical and computed tomography data. Explainable and statistically-validated models are essential because a data-driven identification of subgroups is an inherently multidisciplinary undertaking. In our case, this procedure yielded six distinct patient subgroups with respect to mechanism of injury, severity of presentation, anatomy, psychometric, and functional outcome. This framework for ensemble cluster analysis fully integrates statistical methods at several stages of analysis to enhance the quality and the explainability of results. This methodology is applicable to other clinical data sets that exhibit significant heterogeneity as well as other diverse data science applications in biomedicine and elsewhere.

INDEX TERMS Clustering, ensemble learning, canonical discriminant analysis, multicollinearity, precision medicine, mixed models, explainable AI, hybrid human-machine systems.

I. INTRODUCTION

Traumatic Brain Injury (TBI) is one of the major causes of death and disability. It could result in long term effects of impairments on an individual affecting physical (movement, vision, hearing), emotional (depression, personality changes), and/or cognitive (memory loss) functions. Annually, about 3 million TBI-related incidents cause emergency room visits, hospitalizations, or deaths in the United States [1]. TBI is a heterogeneous neurological disorder in cause, severity, pathology, and prognosis [2]. It can be caused by a number of things, including motor vehicle crashes, falls, assaults, and trauma. It may be associated with penetrating injury, focal contusion, different forms of hematoma (subdural, epidural, subarachnoid, intraparenchymal) or diffuse axonal injury. Some subjects may experience single or repetitive concussion (mild TBI) [3]. Sorting out the heterogeneity present in clinical data, though challenging, has the potential to reveal insights that could aid clinicians [4]. This study investigates an effective framework to characterize heterogeneous clinical data, specifically TBI, using unsupervised learning methods guided by domain expert knowledge and supplemented by statistical analyses to aid in applicability to prognostic and diagnostic analysis.

Unsupervised learning (clustering) separates an unlabeled data set into a discrete set of more homogeneous

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

subgroups [5], [6]. Cluster analysis has been applied to a wide range of problems as an exploratory tool to enhance knowledge discovery. In biomedical applications, unsupervised learning aids disease subtyping i.e. the task of identifying homogeneous patient subgroups that can guide prognosis, treatment decisions and possibly predict outcomes or recurrence risks [7]. In [8], we introduced an ensemble statistical and clustering model and applied it to an Autism Spectrum Disorder (ASD) sample using a set of 27 ASD phenotype features that spanned varied clinical and behavioral categories. Similar to ASD, TBI heterogeneity also presents a major challenge for phenotype categorization. Clinicians need better tools to assess TBI severity and determine key combinations of clinical features that delineate TBI subgroups. Currently, TBI can be classified by Glasgow Coma Scale (GCS) score according to different levels of clinical severity (severe, moderate, and mild), as well as by cranial computer tomography (CT) abnormality [3]. Data-driven decision support holds exceptional promise to enable development of clinical prediction tools for TBI, as well as other heterogeneous disorders, based on its ability to identify hidden correlations in complex data sets. This could elucidate the underlying physiological mechanisms of injury and recovery, and foster therapeutic advances.

This work leverages and formalizes the initial model proposed in [8] and [9] on a reliable TBI clinical data set to detect and characterize novel subtypes, based on discriminant phenotypes. We present a data-driven approach that incorporates not only the GCS score but multiple features extracted from the data set, including CT measurements and varied clinical factors, to identify key predictors. We assess the resulting subtypes using multiple outcome measures to establish clinical relevance. As reviewed in [5], [10], clustering algorithms vary widely based on data domain and problem scenarios. In this work, we utilize an ensemble cluster analysis approach that allows for effective integration of multiple clustering algorithms.

The TBI sample analyzed is drawn from the Citicoline Brain Injury Treatment Trial (COBRIT) [11] data set available from Federal Interagency Traumatic Brain Injury Research (FITBIR) [12] data repository to approved researchers. To the best of our knowledge, the only other previous unsupervised learning task conducted on this data set is a generalized low-rank model for feature selection integrated with cluster analysis in [2]. They identified four clusters with distinct feature profiles that correlated with 90-day functional and cognitive status. In this work, our overall aim is to identify homogeneous subgroups that could predict patterns of patient's prognosis and recovery outcomes. The underlying hypothesis is that novel combinations of statistics and machine learning will yield superior techniques for identifying subtypes of TBI that are amenable to improved outcomes via more precise interventions. This enhanced ensemble statistical and clustering model is an efficient and scalable solution applicable to other disorders that exhibit significant heterogeneity.

The outline of the paper is as follows. We present the ensemble learning framework in Section II. The experimental results obtained are presented in Section III, and a discussion of its clinical relevance in Section IV. Section V summarizes the findings of this work.

II. ENSEMBLE LEARNING FRAMEWORK

This section describes our methodology. The overall learning framework, illustrated in Fig.1, consists of three key phases: data curation, ensemble clustering, and model interpretation. Preliminary variants of this model have been introduced in [8] and [9]. The framework presented here formalizes a principled integration of the necessary components for a multidisciplinary analysis of complex heterogeneous biomedical data sets. The ensemble cluster analysis fully integrates statistical methods at several stages of analysis to enhance the quality and explainability of results.

A. DATA CURATION

Data quality is a major concern in big data processing and knowledge management systems. Data curation refers to preprocessing (cleansing) the data prior to data analysis [5]. This includes input features extraction, data representation, handling of missing data, eliminating redundancy among features and removing possible outliers.

1) INPUT FEATURES (PHENOTYPES) EXTRACTION

We present a domain expert guided process to determine relevant input features (phenotypic characteristics) for cluster analysis. Outcomes of cluster analysis are highly dependent on the set of input features that are selected, since it is an exploratory study. In this work, we are interested in clustering driven by the baseline patient data followed by interpretation of the clusters based on the recovery outcomes. The COBRIT study was a phase 3, double-blind, randomized clinical trial conducted over a span of 4 years to investigate the effects of citicoline compared to placebo on patients with TBI [11]. The study sample consisted of 1213 non-penetrating TBI patients (ages 18-70 years) with diverse severity levels according to GCS scores (3-15). It includes baseline data on demographics, injury information, and metabolic, liver and hematologic functions. Selected vital signs and blood sample data were repeatedly collected at multiple points during the study. Likewise repeated GCS scores and CT scan results were obtained at baseline and during hospitalization (from days 2 - 7). (Note that the COBRIT study did not find that active drug resulted in significant functional or cognitive status improvement [11]). The features investigated here are baseline measurements that span demographics, details of injury, CT scan findings, metabolic, liver and hematologic results obtained from blood samples, GCS scores, and vital signs (discussed in detail in Section III-A). Data types of input features could be numerical, categorical or ordinal. The numerical and ordinal features are normalized to a [0,1] range while categorical data are represented as bits of 0 and 1 using one-hot encoding.

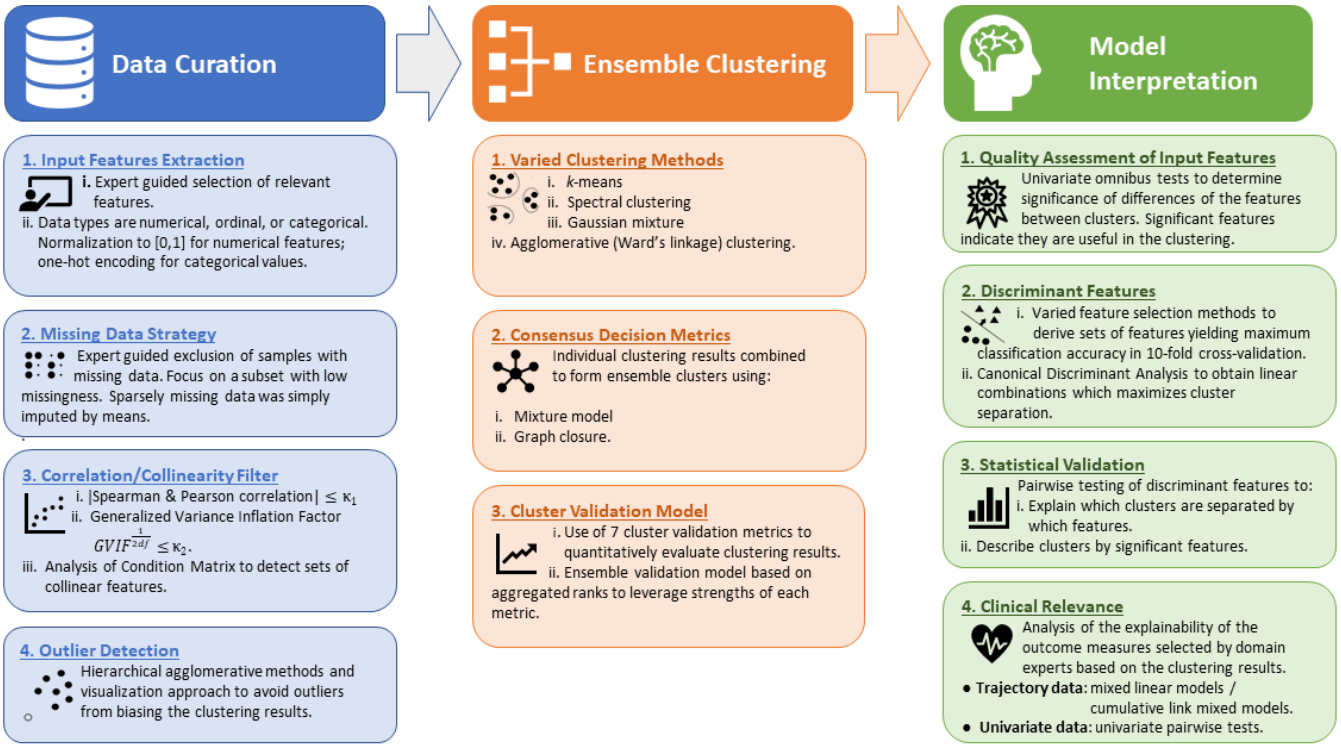


FIGURE 1. **Overview of the Explainable and Statistically Validated Ensemble Clustering Model.**

2) MISSING DATA STRATEGY

Missing data is a known issue with medical studies and multiple methods have been proposed to deal with this issue [5]. Almost all the patients had some level of missing information depending on the feature in question. This work excludes samples that have missing data for the key features (as determined by domain experts). This included patients with any missing CT scans anatomic sites and CT scans volume category measurement information. We focus on a subset of these patients with a consistently low percentage of missing data ($\leq$1%) across the numerical input features considered. These relatively low missing data values are imputed using the mean. Although more advanced techniques are available, these are beyond the scope of this work [13].

3) CORRELATION/COLLINEARITY FILTER

Conducting a robust correlation analysis on the potential set of input features prior to cluster analysis is very useful. It can aid in reducing bias due to multicollinearity [14] by filtering out features with high correlation. The correlation analysis is completed in two parts. First, pairwise correlations are assessed between two features at a time. Next, a metric is utilized that assesses multicollinearity between each feature and all other features.

Pearson and Spearman rank correlations are well suited for quantifying levels of correlation present among the pairs of input numerical features. Pearson measures the strength of linear correlation; whereas Spearman measures the strength of a monotonic relationship and is less sensitive to outliers. Both are calculated for thoroughness. For any pair of features that exhibit a correlation level of greater than a certain threshold (in absolute value), one is dropped. This process should be guided by the domain experts to pick an appropriate threshold for the tolerable level of correlation suited for the data application. (In this work, the allowable correlation threshold value is set at $\kappa_1 \leq 0.8$).

Generalized variance inflation factors (GVIF) assess multicollinearity between each individual feature and all remaining features. This is of relevance, since individual pairwise correlations may be small, but taken together they may still account for much of the information being introduced by the feature in question [15]. The GVIF is a generalized version of the Variance Inflation Factor (VIF) that is more appropriate for analyses in which categorical features are present among the input features, such as in this work. However, these values are not directly comparable as the encoding of different variables implies varying degrees of freedom (df). To compare these values across different dimensions, we consider $\text{GVIF}^{\frac{1}{2\,df}}$. In the case of a continuous scale variable, this is equivalent to computing the square root of the VIF [16]. As a rule of thumb in linear models, a VIF above 5 or 10 is usually indicative of the presence of some or severe multicollinearity, respectively. This cutoff could however be viewed as very lenient since such values correspond to a fraction of 80% and 90% of variance explained by the remaining features [17]. It should also be noted that, while such degrees of multicollinearity may be acceptable in the context of a linear model, different cutoff values may be appropriate in the setting of a

cluster analysis. This work utilizes a cutoff for $GVIF^{\frac{1}{2\,df}}$ of $\kappa_2 \leq 3$ which would correspond to a VIF of 9. After identifying problematic features, the condition indices of the feature matrix are also analyzed. Generally, condition indices > 30 could indicate a multicollinearity problem [18]. Features causing the collinearity problem can be determined by examining the variate involvement. Usually, features with a variance-decomposition proportion > 0.5 are highly probable. The exclusion decision is guided by domain knowledge.

4) OUTLIER DETECTION

An outlier can be defined as an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism [5]. Outliers are known to significantly bias clustering results, given the underlying assumption that every data point belongs to a cluster. This is an inherent challenge when clustering noisy data such as medical data in which outliers can be the result of human or data collection error. We apply a visualization approach using hierarchical agglomerative methods to identify and remove outliers to enhance robustness of results.

B. ENSEMBLE CLUSTERING

Clustering is a multidimensional optimization problem. For a single clustering algorithm, multiple results can be obtained by varying different parameters. It is important to compare and exploit different algorithms, as they can vary significantly in performance and outcome. Ensemble clustering (also known as consensus clustering) is an effective means to aggregate a collection of dissimilar clusterings to yield a more robust solution [19], [20]. Multiple clustering ensemble approaches have been proposed in literature [19]–[24]. Yoon *et al.* [21] developed a heterogeneous clustering ensemble scheme that uses a genetic algorithm to obtain robust clustering results. The ensemble model proposed by Huang *et al.* [24] utilizes an ensemble-driven cluster uncertainty estimation and local weighting strategy in contrast to their previous work that utilized a factor graph [23]. Zheng *et al.* [22] proposed a hierarchical ensemble clustering framework which combines both partitional clustering and hierarchical clustering results.

Greene *et al.* [20] investigated the effectiveness of several ensemble generations and integration techniques using varied synthetic and medical data. Their study, as well as the other ensemble approaches discussed, demonstrate that ensemble clustering provides considerable potential to improve the ability of the model to identify the underlying structure of both synthetic and real data in an unsupervised learning setting. Their results also suggested that diversity among the ensemble input algorithms is necessary, but not sufficient to yield an improved solution without the selection of an appropriate integration method. Hence, to ensure robustness, four individual algorithms (k-means, spectral, Gaussian mixture and agglomerative clustering with Ward's linkage) are utilized in the ensemble clustering model, as well as vigorous consensus decision metrics to facilitate a principled integration method.

To provide some context for this work, a brief description of each algorithm is provided. The k-means algorithm is perhaps one of the best-known and most popular partition-based clustering algorithms. It is regarded as a staple of clustering methods because of its merits and influence on other clustering approaches. K-means is based on an iterative optimization procedure to obtain an optimal k-partition of data by minimizing the sum-of-squared-error criterion [5]. Spectral clustering constructs a similarity graph on the data and derives the clusters by partitioning the embedding of the graph Laplacian [25]. The normalized spectral clustering algorithm [26] is utilized in this work. Agglomerative clustering is a type of hierarchical clustering based on a bottom-up approach. It initially assigns each data point to a cluster and then iteratively merges the clusters based on proximity (or linkage) measures, until a stopping criterion is attained [5]. Gaussian mixture clustering determines the number of components in a mixture and estimates the parameters of each component in a mixture based on the observations [27].

1) CONSENSUS DECISION METRICS

Consensus decision metrics refers to the process of fusing (or finishing) the various partitions obtained from the ensemble clustering scheme. The model utilizes two ensemble finishing methods: Mixture Model (MM) [19] and Graph Closure (GC) [28]. The MM method obtains the consensus partition using a maximum likelihood approach. The Expectation-Maximization algorithm derives maximum-likelihood estimates for model parameters. The k number of desired clusters is determined *a priori*. In contrast, GC obtains the consensus partition by viewing the co-occurrence matrix as a weighted graph. The graph is binarized based on a user-defined threshold to derive cliques (groups of nodes in a network such that every node is connected to each other node) [29]. The cliques are combined in the graph to create unique cluster formations [30]. These metrics are utilized based on their known effectiveness in comparison to other metrics, such as the majority voting method, in obtaining a robust final consensus partition [31].

2) CLUSTERING VALIDATION MODEL

The cluster analysis yields multiple solutions by varying the consensus metrics. Validation helps answer the fundamental question: ''How does one identify the optimal solution that translates to a ''meaningful'' configuration for a given domain application?'' Cluster validation is the process of estimating how well a given partitioning (from a clustering algorithm) aligns with the structure underlying the data [5]. Cluster validation metrics [5], [7] provide an objective measure to evaluate the quality of the clustering configuration derived. This is of critical importance when ground truth is limited. To evaluate the results of cluster analysis in a quantitative and objective fashion, we employ seven commonly used internal clustering validation metrics [32] including the

TABLE 1. Internal cluster validation metrics.

Index (Optimal)	Mathematical Description
SI (max)	$\frac{1}{k}\sum_i \left\{ \frac{1}{n_i} \sum_{x \in C_i} \frac{b(x)-a(x)}{\max_x[b(x),a(x)]} \right\}$ where $a(x) = \frac{1}{n_i - 1} \sum_{y \in c_i, y \neq x} d(x,y)$ $b(x) = \min_{j, j \neq i} \left[\frac{1}{n_j} \sum_{y \in C_j} d(x,y) \right]$
DB (min)	$\frac{1}{k}\sum_i \max_{j, j \neq i} \left[\frac{\frac{1}{n_i}\sum_{x \in C_i} d(x,c_i) + \frac{1}{n_j}\sum_{x \in C_j} d(x,c_j)}{d(c_i,c_j)} \right]$
Dunns (max)	$\min_i \left[\min_j \frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_k \max_{x,y \in C_k} d(x,y)} \right]$
XB (min)	$\frac{\sum_i \sum_{x \in C_i} d^2(x,c_i)}{N \min_{i,j \neq i} d^2(c_i,c_j)}$
CH (max)	$\frac{\sum_i n_i d^2(c_i,c)/(k-1)}{\sum_i \sum_{x \in C_i} d^2(x,c_i)/(N-k)}$
I (max)	$\left[\frac{\sum_{x \in D} d(x,c)}{NC \sum_i \sum_{x \in C_i} d(x,c_i)} \max_{i,j} d(c_i,c_j) \right]^p$
S_Dbw (min)	Scat(NC) + Dens_bw(NC) where $Dens\_bw(NC) = \frac{\sum_i \left[\sum_{j, j \neq i} \frac{\sum_{x \in C_i \cup C_j} f(x,u_{ij})}{\max(\sum_{x \in C_i} f(x,c_i), \sum_{x \in C_j} f(x,c_j))} \right]}{NC(NC-1)}$, $Scat(NC) = \frac{1}{NC}\sum_i \|\sigma(C_i)\| / \|\sigma(D)\|$ with $f(x,C_i) = \begin{cases} 0 & d(x,C_i) > stdev(C_i) \\ 1 & otherwise. \end{cases}$

D: data set; N: number of objects in D; C: center of D; k: number of clusters; C_i: the i^{th} cluster; n_i: number of objects in C_i; c_i: center of C_i; $d(x, y)$: distance between x and y; $\sigma(C_i)$: variance vector of C_i.

Silhouette Index (SI), Dunns index, Xie-Beni index (XB), I index, S_Dbw index, CH index and Davies-Bouldin index (DB). Each metric (Table 1) views the task of determining the optimal clustering configuration from a different perspective. To leverage the strengths of each metric, we utilize the ensemble validation model presented in [7] based on aggregated ranks.

C. MODEL INTERPRETATION

A key component of this framework is model interpretation which is critical for explainability. Cluster analysis is an exploratory tool that identifies subgroups (clusters) based on input features. These results are only meaningful through appropriate clinical interpretation. It is helpful to characterize the clustering result based on input features. This can be done by understanding which sets of features delineate each cluster. Understanding these differences allows clinicians to interpret the nature of the clusters. To establish clinical relevance and explainability of results, a set of pertinent outcome measures, selected by domain experts, should be evaluated to examine how they differ among the clusters. Results could indicate if these clusters have any predictive power for prognosis.

It is important to conduct a thorough statistical analysis to identify significant differences (if any) among the clusters for the input features as well as the outcome measures of interest. The appropriate statistical tests needed depend on the type of data available and hypothesis tested. Multiple testing corrections are required to control the false positive rate across the many tests that are conducted. This framework integrates statistical analysis at every level of the model interpretation phase to quantify significance between the clusters and ensure robustness of interpretation.

1) QUALITY ASSESSMENT OF THE CLUSTERING FEATURES

Univariate global statistical tests (omnibus tests) are useful to assess quality of features (variables) used as input features for cluster analysis. Univariate analyses determine whether there are any differences among the clusters for each feature. The method of analysis differs depending on the measurement scale of the feature. Significant features from these tests suggest that they are useful in the clustering.

For continuous variables, a one-way analysis of variance (ANOVA) is performed to test the null hypothesis for equality of means across the clusters. The homoskedasticity assumption for the ANOVA is assessed using a Levene test with medians (Brown-Forsythe test). A one-way ANOVA with White adjustment [33] is performed when homoskedasticity is not met. For ordinal variables, a Kruskal-Wallis test is employed to test for the equality of mean ranks across clusters, and a χ^2 test for independence is conducted for nominal data. Rejecting the null hypothesis indicates the feature differs significantly among clusters. Table 2 provides a summary of the tests conducted for the different measurement scales. The resulting p-values of these tests are corrected for multiple hypothesis testing using the Holm-adjustment, which controls the family-wise type I error (false positive) rate. The Holm method is uniformly more powerful than the well known Bonferroni-adjustment, which is very conservative especially in the case of testing many hypotheses [34], [35]. The false discovery rate (FDR) adjustment is also performed, which controls the expected proportion of false discoveries rather than directly controlling the family-wise type I error rate.

2) IDENTIFICATION OF SIGNIFICANT DISCRIMINANT FEATURES

Although the univariate analysis examines features individually to determine if they differ across clusters, it does not consider multivariate relationships between features or indicate which features may be most informative in delineating the clusters. Two additional approaches are applied to identify a set of discriminant features for further investigation.

The feature selection approach utilizes three commonly used feature selection methods (Best First Search, Scatter Search and Evolutionary Search) [37]–[40] to determine which features best delineate subgroups. Another method that is sometimes used is Minimum Redundancy Maximum Relevance (MRMR). It selects features by calculating redundancy between features and relevance between features and the class vector. However, it does not consider redundancy and interaction between features. The assumption is that the features are independent from each other [41]. Hence, this method is not used in this work. The best first algorithm is varied using the forward, backward and bidirectional parameters. The most optimal set of discriminant features is selected based on which one yields the maximum classification accuracy

TABLE 2. Overview of all statistical tests used in the model framework.

Assessment	Scale	Statistical Method
Univariate omnibus tests for quality assessment of input features	Continuous Ordinal Nominal	One-way ANOVA F-Test*,† for equality of group means Kruskal-Wallis Test for equality of group mean ranks χ^2 Test for independence
Pairwise testing for statistical validation of discriminant features	Continuous Ordinal Nominal	Two-sample t-Test† Post-hoc Conover Test Two-sided Fisher's Exact Test
Analysis of explainability of outcome measures & other measures of interest	Continuous scale trajectories Ordinal scale trajectories Univariate data	Mixed Model‡,Δ: $Outcome \sim Time * Cluster + (1\|ID)$ Cumulative Link Mixed Model$^{\Delta}$: $Outcome \sim Time * Cluster + (1\|ID)$ Pairwise tests as above

* A Levene Test with medians (Brown-Forsythe Test) is conducted to evaluate the homoskedasticity assumption. In case of a rejection, the F-Test is performed with a White-adjustment to account for heteroskedasticity.
† These tests assume normality of the observations for the test statistic's distribution to be accurate. Power transformations are commonly are used to "normalize" the data. However, these transformation can change the testing hypotheses in unintended ways [36]. Since the distributions are asymptotically accurate and we assess our group sizes to be adequate, we do not control for normality.
‡ This model assumes homoskedasticity and normality. Unlike the case of an ANOVA or t-Test, robustness cannot be argued by asymptotics due to the estimation procedure and inherent distribution assumptions.Tests for homoskedasticity are conducted and log-transforms are performed when these are violated. The normality assumption is visually checked by the use of QQ-plots. More information is contained in section II-C.
Δ Pairwise tests for difference in intercepts for different clusters at fixed times.
ID denotes individual subject random effect.

across 3 classification models (Random Forest, Multi-Layer Perceptron and Support Vector Machine).

In addition, a canonical discriminant analysis (CDA) is utilized to evaluate the discriminative power of multiple linear combinations of the features and capture interactions [42]. The advantage of CDA over linear discriminant analysis (which are equivalent under some circumstances) is that it can be generalized to work with categorical data, discriminant loadings can be easily obtained, and scores are useful for visualization. Rather than utilizing linear combinations of features which best explain the sample covariance (or correlation) structure, linear combinations (canonical discriminant functions) which best discriminate between the classes (clusters) are selected. Similar to principal component analysis, these correspond to eigenvectors (in this case, the "quotient" $S_b S_w^{-1}$ of the between group sum of squares and products matrix S_b and the within group sum of squares and products matrix S_w instead of the covariance or correlation matrix) [42]. Up to $k-1$ canonical discriminant functions can be obtained, where k is the number of classes or groups. The following results are reported for each set of clusters:

1) Squared canonical correlation (SCC). This gives the proportion of variation explained in the cluster grouping variable for each canonical variable (discriminating function).
2) p-values of the test for redundancy of additional canonical discriminant functions. A rejection implies that there is significant discriminative power to be gained in including the respective canonical discriminant function.
3) Pooled within-group correlations between each variable and the standardized canonical discriminant functions. The closer to $+1$ or -1, the more important the variable is in distinguishing the clusters.
4) Plots of the significant canonical variables to visualize their discriminative power. These can be viewed as dimension reduction of the input features to an orthonormal space that hierarchically maximizes the separation of the predefined groups.

The above methods indicate which set or linear combinations of features are useful in classifying an individual into one of the detected clusters. However, this does not provide information about which clusters the features separate between. To obtain these insights, pairwise tests are conducted to identify significant differences between individual clusters in a defined set of features. These results can help clinicians interpret the nature of the clusters based on the input features. The set of investigated features is defined as the union of the feature selection methods output and features which had discriminant loadings of a significant canonical variable exceeding 0.25 in absolute value in CDA. Since the most promising features are selected for further analysis, multiple testing corrections are conducted for all of the obtained p-values for all potentially performed tests (i.e. for all features that went into the selection process). Both the Holm and the FDR-adjustments are reported. For the pairwise tests, the analyses used are two-sample t-tests, post-hoc Conover tests (which can handle ties), and individual two-sided Fisher's exact tests for features of continuous, ordinal, and nominal scale, respectively (see Table 2).

3) OUTCOME MEASURES TO QUANTIFY CLUSTERING CLINICAL RELEVANCE

Clinical interpretation is the process of describing the clinical characteristics of the cluster. For example, is this a cluster of "severe" cases, a cluster of cases indicative of life threatening injury or permanent brain dysfunction, etc. Clinical relevance goes further to address the question "Do the 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 address clinical relevance, varied outcome measures that evaluate functional and cognitive recovery levels are selected with help from domain experts. Other variables of interest

determined by domain experts are also selected to quantify injury severity or unravel underlying data bias. We briefly describe the measures utilized in this work to determine clinical relevance of the resulting TBI subgroups, to provide a context for the statistical analysis.

Recommendations on suitable outcome measures for TBI studies are listed in [43]. Six of these measures were collected in the COBRIT study including the Glasgow Outcome Scale-Extended (GOS-E), California Verbal Learning Test-II (CVLT), Wechsler Adult Intelligence Scale-III Digit Span (DIGIT), Wechsler Adult Intelligence Scale-III Processing Speed Index (PSI), Controlled Oral Word Association Test (COWAT), and the Brief Symptom Inventory 18 (BSI-18) global severity index. The selected outcome measures are indicators of recovery from TBI. Importantly, these outcome measures are not used for cluster creation.

GOS-E is a global outcome assessment of a TBI patient based on 8 possible categories: dead (1), vegetative state (2), lower severe disability (3), upper severe disability (4), lower moderate disability (5), upper moderate disability (6), lower good recovery (7), and upper good recovery (8) [44]. CVLT is a detailed assessment of the patient's verbal learning and memory deficits. It evaluates the recollection and recognition of two lists of words over five learning trials. The patient's free recall and cued recall are assessed after short-term and long-term delay [45]. DIGIT is a neurological assessment of verbal short-term memory, which assists with the evaluation of a patient's cognitive status. It evaluates how a patient can respond to a series of mentioned numbers, recalling and repeating the numbers in the order they were presented [46]. PSI is a score relevant to a patient's ability to identify, discriminate, integrate, derive a choice about information, and to respond to both visual and verbal information. It is broken down into sub-tests: Coding, Symbol Search, and Cancellation, which evaluates overall visual perception, organization, attention focus, and memory [47]. COWAT, a supplemental measure for neuropsychological impairment assessment, is a verbal fluency test in which patients are required to generate words from initial letters [48]. BSI-18 is a self-patient reported scale that reflects a patient's progress, treatment outcome, and psychological assessments.

The available GOS-E data were recorded at different time points ranging from less than 30 days to beyond 180 days post-injury. For other assessments, data were reported at 30, 90, and 180 days post-injury time points. Data at 30 days were unavailable for DIGIT and PSI. Two measures are selected to determine severity of injury by clusters: mortality rate and number of days spent in intensive care unit (ICU) right after injury.

4) STATISTICAL ANALYSIS OF OUTCOME MEASURES

Outcome measures are essential to addressing the primary clinical research objectives which include characterization of the natural course of recovery from TBI, prediction of late outcomes, and comparison of outcomes to other studies [44]. To understand clinical relevance of the results, statistical analyses need to be conducted on the outcome measures to determine if there exist significant differences among the clusters at different post-injury time points. Since CVLT, DIGIT, PSI, COWAT, and BSI-18 are collected at two or three fixed post-injury time points, pairwise comparisons at each time point are conducted to determine which clusters differ significantly on the outcome measure. A mixed model analysis is performed with a random subject effect to account for missing data [49] and correlation between data collected on the same subject at multiple times. Fixed categorical effects of the model include cluster, time, and the cluster by time interaction.

The mixed model requires homoskedasticity and normality. To determine if the homoskedasticity assumption is met, an ANOVA can be applied on the squared residuals of the fitted model. A log-transform can be used to improve homoskedasticity, which is needed for all outcomes in this work. It should be noted that transformations can lead to unexpected changes in the actual hypotheses of the tests performed [36] and ought to be used with caution. In contrast to the omnibus and pairwise tests described in Table 2, an argument of asymptotic normality does not apply to the mixed model. Thus, normality is ascertained by the use of QQ-plots to avoid severe deviations. Log-transforms, used to improve homoskedasticity, also result in improved normality.

GOS-E is also collected at multiple post-injury time points, but unlike the previously mentioned outcome measures the time points are not fixed and may vary for different subjects. The GOS-E is also reported on an ordinal scale, which violates the necessary assumption of normality. Hence, a cumulative link mixed model (CLMM) is employed to examine the pairwise differences of the fixed effects at each time point. The model tests for equality of mean fixed intercepts across clusters at specified time points while incorporating the interactions of time and cluster. The model formulas for both models are given by $Outcome \sim Time * Cluster + (1|ID)$, where $Outcome$ is the response, $Time$ and $Cluster$ are modeled as fixed effects including interactions, and $(1|ID)$ indicates random intercepts at the individual subject level.

Other outcomes of interest (percentage of patients admitted to the ICU, days spent in ICU, and mortality rate) are also investigated by the same pairwise testing procedure described in Section II-C2 and Table 2.

Holm and FDR multiple testing corrections are performed in two ways for the outcome measures analysis. Simultaneous testing corrections of the unadjusted pairwise p-values are applied across all outcome measures to control the familywise false positive rate (Holm) or expected proportion of false discoveries (FDR) across all outcome measures. Since these adjusted p-values will depend on the number of outcome measures tested, corrections are also applied and reported for pairwise comparisons within each outcome measure. This allows for comparisons with future studies even if only a subset of outcome measures is chosen to study independently of the work presented here. The goal of these analyses is to provide domain experts with the information needed to

derive the clinical interpretation of the clustering results. It is important to note that the nature of this work is exploratory and aimed at showing the potential for clustering to identify clinically relevant subgroups of patients. In this context, the FDR adjustment is more meaningful since it allows for false discoveries (type I errors) but controls their proportion to true discoveries. This can increase power (lower type II error) at the expense of allowing some type I errors to occur. Regarding a confirmatory interpretation, the Holm adjustment is more suited as it controls the probability of any type I error occurring. We also suggest additional studies before potential clinical use.

III. RESULTS

A. DATA CURATION AND EXPERIMENT SETUP

In this work, a set of 35 features is initially considered, based on the attributes available from the COBRIT study and domain expert guidance (D.H and B.H). These baseline measurements include demographics, injury information, CT scans, metabolic, liver and hematologic functions obtained from blood samples, GCS scores, and vital signs. The missing data exclusion criteria (see Section II-A) reduced the study sample to 859 patients. Ten of these features (see Table 3: F2, F14, F15, F20, F21, F23, F24, F25, F27, F29) have some missing data ($\leq$ 0.1%) which are imputed using the mean values across all patients. The following features are excluded from further analysis based on the correlation/collinearity analysis: CT epidural lesion anatomic site, CT subdural lesion anatomic site and CT intraparenchymal lesion anatomic site. As a result of high correlation between the lowest prothrombin time and highest prothrombin time, only lowest prothrombin time is retained. CT lesion high mixed density feature is dropped as it displayed no variation among the patient sample considered. Three patients are identified as outliers and are excluded. The list of 30 features on the sample of 856 patients used for subsequent cluster analysis are described in Table 3.

TABLE 3. Description of input features and evaluation of significance by clustering output (k6 vs. k5).

Description of Feature	Tag	Data Type	*p*-value k6 result	*p*-value k5 result
Pre-injury Factors/Demographics				
$*$ $\diamond$Age	F1	Numerical	S^{hf}	S^{hf}
Weight	F2	Numerical	S^{hf}	S^{hf}
Injury Related factors				
$*$ $\diamond$Hours between injury & scan	F3	Numerical	S^{hf}	S^{hf}
$*$ $\diamond$Mechanism of injury	F4	Categorical	S^{hf}	S^{hf}
$*$ $\diamond$Glasgow Coma Score (total)	F5	Numerical	S^{hf}	S^{hf}
Radiology/CT Imaging				
Epidural lesion volume†	F6	Ordinal	S^{hf}	S^{hf}
Midline shift†	F7	Ordinal	S^{hf}	S^{hf}
$*$ $\diamond$Subarachnoid hemorrhage type	F8	Ordinal	S^{hf}	S^{hf}
Intraventricular hemorrhage	F9	Binary	S^{hf}	S^{hf}
Subdural lesion volume†	F10	Ordinal	S^{hf}	S^{hf}
$*$Intraparenchymal lesion volume†	F11	Ordinal	S^{hf}	S^{hf}
Mesencephalic cisterns type†	F12	Ordinal	S^{hf}	S^{hf}
Hydrocephalus present	F13	Binary	NS	NS
Clinical Factors				
Clinical Assessment				
Heart rate (highest)	F14	Numerical	S^{hf}	S^{hf}
$*$ $\diamond$Heart rate (lowest)	F15	Numerical	S^{hf}	S^{hf}
Systolic blood pressure (highest)	F16	Numerical	NS	NS
Systolic blood pressure (lowest)	F17	Numerical	S^{hf}	S^{hf}
Body temperature (highest)	F18	Numerical	S^{hf}	S^{hf}
Oxygen saturation (lowest)	F19	Numerical	NS	NS
Laboratory Test Results				
$\diamond$Prothrombin Time INR (lowest)	F20	Numerical	S^{f}	S^{f}
$*$White blood cell count (highest)	F21	Numerical	S^{hf}	S^{hf}
$*$ $\diamond$Hematocrit level (highest)	F22	Numerical	S^{hf}	S^{hf}
$*$ $\diamond$Hematocrit level (lowest)	F23	Numerical	S^{hf}	S^{hf}
Glucose level (highest)	F24	Numerical	S^{hf}	S^{hf}
Glucose level (lowest)	F25	Numerical	S^{f}	S^{f}
$*$ $\diamond$Sodium level (highest)	F26	Numerical	S^{hf}	S^{hf}
Sodium level (lowest)	F27	Numerical	NS	NS
Platelet count (lowest)	F28	Numerical	S^{hf}	S^{f}
$*$ $\diamond$Hemoglobin count (lowest)	F29	Numerical	S^{hf}	S^{hf}
Miscellaneous				
Hypertonic saline total volume	F30	Numerical	NS	NS

CT: Computed Tomography; INR: International Normalized Ratio.
†CT volumes converted to ordinal values
Significance criteria: Holm: S^{h}; False discovery rate: S^{f}; NS: Not significant
Discriminant features using feature selection methods: $*$: k6 result; $\diamond$: k5 result.

The ensemble cluster analysis is implemented using Open Ensembles Python library [28]. Three different combinations of algorithms for the ensemble cluster model are implemented. The first model (E1) is a combination of k-means, spectral, and agglomerative algorithms. The second model (E2) combines k-means, spectral, agglomerative, and Gaussian mixture while the third model (E3) aggregates spectral, agglomerative, and Gaussian mixture algorithms. For each algorithm, the number of clusters k is varied from 2 to 10, with 20 iterations per k, using each algorithm's default hyper-parameters. Applying both mixture model (MM) and graph closure (GC) consensus decision metrics on all the outcomes results in 33 different clustering outputs across the three ensemble models. The clustering outputs with very small cluster sizes (n<5) are discarded. The following output naming scheme is adopted: <model#><finishing method> k <#clusters>. For example, a 5-cluster solution from graph closure using E1 ensemble model is denoted by E1:GC:k5 while a 5-cluster solution from mixture model using E2 model is denoted by E2:MM:k5. Each ensemble model experiment is implemented once. Within each run, the base algorithms (k-means, Gaussian mixture, spectral, agglomerative), prior to obtaining the overall partition, are repeated multiple times by varying multiple parameters (number of clusters for single base algorithms, number of clusters for mixture model consensus function, threshold values for graph closure consensus function). For simulation comparisons, to assess the effectiveness of the ensemble clustering approach, we also conducted clustering using each base algorithm as a stand alone run.

The outcomes of the cluster quality evaluation using the ensemble cluster validation model is presented in Table 4. The Table illustrates the most optimal clustering results for each of the three ensemble models as well as the top three optimal results across the base clustering algorithms. The adjusted rank scores for the ensemble validation model are obtained using r=5. Interestingly, the E3 and E2 models both produced

TABLE 4. Cluster validation model ranking outcome.

Clustering Output	Overall Score	Adjusted Rank Score per CVM						
		Sil	Db	Xb	Dunn	CH	I	S_Dbw
Highest ranked result for each ensemble model								
E2:GC:k6	20	5	4	5	5	1	0	0
E3:GC:k6	20	5	4	5	5	1	0	0
E1:GC:k5	14	2	0	4	5	3	0	0
Highest ranked results across all base algorithm runs								
G:MM:k6	18	3	5	5	5	0	0	0
G:GC:k6	16	5	3	3	5	0	0	0
K:GC:k6	12	4	4	4	0	0	0	0

GC: Graph Closure; MM: Mixture Model; G: Gaussian mixture; K: k-means

the same set of 6 clusters that are the highest ranked for each: (E3:GC:k6 and E2:GC:k6). Thus, this result and the highest ranked E1 result (E1:GC:k5) are selected for further analysis. They are denoted by k6 and k5 respectively, in the remainder of this paper. A comparative visualization inspection of both results are illustrated in Fig. 2 using Isometric Feature Mapping (ISOMAP) algorithm [5].

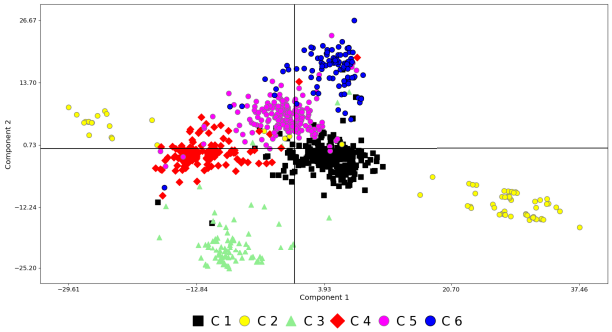
We can also observe from Table 4 that the top 3 results from the individual clustering algorithms are also 6-cluster results: G:MM:k6, G:GC:k6 and K:GC:k6. These three results varied slightly in cluster assignment of the samples, but are comparable to the 6-cluster result obtained from the ensemble model. The ISOMAP visualization from the base clustering algorithm runs are illustrated in the supplementary materials (Fig. S.1). However, the overall validation ranking is lower compared to the ensemble model's outcome. These comparisons provides further evidence in favor of the ensemble clustering approach's ability to yield a more robust clustering outcome. Subsequent analyses are conducted on the top two ensemble model results.

B. OUTCOME OF MODEL INTERPRETATION PHASE

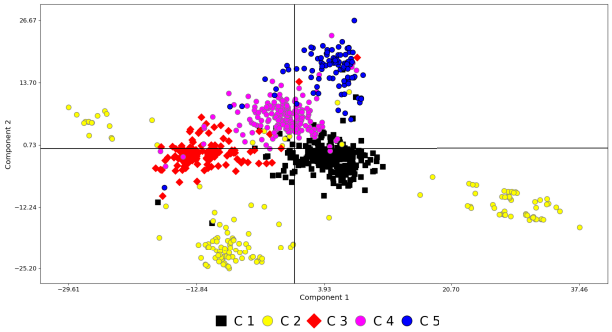
1) QUALITY ASSESSMENT OF INPUT FEATURES AND IDENTIFICATION OF SIGNIFICANT DISCRIMINANT FEATURES

Results from the global univariate statistical analysis are shown in Table 3. A majority of input features are identified as significant with both Holm and FDR approaches for both k5 and k6 clustering results. This validates the quality of the input features used. However, these results do not aid in identifying which features are most informative in the clustering, and the significance values do not take into account the interactions among the features.

The discriminant features identified using feature selection methods are shown in Table 3 for both k5 and k6 clustering results. The feature selection methods inform which features are most useful in explaining the resulting clusters. The features identified as discriminant had to have been selected by at least two of the feature selection methods (Best First Search, Scatter Search and Evolutionary Search). (A 10-fold cross-validation is applied for the classification models utilized to generate the results.) Out of the 30 input features, 11 are selected as discriminant for the k5 result, and 12 for



(a) k6 (6-cluster output).



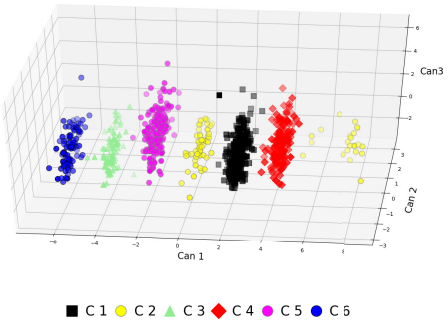
(b) k5 (5-cluster output)

FIGURE 2. ISOMAP visualization of top two clustering outputs from the ensemble clustering model.

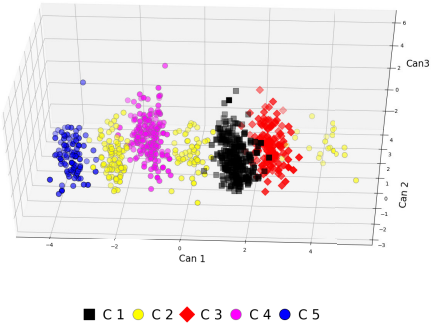
the k6 result, with 10 of these common between both. Intraparenchymal lesion volume, highest white blood cell count are discriminant only for the k6 result while the lowest prothrombin time INR is discriminant only for the k5 result. All discriminant features are also significant in the univariate analysis (either Holm or FDR or both) for both the k5 and k6 results.

The CDA results are presented in Table 5. For each clustering result, there are k-1 canonical variables. Features with discriminant loadings exceeding 0.25 in absolute value in CDA are listed for each canonical variable. In both the k5 and k6 cluster results, the first three canonical variables provide significant discriminatory power (p-value<0.05) while the remaining are not significant. The canonical plots (Fig. 3(a) and 3(b)), are generated using the first three canonical variables. The plots demonstrate that these variables almost perfectly explain the separation between the clusters. Only features on the first first three (significant) canonical variables with discriminant loadings exceeding 0.25 in absolute value are identified as being important in delineating the clusters and are further analyzed. There is a large overlap among features identified as important in distinguishing clusters in the CDA and those identified in the discriminant feature selection for both clustering outputs (Table 3), indicating an agreement between both methods.

Both the k5 and k6 clustering results exhibit similarity in terms of important features and visual inspections. We conferred with domain experts to determine which result seemed most meaningful to utilize for further analysis. The k6 result



(a) k6 (6 cluster) output: Can 1 vs. Can 2 vs. Can 3.



(b) k5 (5 cluster) output: Can 1 vs. Can 2 vs. Can 3.

FIGURE 3. Canonical plots of top two clustering outputs from the ensemble clustering model.

TABLE 5. Canonical Discriminant Analysis (CDA) results for top two optimal clustering outputs.

Clustering Output	Features important for distinguishing clusters	SCC	*p*-value
k6	Can 1: F4**, F29	0.92	<0.001
	Can 2: F1*, F2, F14, F15, F18, F21	0.20	<0.001
	Can 3: F1, F6, F9, F10, F13, F14, F15, F21, F22, F23, F24, F25	0.07	0.002
	Can 4: F1, F2, F7, F11, F21, F22, F23, F29	0.04	0.108
	Can 5: F3, F8, F12, F15	0.04	0.247
k5	Can 1: F4**, F5, F14, F21, F29	0.78	<0.001
	Can 2: F1*, F2, F14, F15, F18	0.19	<0.001
	Can 3: F1, F6, F9, F10, F12, F13, F14, F15, F21, F22, F23, F24, F25	0.07	<0.001
	Can 4: F1, F2, F8, F15, F21, F22, F23, F29	0.04	0.157

SCC: Squared Canonical Correlation.
**: Absolute value of canonical variable discriminant loading (cvdl) > 0.9.
*: Absolute value of cvdl > 0.7.
All other listed features have absolute value of cvdl > 0.25.

is chosen based on an examination of pairwise differences in feature means between clusters for identified discriminant features (Fig. 4). These pairwise results are a union of input features identified from CDA and feature selection methods that demonstrated significance using the multiple testing criterion (Holm & FDR). Each sub-figure presents the results for two features, with the mean and standard deviation for each cluster provided below each feature name for numerical variables. For non-numerical variables (Fig. 4h), the percentage distribution per cluster is given for intraventricular hemorrhage presence and the highest level of subarachnoid hemorrhage severity (convexities/sulci and cisterns present). For mechanism of injury (Fig. 4i), only one mechanism is present in each cluster except C2 and this mechanism is reported in the Figure. For C2, three mechanisms are present (vehicle assault, sports, and non-assault object strike) and none of these are present in the other clusters. The mechanism with the highest frequency (vehicle assault) in C2 is reported in the Figure. Results in the lower left triangle give the pairwise significance for the feature on the left side of the sub-figure; whereas the upper right triangle gives the results for the feature on the right side. The grayed out cells containing an asterisk separate these two sets of results. Brown denotes a statistically significant difference between two clusters by both FDR and Holm. Tan denotes significance by FDR only. Note that there are some features that are deemed discriminant according to CDA and the feature selection methods (intraparenchymal lesion volume (both), subdural lesion volume (both), hydrocephalus present (CDA only)), however they are not listed in Fig. 4 since they failed the multiple testing significance criterion. Likewise there are some features that are not selected based on the discriminant features analysis but exhibited some pairwise significance based on multiple testing using FDR (platelet count, systolic BP, prothrombin time INR, and mesencephalic cisterns type).

2) STATISTICAL ANALYSIS OF OUTCOME MEASURES

Table 6 shows the results for four TBI outcome assessments (BSI-18, CVLT total adjusted score, DIGIT, and PSI) that demonstrated pairwise significance between clusters in the mixed model for at least one time point. The mean and standard deviation are reported along with the percentage of missing data at each time point. For all assessment scores, higher values indicate better outcomes. Multiple testing adjustments are done both within a single outcome (Holm P) and across the entire span of 8 outcome measures considered (Holm MT and FDR MT). Only time points with a statistically significant difference for at least one of these methods are reported. Note that the trajectory of COWAT (see Supplementary Materials) did not display a significance difference for any time point.

The recovery trajectory plot of GOS-E across clusters is illustrated in Fig. 5. The time points are binned since there was high variability in the different time points at which the GOS-E scores were obtained from the patients. Thus, time is modeled as a categorical, rather than continuous, variable. Table 7 gives the pairwise results for mean differences between clusters at each of the GOS-E binned time points from the CLMM model. At least one pairwise difference is significant at each time point with the exploratory FDR MT adjustment, although most of these are not confirmed using the Holm MT adjustment. Further studies may prove insightful. Additional outcomes of interest, that quantify severity of injuries, are reported in Table 8. Since the data available on days spent in ICU exhibit strong skewness, the median is reported as a measure of location, and the 25% and 50% quantiles, as an indication of spread.

IV. DISCUSSION OF CLINICAL RELEVANCE

Patients that suffer from TBI are a heterogeneous population that exhibit diverse pathologies, prognoses, and recovery trajectories. Sorting out this heterogeneity is challenging. However, a verifiable and explainable model has the

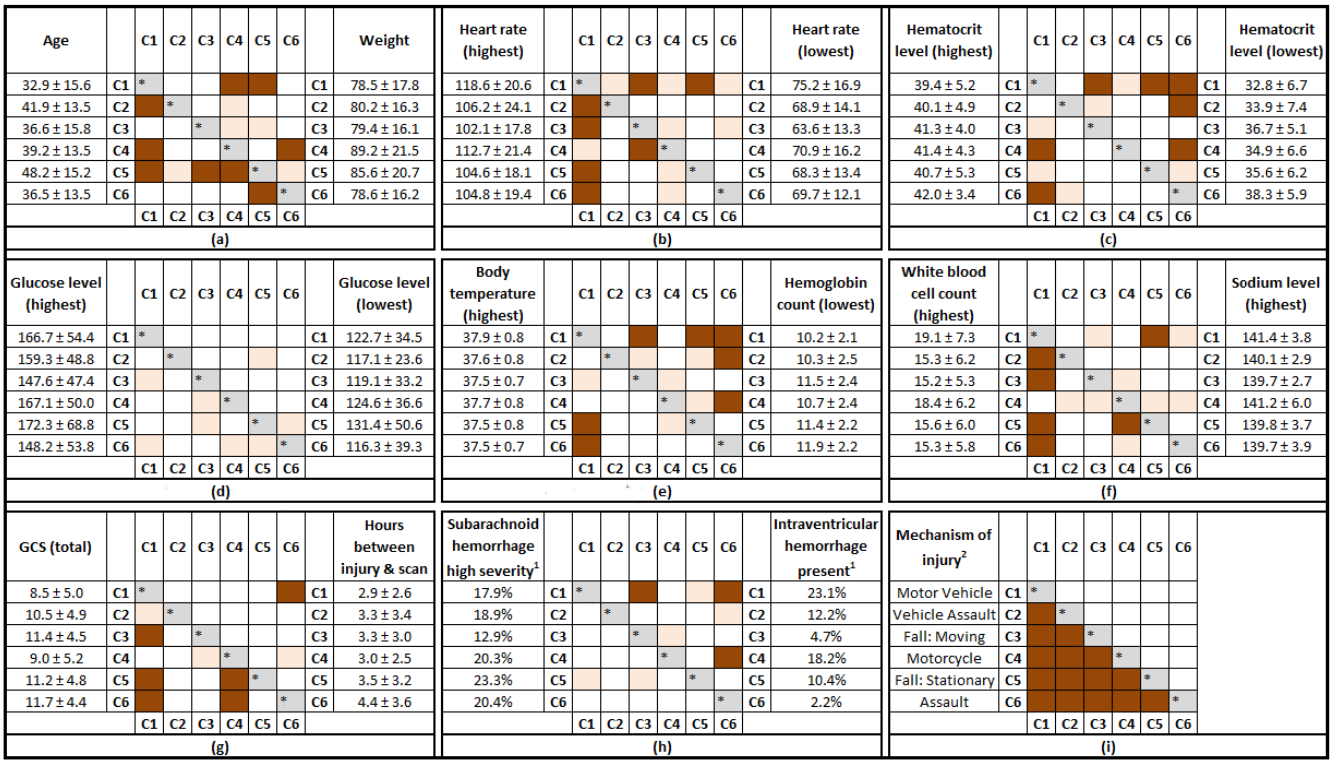


FIGURE 4. Overview of clusters based on union of input features from canonical discriminant analysis and feature selection methods that exhibited significance using the multiple testing criterion (Holm & False discovery rate: FDR). Brown denotes significance by both FDR and Holm. Tan denotes significance by FDR only. For features on the left side of a box refer to the lower left triangle and for features on the right side of a box, similarly, refer to the upper right triangle of the matrix. The two are separated by the grayed out cells containing asterisk (∗). For numerical variables, mean±standard deviation is reported. For the non-numerical variables, [1] indicates percentage distribution per cluster for the highest level of severity (convexities/sulci and cisterns present for subarachnoid hemorrhage). [2] indicates the highest frequency of mechanism of injury modality for each cluster.

TABLE 6. Mean and Standard Deviation of Outcome Assessment Measures per Subgroup.

Cluster (size)	BSI-18 90 days		California Verbal Learning Test-II 90 days		California Verbal Learning Test-II 180 days		Digit Span 90 days		Digit Span 180 days		Processing Speed Index 90 days		Processing Speed Index 180 days	
	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR
C1 (268)	56.6 (11.2)	32%	41.5 (13.6)	36%	44.9 (15.3)	47%	9.1 (2.6)	36%	9.1 (2.7)	47%	86.1 (15.2)	37%	90.7 (17.8)	49%
C2 (74)	58.7 (11.7)	35%	44.9 (13.9)	36%	50.7 (15.4)	45%	9.6 (3.1)	38%	10.5 (3.2)	46%	94.7 (16.6)	41%	99.4 (19.1)	47%
C3 (85)	56.7 (11.8)	21%	47.0 (14.0)	28%	53.0 (16.0)	32%	9.9 (2.7)	28%	10.1 (2.6)	33%	93.9 (17.1)	29%	98.8 (19.8)	33%
C4 (143)	54.4 (11.1)	36%	43.9 (15.4)	36%	47.4 (14.9)	48%	9.0 (3.0)	36%	9.6 (2.9)	48%	86.5 (15.1)	41%	90.4 (16.2)	48%
C5 (193)	56.8 (12.6)	36%	46.4 (14.7)	39%	50.3 (15.4)	55%	9.3 (2.5)	40%	9.5 (2.9)	55%	91.2 (16.0)	41%	95.0 (17.3)	55%
C6 (93)	60.2 (13.3)	27%	41.1 (13.4)	30%	47.4 (15.6)	47%	8.2 (2.6)	29%	8.5 (3.1)	47%	87.5 (14.1)	32%	92.6 (16.0)	48%
Holm P	NS		NS		NS		NS		C6:C2, C3		C1:C2		NS	
FDR MT	C4:C6		C1:C3, C5		C1:C3		C3:C6		C6:C2, C3		C1:C2, C3 C4:C2, C3		C1:C2, C3	

BSI-18: Brief Symptom Inventory-18; NR: Not Reported Data; SD: Standard Deviation;
Holm P: p-values are adjusted for multiple testing with Holm for pairwise comparisons within single outcome.
FDR MT: p-values are adjusted for multiple testing with FDR for pairwise comparisons across all outcomes.
The Holm multiple testing (MT) results are not shown since none of the comparisons are significant using the Holm adjusted p-values across all outcomes.

potential to reveal insights that could aid clinicians. This work investigates such a model that identifies phenotype features which delineate patients into more homogeneous subgroups, and characterizes these groups in terms of severity of injury and recovery.

The data utilized in this study were obtained from the COBRIT data of 1213 TBI subjects accumulated over a 4-year period. Data curation (including elimination of outliers and subjects with excessive missing data) yielded 856 usable subjects for ensemble clustering. Thirty features (Table 3) were used for the cluster analysis. We combined four different clustering algorithms (agglomerative, Gaussian mixture, spectral, and k-means) into three different ensembles (E1, E2, and E3) and two different consensus models (GC and MM). A key component of our framework is intentional assessment of clustering quality and statistical validation at every phase. To assess the clustering quality and ensure that the optimal solution is selected, we employ an ensemble cluster

TABLE 7. Recovery trajectory of Glasgow Outcome Scale-Extended (GOS-E) distribution per cluster from less than 30 days to 6 months post-injury.

Cluster (size)	≤30 days		31 - 60 days		61 - 90 days		91 - 120 days		121 - 180 days		>180 days	
	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR	Mean (SD)	NR
C1 (268)	3.7 (1.9)	59.0%	3.9 (1.8)	45.5%	4.7 (2.2)	55.2%	4.9 (2.3)	51.9%	4.8 (2.5)	64.2%	5.1 (2.3)	54.9%
C2 (74)	3.7 (1.6)	56.8%	4.2 (2.0)	48.6%	4.4 (2.1)	62.2%	4.9 (2.1)	50.0%	4.8 (2.1)	60.8%	5.1 (2.3)	52.7%
C3 (85)	4.5 (1.8)	61.2%	5.2 (2.0)	52.9%	6.1 (1.9)	50.6%	5.7 (2.1)	63.5%	5.9 (2.0)	58.8%	6.2 (2.2)	57.6%
C4 (143)	4.0 (2.1)	60.1%	4.2 (2.0)	43.4%	5.0 (2.5)	62.2%	4.6 (2.3)	49.7%	5.1 (2.4)	60.1%	5.1 (2.5)	60.1%
C5 (193)	4.3 (2.2)	57.5%	4.0 (1.9)	49.7%	4.6 (2.4)	60.1%	4.9 (2.4)	47.7%	4.9 (2.7)	57.0%	4.9 (2.7)	56.5%
C6 (93)	4.7 (1.4)	55.9%	5.4 (1.8)	58.1%	5.5 (1.6)	55.9%	6.2 (1.5)	65.6%	5.8 (1.8)	71.0%	6.5 (1.4)	63.4%
Holm P	NS		C1:C6		C1:C3		NS		NS		NS	
Holm MT	NS		C1:C6		NS		NS		NS		NS	
FDR MT	C1:C6		C6:C1, C2, C4, C5 C3:C1, C5		C3:C1, C2, C4, C5		C6:C1, C2, C4, C5 C3:C2		C3:C1, C2 C6:C2		C3:C1, C2 C6:C1, C2, C5	

Holm P: p-values are adjusted for multiple testing with Holm for pairwise comparisons within single outcome.
Holm (FDR) MT: p-values are adjusted for multiple testing with Holm (FDR) for pairwise comparisons across all outcomes.
Values are rounded to one decimal place and may not reflect differences at higher order place values, as shown more clearly in Fig. 5.

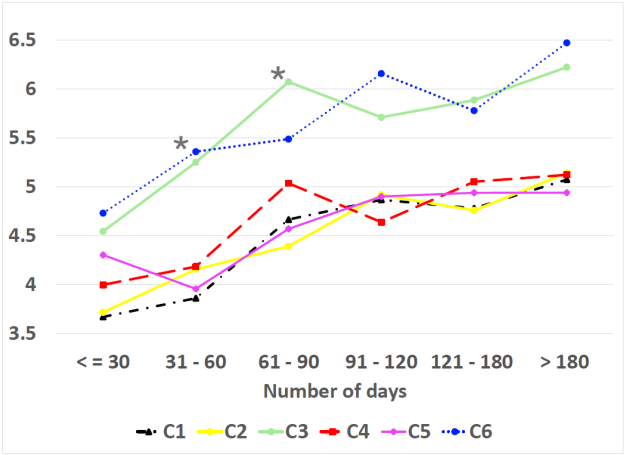


FIGURE 5. Mean trajectory of Glasgow Outcome Scale-Extended scores across clusters.The asterisk (∗) markers indicate the time points where there is at least one significant pairwise difference (adjusted with Holm for pairwise testing, not for multiple testing across other outcomes) in the GOS-E output between the clusters. For further information refer to Table 7.

TABLE 8. Severity view of clusters based on number of days in intensive care unit (ICU) (Median [25% Quantile, 75% Quantile]), and mortality information.

Cluster (size)	% in ICU	Days in ICU Median [Q1, Q3]	Mortality Info. % Death Reported
C1 (268)	97.0%	5.0 [2.0, 14.25]	7.1% (1.6)
C2 (74)	97.3%	4.5 [2.0, 15.0]	5.4% (2.6)
C3 (85)	94.1%	2.0 [1.0, 6.25]	3.5% (2.0)
C4 (143)	94.4%	5.0 [2.0, 13.5]	7.0% (2.1)
C5 (193)	92.8%	3.0 [2.0, 9.5]	11.9% (2.3)
C6 (93)	97.9%	2.0 [1.0, 4.5]	1.1% (1.1)
Holm P		C1:C3, C5, C6 C2:C3, C6; C4:C3, C6	C5:C6
Holm MT		C1:C3; C6:C1, C2, C4	NS
FDR MT		C1:C3, C5, C6 C3:C2, C4; C6:C2, C4	C5:C6

Q1: 25% Quantile, Q3: 75% Quantile.
All but 3 patients (1 each from C4, C5 and C6) were hospitalized. Hospital information is missing for these 3.
Not all patients ended up in ICU, as shown in column 2.
Holm P: p-values are adjusted for multiple testing with Holm for pairwise comparisons within single outcome.
Holm (FDR) MT: p-values are adjusted for multiple testing with Holm (FDR) for pairwise comparisons across all outcomes.

validation paradigm [7]. As discussed in [7], limiting the selection of the optimal clustering solution to one prior to domain experts' guidance might be misleading. These measures of cluster quality are agnostic as to whether the derived clusters are clinically ''meaningful'', that is, do these clusters assist physicians in thinking about a heterogeneous disorder like TBI in a more insightful way? Hence, we presented the top two clustering solutions (k6 and k5 results), as indicated by the cluster assessment phase (Table 4), to the domain experts for inspection based on visualization (Fig. 2, Fig. 3) and descriptive means (or mode for ordinal and categorical variables). The selection of the k6 result as the optimal result is based on a qualitative evaluation and comparison. Visual inspection of the clusters indicates excellent cluster separation with both k5 and k6 clustering solutions.

The clinical severity presentation, functional and cognitive outcomes of the resulting six clusters differ on multiple fronts (see Fig. 4). It is useful to examine the features selected for the CDA equations (Table 5) as most important in delineating the clusters for the k6 clustering result. These discriminating features can be grouped as disease mechanism (mechanism of injury), demographic features (age and weight), features that reflect brain injury on CT scan (hydrocephalus, epidural blood volume, subdural blood volume, intraventricular hemorrhage (IVH)), and features that reflect patient criticality (heart rate (highest and lowest), highest body temperature, highest white blood cell count (WBC), hematocrit (highest and lowest), glucose (highest and lowest), lowest hemoglobin count). It is interesting to observe that the feature selection method approach (Table 3) also strongly aligned with the CDA results. It identifies additional features beyond the CDA: total GCS score as well as additional CT scan features (subarachnoid hemorrhage (SAH) type, intraparenchymal lesion volume), and other features that reflect patient criticality (hours between injury and CT scan, highest sodium). This suggests the usefulness of combining both approaches to determine the set of discriminant features.

Examination of Fig. 4, Fig. 5, and Table 7 suggest that the six derived clusters do differ from each other in clinically meaningful ways. The resulting clusters suggest that mechanism of injury plays a key role in predicting both initial severity and long-term outcome. C1 consists of patients injured as a result of occupant motor vehicle accident (MVA), either a driver or passenger in a motor vehicle. C2 is a heterogeneous mechanism of injury group whose injury resulted either from pedestrian MVA, sports or being struck on head by an object but not assault. Pedestrian MVA mechanism is the most common in C2 (70% of the group). The C3 patients were injured due to a fall from a moving object such as bike, skateboard or horse while C4 patients, as a result of being an occupant (passenger/driver) in a motorcycle or all terrain vehicle or golf cart accident. For C5 cluster, the injuries were triggered by a fall from a stationary object (such as roof/ladder) while for C6 cluster, by any means of assault.

In terms of modal (most common) mechanism of injury clusters C1, C2, C4 involve motor vehicles or motorcycles, clusters C3 and C5 involve falls, and cluster C6 involves assaults. C1 patients have the highest group mean values for WBC counts and lowest hematocrits and hemoglobin counts, which could be indicative of a high level of trauma and blood loss associated with the injury. This is also aligned with having lowest initial total GCS score (group mean: 8.5) and shortest time in between injury and scan. Usually, patients with a greater degree of severity of injury tend to get to the CT scans quickest. However, this cluster appears to have a significantly better recovery outcome over time, as quantified by the GOS-E trajectory compared to the oldest cluster (C5). Though C5 has a relatively high initial total GCS score (group mean: 11.2) and lowest WBC count, they appear to make the slowest recovery over time, as evident by their GOS-E trajectory. In terms of demographics, C1 cluster consist of the youngest age group while C5 cluster is relatively the oldest compared to all others. This appears to substantiate the cluster analysis as this aligns with the widely accepted notion that younger age results in better TBI outcomes, even for higher levels of severity.

The cluster results seem to suggest that controlled blood sugar levels (as indicated by relatively low glucose levels) could be associated with better recovery outcome. C5 cluster have the most elevated blood sugar levels compared to the rest of the clusters. Though its initial GCS score (group mean: 11.2) and GOS-E at less than 30 days (group mean: 4.3) align closely with C3 (group means: GCS - 11.4, GOSE-E - 4.5), it lags in recovery compared to C3 (see Fig. 5 and Table 7). C3 and C5 differ significantly in their glucose levels (highest and lowest). The findings seem to suggest that poorly controlled glucose levels might blunt or delay recovery from brain injury. Clinically, elevated blood sugar levels are not good for the brain, as it increases the intracranial pressure. C3 cluster's faster paced recovery compared to C5 might be associated with the lower glucose levels.

Presence of IVH is indicative of severity of bleeding inside the brain. C1 cluster has the highest percentage (23.1%),

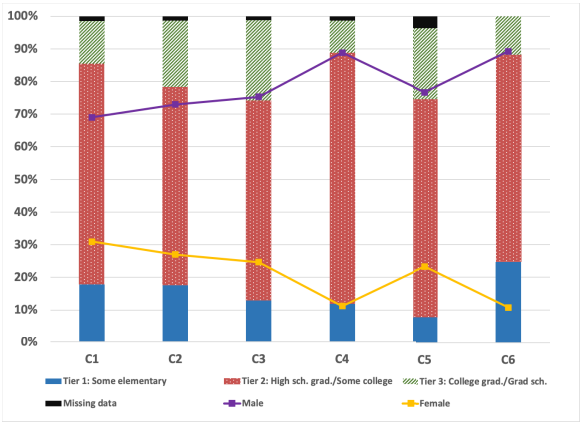


FIGURE 6. Distribution of education level and gender across clusters. The blue (top) line denotes the percentage of male per cluster while the yellow (bottom) line denotes likewise for female.

providing further evidence of the initial severity of injury of this subgroup while C6 has the lowest percentage (2.2%) indicating minimal trauma, as also verified by its highest initial GCS score (group mean: 11.7). SAH type quantifies the distribution of bleeding in the space surrounding the brain. SAH present in convexities/sulci and cisterns implies the most critical region. C5 (highest percentage of most severe type - 23.3%) differs significantly from C1 (17.9%) and C3 (12.39%) clusters. The findings also indicate that mesencephalic cisterns type and midline shift, which both quantify cerebral compression, do not play a significant role in delineating the clusters. This could be due to how these features are encoded in the cluster analysis. Midline shift (ms) is represented on an ordinal scale: none, 0 $<$ms$\leq$5mm, 5 $<$ms$\leq$10mm, ms$>$10mm. About 2.7% of the entire patient sample fell into the '$>$10mm' severity group. Likewise, mesencephalic cisterns type is also encoded as ordinal: normal (0), 'blood in cisterns but no compression' (1), 'effaced/compressed but visible' (2), 'obliterated/absent' (4). The obliterated/absent (most severe) group are about 6.8% of the overall sample size. The C1 cluster does contain the highest percentage of the most severe cases (8%) while C2 and C3 clusters both have the lowest percentage (4%). Future analysis using a binary encoding for these variables might result in a more meaningful assessment of their effect.

Regarding the highest sodium level and body temperature, the statistical testings suggests that there is some significance in the group means and distribution among the clusters. However, the domain experts viewed these differences as not clinically meaningful, as the values are within a uniform range. It might have been more useful to quantify these features using bins based on ranges rather than actual numerical values to reduce the system's sensitivity to small ranges of difference. Although standard statistical methods (Fig. 4, Table 6 and Table 7) have been employed to assist in the interpretation of the cluster ''meaning'', it is important to note that not all of the features that demonstrate statistical significance between clusters may provide clinically useful information since the

mean differences are very small. This is a recognized limitation of statistical significance, as determined by *p*-values, in that it does not measure the size of an effect and may not always translate to practically significant results [50], as in the case of the sodium and body temperature values.

Table 6 illustrates the outcome of the clusters as quantified by their performance on the cognitive and functional assessments at multiple time points. Given that education level could impact performance of patients on these type of assessment tests, we also examined the clusters in terms of education levels (Fig. 6) and other demographics that are not necessarily outcome measures. This includes gender and distribution of study treatment category (placebo vs. citicoline drug), since the sample is drawn from a drug trial study. Multiple testing corrections are done across all three variables. None of these demographics exhibited a significance difference across the clusters except for gender (Fig. 6). Nonetheless, educational level did vary across clusters. Interestingly clusters C2, C3, and C5 had more advanced education. Cluster 3 is somewhat different than the others in that it is characterized by higher educational level (Fig. 6), better outcome on the GOS-E (Table 7), and better scores on the PSI and CVLT-II (Table 6). Educational level has been frequently linked to better health outcomes in a variety of diseases [51]. The C5 cluster is the oldest group, age-wise. Thus, the more advanced education might be more reflective of the age of the population.

Revisiting Fig. 3(a), it is of interest to look at canonical variable 1 which combines F4 (mechanism of injury) and F29 (lowest hemoglobin). Lowest hemoglobin is indicative of overall blood loss during an injury which could suggest higher trauma. Along the x-axis of Fig. 3(a), the clusters sort out from left to right as C6-C3-C5-C2-C4-C1-C2 with C2 appearing twice along the axis. This ordering is of interest because C6 and C3 to the far left have the best outcomes on the GOS-E (Table 7) and have the lowest death rate and shortest median ICU stay times (Table 8). On the other hand, clusters C4 and C1 to the right hand side of x-axis of Fig. 3(a) have longer ICU median stays and high death rates (Table 8). If we consider the x-axis of Fig. 3(a) as representing patient ''criticality'' from lower to higher, than the bifurcation of cluster 2 may be interpreted as reflecting more heterogeneity in this cluster with some patients more ''critically'' ill and others less so. More importantly, ensemble clustering of this heterogeneous data set of TBI subjects allows the development of hypotheses as to how different subtypes of patients may differ in injury mechanism, severity, and outcome. Again, though the C1 cluster has a very high initial severity of presentation, they rebound the fastest, compared to C2, C4 and C5.

Primary limitations of this study are the relatively low number of features utilized as well as inconsistent patterns in unreported data for varied outcome measures used for assessing recovery progress of the patients. The clustering solution is a reflection of the selected features. If different features, such as biomarker levels [52] or neurological signs or symptoms, had been utilized, clustering solutions that differ both in size and composition would likely result. This is a key trait of unsupervised machine learning. Cluster quality is assessed by metrics (Table 1) that did not require ground truth labels for the subjects. Such measures are necessary when insufficient ground truth is available. Although these metrics are widely accepted to measure cluster quality, interpretation of cluster ''meaning'' remains part art and part science. The primary strength of this study is the participation of an interdisciplinary team of computational scientists, statisticians, a neurosurgeon, and a neuroscientist to ensure a robust statistical and clinical approach that confirmed validity of cluster methodology.

V. CONCLUSION

This paper presents an effective, validated, explainable framework, enabling the combination of automated data analytics with human domain expert knowledge. In this application, it was used to characterize TBI phenotype data for applicability to prognostic and diagnostic analysis. The results yielded six distinct patient subgroups with respect to mechanism of injury, severity of presentation, anatomy, psychometric, and functional outcomes. The findings suggest that younger age and controlled blood sugar levels may contribute to more favorable and quicker recovery trajectory regardless of a very severe initial presentation of injury. This enhanced ensemble statistical and clustering model is applicable to other disorders that exhibit significant heterogeneity.

ACKNOWLEDGMENT

(Dacosta Yeboah and Louis Steinmeister contributed equally to this work.) The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Leonard Wood Institute, the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] A. B. Peterson, L. Xu, J. Daugherty, and M. J. Breiding, ''Surveillance report of traumatic brain injury-related emergency department visits, hospitalizations, and deaths, United States, 2014,'' Center Disease Control Prevention, Atlanta, GA, USA, Tech. Rep., 2019.

[2] A. J. Masino and K. A. Folweiler, ''Unsupervised learning with GLRM feature selection reveals novel traumatic brain injury phenotypes,'' 2018, *arXiv:1812.00030*. [Online]. Available: http://arxiv.org/abs/1812.00030

[3] K. K. Wang, Z. Yang, T. Zhu, Y. Shi, R. Rubenstein, J. A. Tyndall, and G. T. Manley, ''An update on diagnostic and prognostic biomarkers for traumatic brain injury,'' *Expert Rev. Mol. Diag.*, vol. 18, no. 2, pp. 165–180, 2018.

[4] D. Cherezov, D. Goldgof, L. Hall, R. Gillies, M. Schabath, H. Müller, and A. Depeursinge, ''Revealing tumor habitats from texture heterogeneity analysis for classification of lung cancer malignancy and aggressiveness,'' *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019.

[5] 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.

[6] R. Xu and D. C. Wunsch, ''Clustering algorithms in biomedical research: A review,'' *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 120–154, 2010.

[7] 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.
[8] K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, T. N. Takahashi, S. Kanne, and D. Wunsch, "Ensemble statistical and subspace clustering model for analysis of autism spectrum disorder phenotypes," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 3329–3333.
[9] T. Obafemi-Ajayi, K. Al-Jabery, L. Salminen, D. Laidlaw, R. Cabeen, D. Wunsch, and R. Paul, "Neuroimaging biomarkers of cognitive decline in healthy older adults via unified learning," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–9.
[10] R. Xu and D. Wunsch, *Clustering*, vol. 10. Hoboken, NJ, USA: Wiley, 2009.
[11] R. D. Zafonte, E. Bagiella, B. M. Ansel, T. A. Novack, W. T. Friedewald, D. C. Hesdorffer, S. D. Timmons, J. Jallo, H. Eisenberg, T. Hart, and J. H. Ricker, "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.
[12] National Institute of Health. *Federal Interagency Traumatic Brain Injury Research (FITBIR)*. U.S. Department of Health & Human Services. Accessed: Jun. 2019. [Online]. Available: https://fitbir.nih.gov/
[13] L. Steinmeister, D. Yeboah, G. Olbricht, T. Obafemi-Ajayi, B. Hadi, D. Hier, and D. C. Wunsch, "Handling missing data for unsupervised learning with an application on a fitbir traumatic brain injury (TBI) dataset," in *Proc. Mil. Health Syst. Res. Symp. (MHSRS)*, 2020, doi: 10.13140/RG.2.2.22914.71361.
[14] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. Marquéz, B. Gruber, B. Lafourcade, P. J. Leit ao, T. Münkemüller, C. Mcclean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, pp. 027–046, 2013.
[15] F. John, *Applied Regression Analysis and Generalized Linear Models*. Newbury Park, CA, USA: Sage, 2015.
[16] J. Fox and G. Monette, "Generalized collinearity diagnostics," *J. Amer. Stat. Assoc.*, vol. 87, no. 417, pp. 178–183, Mar. 1992.
[17] T. A. Craney and J. G. Surles, "Model-dependent variance inflation factor cutoff values," *Qual. Eng.*, vol. 14, no. 3, pp. 391–403, Mar. 2002.
[18] D. Belsley, "A guide to using the collinearity diagnostics," *Comput. Sci. Econ. Manage.*, vol. 4, pp. 33–50, Feb. 1991. [Online]. Available: https://link.springer.com/article/10.1007/BF00426854
[19] 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.
[20] 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.
[21] H. S. Yoon, S. Y. Ahn, S. H. Lee, S. B. Cho, and J. H. Kim, "Heterogeneous clustering ensemble method for combining different cluster results," in *Proc. Int. Workshop Data Mining Biomed. Appl.* Berlin, Germany: Springer, Apr. 2006, pp. 82–92.
[22] L. Zheng, T. Li, and C. Ding, "Hierarchical ensemble clustering," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 1199–1204.
[23] D. Huang, J. Lai, and C.-D. Wang, "Ensemble clustering using factor graph," *Pattern Recognit.*, vol. 50, pp. 131–142, Feb. 2016.
[24] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1460–1473, May 2018.
[25] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.
[26] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[27] H. Zeng and Y.-M. Cheung, "Iterative feature selection in Gaussian mixture clustering with automatic model selection," in *Proc. Int. Joint Conf. Neural Netw.*, Aug. 2007, pp. 2277–2282.
[28] T. Ronan, S. Anastasio, Z. Qi, R. Sloutsky, K. M. Naegle, and P. H. S. V. Tavares, "Openensembles: A python resource for ensemble clustering," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 956–961, 2018.
[29] F. Reid, A. McDaid, and N. Hurley, "Percolation computation in complex networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 274–281.
[30] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, p. 814, Jun. 2005.
[31] R. Ghaemi, M. N. Sulaiman, H. Ibrahim, and N. Mustapha, "A survey: Clustering ensembles techniques," *World Acad. Sci. Eng. Technol.*, vol. 50, pp. 636–645, Feb. 2009.
[32] 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.
[33] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, no. 4, pp. 817–838, 1980.
[34] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandin. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
[35] W. Haynes, *Holm's Method*. New York, NY, USA: Springer, 2013, p. 902.
[36] C. Feng, H. Wang, N. Lu, T. Chen, H. He, Y. Lu, and X. M. Tu, "Log-transformation and its implications for data analysis," *Shanghai Arch. Psychiatry*, vol. 26, no. 2, pp. 105–109, 2014.
[37] L. Xu, P. Yan, and T. Chang, "Best first strategy for feature selection," in *Proc. 9th Int. Conf. Pattern Recognit.*, May 1988, pp. 706–708.
[38] Y. Kim, W. N. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Boston, MA, USA, 2000, pp. 365–369.
[39] F. C. G. López, M. G. Torres, J. A. M. Pérez, and J. M. M. Vega, "Scatter search for the feature selection problem," in *Proc. Conf. Technol. Transf.*, Springer, 2003, pp. 517–525.
[40] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *Proc. 38th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2015, pp. 1200–1205.
[41] I. Jo, S. Lee, and S. Oh, "Improved measures of redundancy and relevance for mRMR feature selection," *Computers*, vol. 8, no. 2, p. 42, May 2019.
[42] Y. Fujikoshi, V. V. Ulyanov, and R. Shimizu, *Multivariate Statistics: High-Dimensional Large-Sample Approximations*, vol. 760. Hoboken, NJ, USA: Wiley, 2011.
[43] E. Bagiella, T. A. Novack, B. Ansel, R. Diaz-Arrastia, S. Dikmen, T. Hart, and N. Temkin, "Measuring outcome in traumatic brain injury treatment trials: Recommendations from the traumatic brain injury clinical trials network," *J. Head Trauma Rehabil.*, vol. 25, no. 5, p. 375, 2010.
[44] E. A. Wilde, G. G. Whiteneck, J. Bogner, T. Bushnik, D. X. Cifu, S. Dikmen, L. French, J. T. Giacino, T. Hart, J. F. Malec, and S. R. Millis, "Recommendations for the use of common outcome measures in traumatic brain injury research," *Arch. Phys. Med. Rehabil.*, vol. 91, no. 11, pp. 1650–1660, 2010.
[45] *California Verbal Learning Test (CVLT)*. Accessed: Dec. 2019. [Online]. Available: http://dictionary.apa.org/california-verbal-learning-test
[46] *Digit Span*. Accessed: Dec. 2019. [Online]. Available: https://www.cambridgebrainsciences.com/science/tasks/digit-span
[47] J. E. Kennedy, P. F. Clement, and G. Curtiss, "WAIS-III processing speed index scores after TBI: The influence of working memory, psychomotor speed and perceptual processing," *Clin. Neuropsychologist*, vol. 17, no. 3, pp. 303–307, Aug. 2003.
[48] *Controlled Oral Word Association Test (COWAT)*. Accessed: Dec. 2019. [Online]. Available: https://healthabc.nia.nih.gov/sites/default/files/COWATOMY16_0.pdf
[49] G. L. Gadbury, C. S. Coffey, and D. B. Allison, "Modern statistical methods for handling missing repeated measurements in obesity trial data: Beyond LOCF," *Obesity Rev.*, vol. 4, no. 3, pp. 175–184, Aug. 2003.
[50] R. L. Wasserstein and N. A. Lazar, "The ASA statement on *p*-values: Context, process, and purpose," *Amer. Statistician*, vol. 70, no. 2, pp. 129–133, 2016, doi: 10.1080/00031305.2016.1154108.
[51] R. A. Hahn and B. I. Truman, "Education improves public health and promotes health equity," *Int. J. Health Services*, vol. 45, no. 4, pp. 657–678, Oct. 2015.
[52] Z. Roy, S. Subhash, L. A. Bui, B. Hadi, D. B. Hier, D. Wunsch, G. R. Olbricht, and T. Obafemi-Ajayi, "Exploratory analysis of concussion recovery trajectories using multi-modal assessments and serum biomarkers," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 5514–5518.

DACOSTA YEBOAH received the B.Sc. degree in electrical and electronics engineering from the Kwame Nkrumah University of Science and Technology, Ghana, in 2018. He is currently pursuing the master's degree with the Computer Science Department, Missouri State University (MSU). He is also a member of the Computational Learning Systems (CLS) Laboratory, MSU. His research interests are in machine learning and robotics.

LOUIS STEINMEISTER received the B.Sc. degree in business mathematics from the University of Hamburg, having emphasized in finance, stochastic processes, and statistics, in 2016, and the M.Sc. degree in applied mathematics from the Missouri University of Science and Technology (Missouri S&T), through a joint program with Ulm University, Germany, with an emphasis in mathematical statistics and machine learning. He is currently pursuing the Ph.D. degree in statistics with Missouri S&T. His research interests are in algorithmic trading, reinforcement learning, machine learning, financial mathematics, and mathematical statistics. He is a member of the Applied Computational Intelligence Laboratory, Missouri S&T.

DANIEL B. HIER is currently an Adjunct Professor of Electrical and Computer Engineering with the Missouri University of Science and Technology (Missouri S&T) and a Professor Emeritus of Neurology and Rehabilitation with the University of Illinois at Chicago (UIC). He is also a Faculty Member of the ACIL Research Group. He had previously served as a Physician Executive at the Cerner Corporation, Kansas City, MO, USA, and the Head of the Neurology and Rehabilitation, UIC. He is board-certified in both neurology and medical informatics.

BASSAM HADI received the B.Sc. degree in physics/math minor from the Grainger College of Engineering, University of Illinois, in 1989, and the M.D. degree from the St. Louis University School of Medicine, in 1994. He is currently a board-certified Neurosurgeon at Mercy Clinic South, St Louis, MO, USA, where he also serves as the Department Chair of Surgery. His Neurosurgery Residency was at the University of Louisville in 2000.

DONALD C. WUNSCH II (Fellow, IEEE) received the B.S. degree in applied mathematics from the University of New Mexico, Albuquerque, NM, USA, the M.S. degree in applied mathematics from the University of Washington, Seattle, WA, USA, the Ph.D. degree in electrical engineering from the University of Washington, the Executive MBA degree from the Washington University in St. Louis, and Jesuit Core Honors Program from Seattle University. He is currently the Mary K. Finley Missouri Distinguished Professor at the Missouri University of Science and Technology (Missouri S&T). He is also the Interim Director of the Intelligent Systems Center and the Director of the ACIL Research Group. He is the previous International Neural Networks Society (INNS) President, INNS Fellow, NSF CAREER Awardee, 2015 INNS Gabor Award recipient, and 2019 Ada Lovelace Service Award Recipient.

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

TAYO OBAFEMI-AJAYI (Member, IEEE) received the B.S. degree in electrical engineering, the M.S. degree in electrical engineering, and the Ph.D. degree in computer science from the Illinois Institute of Technology. She is currently an Assistant Professor of Electrical Engineering in the Engineering Program with the Michigan State University (MSU), where she is also the Director of the Computational Learning Systems Laboratory, and a Faculty Member of the ACIL Research Group, Missouri University of Science and Technology (Missouri S&T). Her research interests include machine learning, data mining, biomedical informatics, and control systems.