IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 26, 2021, accepted March 22, 2021, date of publication March 30, 2021, date of current version April 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3069684

Cluster Analysis of Mixed and Missing Chronic Kidney Disease Data in KwaZulu-Natal Province, South Africa

PETER A. POPOOLA[1], JULES-RAYMOND TAPAMO[2], (Member, IEEE), AND ALAIN G. ASSOUNGA[3,4]

[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus, Durban 4041, South Africa
[2]School of Engineering, University of KwaZulu-Natal, Howard College Campus, Durban 4041, South Africa
[3]Department of Nephrology, Nelson Mandela School of Medicine, University of KwaZulu-Natal, Durban 4001, South Africa
[4]Inkosi Albert Luthuli Central Hospital, Durban 4091, South Africa

Corresponding author: Jules-Raymond Tapamo (tapamoj@ukzn.ac.za)

ABSTRACT Real-world datasets, particularly Electronic Health Records, are routinely found to be mixed (comprised of both categorical and continuous variables) and/or missing in nature. Such datasets present peculiar challenges related both to their clustering and the evaluation of the clusterings obtained. In this paper, we discuss these challenges in detail, as well as the solution approaches applied to them in the literature. We then apply some of these approaches to a multi-racial Chronic Kidney Disease (CKD) dataset comprising of 20 continuous and 12 categorical variables with an over 30% missingness ratio, evaluating our results through external and internal validation as well as cluster stability testing. From the results of our study, the Ahmad-Dey distance measure consistently outperformed Gower's distance on our mixed and missing dataset. In addition, our results show that advanced imputation methods like multiple imputation, which take into consideration the uncertainty inherent in imputation, should be explored when clustering missing datasets. Three clusters were identified from our dataset which were significantly differentiated by age, sex, estimated Glomerular Filtration Rate (eGFR), creatinine, urea, and hemoglobin, but not by race or blood pressure. The fact that, through proper cluster analysis, we were unable to identify five clusters corresponding to the five CKD stages usually used to classify CKD patients indicates that datasets with more than the usual four/six variables used for computing eGFR may contain a latent structure different from this five-group structure, the identification of which will provide valuable insights peculiar to each cohort for medical practitioners.

INDEX TERMS Chronic kidney disease, cluster analysis, electronic health records, missing data, mixed data, Gower's distance, Ahmad-Dey distance.

I. INTRODUCTION

Data clustering is an important approach which can be used to, in an unsupervised way, decipher some inherent and practically relevant structure in a dataset. With more than 404,000 documents related to cluster analysis in the literature [1], this approach has been widely used to extract meaning from data in the scientific literature over the years, with applications ranging from data compression, tumor categorization, video segmentation, and recommender systems to the grouping of galaxies by their shape [2]. In simple terms, clustering is the segmentation/partitioning of a dataset into groups. Two major desirable characteristics of the resultant grouping are homogeneity, which maximizes intra-cluster compactness or similarity, and heterogeneity, which seeks to achieve as much separation as possible between clusters. In essence, we desire to partition a body of data into groups that are as distinct as possible.

Clustering real-world datasets poses several challenges because they do not present in as ideal a format as would make for easy clustering. Apart from complexities introduced by high dimensionality, noise, and outliers, missingness and a combination of multiple variable types within a single dataset are significant challenges that are routinely encountered when clustering real-world datasets. Though several clustering algorithms have been proposed in the literature to

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Salehzadeh-Yazdi.

handle missing or mixed datasets, it is more difficult to find approaches which are designed to properly cluster data which is both mixed and missing.

An important aspect of data clustering is cluster evaluation, where we seek to either obtain some level of confidence regarding the reliability of the results obtained from cluster analysis (validation), or select the most suitable algorithm from a set of algorithms available for a given problem (comparison) [2]. Cluster evaluation of mixed and missing datasets also has its unique challenges. A widely-used approach to clustering mixed and/or missing data in the literature is the application of mixture models, which view the dataset as being comprised of a mix of multiple parameterized probability distributions which are seen to correspond to clusters, with the major aim of discovering the parameters which describe these distributions [3]. These approaches do not inherently apply a sense of pairwise distance per se as is the case with most internal validation measures, and thus, performing internal validation on their results is impractical. A special class of validation measures, e.g. the Bayesian Information Criterion and Akaike Information Criterion [4], which take advantage of the probabilistic nature of the distributions, can be used to evaluate these methods, as is usually done in the literature. However, this does not help, either, as the question of fairness in comparing them with internal validation scores obtained for other algorithms based on the computation of Euclidean or other distances arises [5].

Electronic Health Record (EHR) data are prime examples of data which can usually be found to be both missing (especially due to attrition in longitudinal studies) and mixed – containing both demographic information like sex, age, and weight, and laboratory measurement or treatment intervention indicators, which are either categorical or continuous in nature. Thus, a bid to properly and reliably perform cluster analysis on such datasets pits the researcher against the challenges earlier alluded to. In this paper, we discuss and apply some solutions existing in the literature to tackle these challenges, particularly with regards to our Chronic Kidney Disease (CKD) dataset, presenting some comparative analyses of a number of these approaches, and insights which can be drawn from them. Based on the outcome of our comparisons of these methods, we present analyses of the clustering obtained. This paper is structured as follows: Section II presents a review of the approaches existing in the literature for clustering mixed, missing, and mixed and missing datasets, cluster evaluation for mixed and missing data, as well as some studies which have applied cluster analysis to EHR data and the results they obtained. Given the wide range of clustering and cluster evaluation approaches available in the literature [2], Section III presents our considerations and justifications for the selections we made in our specific CKD case study. We also present an overview of our dataset in this section. In Section IV, we present the results of our experiments and a discussion of our inferences from them. Section V presents a conclusion of the study, its implications, and suggestions for future research.

II. LITERATURE REVIEW

A. CLUSTERING MIXED DATA

Mixed data clustering is a highly active field of research, given that it presents unique challenges due to the presence of categorical variables, which, being non-numeric, do not lend themselves to arithmetic operations like addition, subtraction, squares/square roots, etc. These operations are integral to the computation of Euclidean or other distances. Put another way, categorical variables convey value/meaning in a non-numerical sense, or at least, in a sense which is difficult to quantify numerically. Thus, as Christian Hennig puts it, ‘‘Euclidean intuition is irrelevant in . . . clustering problems with categorical variables or non-Euclidean dissimilarities [5, p. 58]. This immediately discountenances the naïve practice of simply converting categorical variable values to integers and treating them as numerical, one which is not uncommon, at least not among those who are less statistically inclined. Thus, the major and widely-known methods for approaching this challenge in the literature include either converting variables from categorical to continuous or continuous to categorical, developing new distance measures suitable for mixed data, and applying mixture models [3], [6].

1) VARIABLE CONVERSION

Converting continuous variables into categorical ones generally involves binning into a pre-determined number of representative categories, which then allows for the application of clustering methods developed for categorical data such as CACTUS [7], Squeezer [8], Clarke *et al.*’s [9] ensemble method, and a plethora of others [10]. However, this form of discretization results in loss of information and poses a new non-trivial challenge of selecting an appropriate discretization scheme, a critical choice that directly determines the resulting similarity matrix and/or clustering [3]. One of the most popular methods of converting from categorical to continual variables is dummy coding, where each category is converted into a binary variable, with a value of 1 indicating the allocation of that category level to a record, and zero, otherwise. This is also known as one-hot-encoding in the machine learning community. Once this is done, the entire data is (optionally) scaled, and Euclidean intuition can be used. However, Foss *et al.* [11] show through theoretical calculations and simulations that apart from the high dimensionality and consequent high computational burden this introduces to the clustering process, the domain-specific meaning conveyed by such transformed variables is lost. Furthermore, dummy-coded variables tend to get dominated by continuous variables during clustering, and no weighting scheme can overcome this generally. Another conversion approach which has been proposed and used in the literature to overcome some of these shortcomings leverages dimensionality reduction. In some cases, dimensionality reduction is performed solely on the categorical variables, after which the resulting principal components (now numerical) are combined with the continuous part of the mixed data, and clustering is done. In other cases, the entire mixed dataset undergoes dimensionality reduction, and the resulting top d components are selected and used for clustering. This approach, where dimensionality reduction is done as a separate process preceding clustering, is known as the tandem approach [3]. For example, Factor Analysis of Mixed Data [12] can be applied as a dimensionality reduction technique, followed by k-means, hierarchical clustering, etc. However, apart from the fact that the selection of d presents a separate optimization problem given that it invariably determines the cluster results, this approach leads to a problem known as cluster masking. The cluster masking problem arises because each of these two steps – dimensionality reduction and clustering – optimize a different objective, and in some cases, dimensionality reduction hides the underlying cluster structure. To address this problem, solutions have been proposed in the literature which optimize a combination of these two objectives [13]. However, it is important to note that both approaches depend heavily on the amenability of the dataset to dimensionality reduction where a few principal components account for a high percentage of variability in the dataset. Where this is not the case, these methods cannot be reasonably applied.

2) HYBRID DISTANCE MEASURES

The extension of distance measures to accommodate mixed data is an area of active research in the literature, as "finding an appropriate similarity measure and cost function to handle mixed data remains a challenge in partitional clustering algorithms." [6, p. 11]. One of the most popular of such hybrid distance measures is Gower's distance [14], [15], which is widely used for clustering mixed datasets in the literature, and is given as follows:

$$\delta_G(X, Y) = \frac{\sum_{j=1}^{m} w_j f_j(x_j, y_j)}{\sum_{j=1}^{m} w_j} \tag{1}$$

where $f_j(x_j, y_j) = |x_j - y_j|/r_j$ if j is continuous (r_j being the sample range of variable j), and simple matching if j is categorical, m is the number of variables, and w_j is the weight associated with each variable.

As can be seen from (1), Gower's distance measure computes distances for continuous and categorical variables separately, summing the results. In simple matching, 1 is returned when the categorical values are the same, and 0 otherwise. The weight, w_j is user-defined both to help balance the contributions of both continual and categorical variables, and to implement any expert knowledge regarding variable importance. A key weakness associated with this weighting approach in general is that the choice of optimal weights is a difficult/impossible one, in addition to the fact that, in the case of Gower's distance in particular, wrongly-specified weights by the user will invariably lead to inaccurate clustering results, given the fact that the weights strongly affect the clustering outcome. This is more evident for categorical variables, where, being multiplied by 1, the weight is essentially the variable value. Foss *et al.* [15] also point out that when Gower's distance is combined with Partitioning Around Medoids (PAM), the distance function produces unchanging values even with changing cluster-specific underlying categorical level probabilities. They also show that the distance is dominated by the weight, even in cases where the weight gives no information about cluster membership.

Huang's [16] distance measure for mixed datasets was developed to be used in conjunction with the k-prototypes clustering algorithm. Being a centroid-based method, distance/similarity is measured from each data point to its pre-determined cluster prototype, and is given as follows:

$$\delta_H(X_i, Q_l) = \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{j=1}^{m_c} (x_{ij}^c, q_{lj}^c) \tag{2}$$

where X_i and Q_l are the ith data point and the prototype for cluster l, respectively, m_r and m_c are the number of continuous and categorical variables, respectively, γ_l is a weight for cluster l's categorical variables, q_{lj}^r is the mean, and q_{lj}^c, the mode value for attribute j in cluster l.

In a similar fashion to Gower's distance, Huang's distance computes continuous and categorical distances separately and sums the results, using Euclidean distance for continuous variables, and simple matching for categorical variables. Notably, the weight, $\gamma_{l,}$ is here applied on a cluster-by-cluster, rather than variable-by-variable basis, as is the case with Gower's distance. γ_l is also automatically computed as part of the k-prototype algorithm. This is probably aimed at addressing some of the weaknesses of Gower's distances mentioned earlier. However, in doing so, the flexibility of portraying and specifying varying variable importance is lost, and this is more so given that the weights are only associated with categorical variables. Ahmad and Dey [17] also point out another weakness associated with the use of the mode as the prototype for categorical variables, explaining that it leads to loss of information, especially in cases where the differences in frequency between categorical levels are small. In addition, since distances are computed with respect to cluster prototypes, a square distance matrix reflecting pairwise distances between all data point cannot be obtained from associated k-prototype algorithm, leading to difficulties in applying internal validation indices which generally require a distance matrix. However, taken alone, the distance function can be easily modified to replace cluster centroids with another data point, making pairwise distance computations achievable. Modha and Spangler [18] proposed a distance measure which sought to address the issue of balancing the contribution between categorical and continuous variables through a brute-force approach which adaptively assigns weights to either variable type based on the quality of their underlying cluster structure, that is, their contribution to the separation and compactness of the resulting clusters. Their algorithm and its associated mathematical formulations are summarized in [15]. Though Foss *et al.* [15] assert that in most cases, the Modha-Spangler approach is able to balance the contributions of both variable types, leading to great results, Ahmad and Dey [17] point out the computational burden associated with the method due to its brute force approach. Foss and Markatou [19] also identify another weakness of the method which lies in its inability to up- or down-weigh individual variables, given that it uses a single weight value to balance categorical vs continuous variable contributions. A number of other hybrid distance measures are highlighted in [15] and [17], but we lastly discuss Ahmad and Dey's [17] distance measure, due to its unique approach to computing categorical distances based on their co-occurrence. Their distance function, an extension of Huang's distance measure is as follows:

$$\varphi = \sum_{i=1}^{n} \delta_A(d_i, C_j) \tag{3}$$

where

$$\delta_A(d_i, C_j) = \sum_{t=1}^{m_r} (w_t(d_{it}^r - C_{jt}^r))^2 + \sum_{t=1}^{m_c} \Omega\left(d_{it}^c, C_{jt}^c\right)^2 \tag{4}$$

where m_c and m_r are the number of categorical and continuous variables, respectively, C_j is the closest cluster center to d_i and w_t is a weight associated with each continuous variable.

Though the Ahmad-Dey distance measure adopts a very similar approach to Huang's in that distances are computed from cluster centers and categorical and continuous distances are computed separately, a key difference is that they do not compute categorical distances as simple binary matching. Categorical distances are computed as a function of the overall distribution of categorical values in a variable, as well as their co-occurrences with other categorical variable values. By so doing, they provide a richer value spectrum for categorical variables, removing the narrow limitations of binary values. Similarly, the weights for numeric variables are automatically computed based on value co-occurrence, though this computation necessitates the discretization of the continuous values. The authors of the algorithm emphasize, though, that this discretization only occurs during weight computation, and the original continuous values are used in computing squared Euclidean distance. The idea for the use of co-occurrence in computing categorical distance, which was taken from Stanfill and Waltz [75], has also been applied in [7]. One weakness of this approach, however, is the computational burden associated with computing co-occurrence for all combinations of categorical values and variables.

3) MIXTURE MODELS

As has been earlier mentioned, a statistical approach to mixed data clustering involves viewing the dataset as a mixture of parametric distributions, also known as a finite mixture model. That is, assumptions are made about underlying/latent distributions in the data, and methods (usually variants of EM) are used to estimate the parameters that describe those distributions, each of which corresponds to a cluster [20]. Upon convergence, these methods produce estimations of cluster membership probability for each data point [21], which can then be converted into crisp cluster memberships by assigning each data point to the cluster with the maximum associated conditional probability. These mixtures are generally modelled as

$$f\,(x_i, \phi) = \sum_{k=1}^{K} \pi_k f_k(x_i; \theta_k) \tag{5}$$

where $\sum_{k=1}^{K} \pi_k = 1$, $0 < \pi_k \leq 1 \forall k$, θ_k is the parameter vector, $\phi = (\theta, \pi)$, and K is the number of distributions.

In (5) above, π_k will eventually represent the probability of data point x_i being in cluster k upon convergence. Viewing the data as a mixture of homogenous distributions allows both simple and complex components to be modeled. In addition, mixture models are usually able to achieve a good balance between continuous and categorical variables. However, such methods can perform poorly when parametric assumptions are violated. Moreover, some of the weaknesses of the EM algorithm like its tendency to get stuck on local optima and the possibility of intractable integrals in the E step [15] are also associated with this approach, given that EM is its most widely used method. A number of studies and a review on finite mixture models are listed in [21], and studies applying finite mixture models to clustering mixed data in particular are reviewed in [6]. We briefly highlight two which are somewhat popular in the literature. The KAy-means for MIxed LArge datasets (KAMILA) algorithm borrows ideas from the classical k-means algorithm as well as Gaussian-multinomial mixture models [11]. Thus, it is able to balance categorical and continuous variable contributions and avoid the weaknesses of the k-means algorithm without making the strong parametric assumptions required by mixture models. It achieves the relaxation of these assumptions by computing its density estimator directly from the data. From the extensive tests carried out on both simulated and real-world datasets by the authors, they showed that their algorithm performs well on elliptical and non-elliptical data, and was the only algorithm to perform well across all test conditions, outperforming the Modha-Spangler method, which failed to achieve a good balance between categorical and conditional variables in some cases. Another popular clustering algorithm which applies mixture modelling is McParland and Gormley's [22] ClustMD, which consists of a suite of six latent variable mixture models with varying levels of parsimony achieved by imposing varying constraints on the model parameters, all of which are assumed to be Gaussian – including categorical variables, which are assumed to arise from a latent Gaussian continuous model. The method also assumes diagonality of the covariance matrix for the models, implying conditional independence of the variables. This assumption, generally known as the local independence assumption, is universal to mixture models involving both categorical and continuous data [20]. Apart from the fact that this assumption imposes restrictions on the generalizability of the model, the ClustMD model's computational efficiency decreases rapidly with the addition of more categorical variables. They performed simulations to evaluate the performances of the six parsimonious models over a range of component numbers, and the VII model with two components, which was derived by constraining the sum of all cluster volumes associated with categorical (nominal) variables to 1, outperformed all others. No explanation was provided for the meaning of the acronyms like VII and EVI used to represent the six models. When tested on a real-life prostate cancer dataset, the EVI model performed best. The formulation of the EVI model, as well as the other four, is detailed in their report.

B. CLUSTERING MISSING DATA

Missingness in a dataset has been classified in the literature as Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) [23]. The MCAR missingness pattern describes missingness which is independent of observed or missing data points within the dataset. Thus, any observation is as likely to be missing in one variable as the other [24]. The MAR missingness pattern is a step less restrictive than MCAR, in that it assumes that the missingness in the dataset is dependent only on the observed data. For a dataset to be MNAR, however, the probability of missing values occurring in a particular variable must be dependent on those same unobserved values [25]. In practice, the MCAR condition is more difficult to prove, given its strong assumptions. On the other hand, the MNAR condition, though requiring the weakest assumptions of the three, is much more difficult to handle, given that the underlying probability model for missingness must first be found, and that the MNAR condition itself can be difficult to prove, since it depends on values which are not observed [25]. MAR is, thus, the condition which lies reasonably between these two in practical applications, and there are ways through which its assumptions can be made even more plausible from the data [26]. More so, the MAR assumption subsumes the MCAR assumption. This has led most state-of-the-art missingness treatment algorithms to hold MAR as an assumptive basis for their computations [25].

The various missingness treatment methods in the literature can be broadly classified as either deletion, or imputation, or direct estimation methods. The two known deletion methods, complete-case analysis and pair-wise deletion are strongly discouraged in the literature, as at best, they lead to a reduction in statistical power, and at worst, introduce serious bias into analysis results. This is due to the fact that they depend on the MCAR assumption. Complete-case analysis, in particular, is almost always violated with EHR data [27], and thus, is almost guaranteed to introduce bias into statistical/research inference. A plethora of imputation-based methods exist in the literature, including single and multiple imputation methods. However, the most popular and ''state-of-the-art'' method is Multiple Imputation (MI), due to its relative ease of implementation, as well as its modelling of the uncertainty that is inherent in data imputation [28]. MI is carried out in three major steps: imputation, analysis, and pooling. In brief, these steps involve the imputation of m independent values for each missing data point, resulting in m imputed datasets; the individual and independent analysis of each of these datasets to obtain the desired statistic; and the subsequent pooling/combination of the m results into one final desired analysis result/statistic [29]. It is the independent analyses and subsequent pooling of analysis results that distinguish MI from SI as taking into consideration the uncertainty associated with imputed values.

In the context of clustering, there are two major approaches to handling missingness: one which treats missingness first and then carries out clustering on the resultant complete dataset, and another which merges the two problems into one, performing both clustering and missingness treatment simultaneously. These two approaches can be referred to as multi-stage and direct ways of clustering missing data, respectively.

1) MULTI-STAGE CLUSTERING

Under multi-stage approaches, Zhang and Fang [30] carried out a study which showed the superiority of MI over SI or non-imputation in fuzzy clustering accuracy, while Goel and Tushir [31] showed the superiority of linear interpolation (single) imputation combined with the incorporation of the Mahalanobis distance in the clustering step over some other fuzzy clustering approaches. In a similar vein, Tuikkala *et al.* [32] had earlier reported the superiority of advanced imputation techniques to basic ones like mean imputation in clustering gene expression data. However, Souto *et al.* [33] later argued that from their experiments, this superiority is non-existent, backing up their conclusion with the rationale that gene expression being highly correlated and characterized by very close values, imputing with a mean will have minimal effect on the shape of the data's distribution. Løkse *et al.* [34] introduced a new kernel function which learns the similarities between data points from the data's fitted mixture models, inherently taking care of the missing value problem. They then use this kernel function for spectral clustering, performing k-means clustering on the spectral clustering output. They show that their approach outperforms other baseline imputation methods in clustering accuracy. Finally, Yu *et al.* [35] use optimally designed variational encoder networks and high-order fuzzy c-means to first perform clustering, after which missing values are recovered from the clustering results.

2) DIRECT CLUSTERING

Direct approaches to clustering in the face of missingness can be divided into those that extend existing methods like k-means, and those that modify existing distance measures to produce novel ones which are designed to compute distances on missing data. Among the extension approaches, Chi *et al.* [36] introduced k-POD, a method which leverages an underlying assumption of k-means that every member of a cluster is a noisy instance of the cluster centroid to develop a simple majorization-minimization algorithm which, in a manner similar to expectation-maximization, iteratively improves both imputation and clustering quality. They show from their experimentations that k-POD produces comparably accurate (externally validated) clustering results with much faster times than multi-stage approaches which combine imputation and clustering time, and that this advantage becomes more evident with higher dimensionality and missingness rates. They note, however, that, given that it is a k-means-based algorithm, k-POD has the same weaknesses that k-means. The essence of this statement also applies to all extension methods in this context – they generally possess the same weaknesses as the original/classical methods they extend to handle missingness. A similar approach is used by Li *et al.* [37], who, in a bid to overcome the uncertainty associated with imputing missing values, modify the clustering objective function to incorporate a representation of missing data as intervals. They then incorporate an adversarial factor into the k-means and k-medoids clustering algorithms to produce 'robust' clustering decision makers. They show that their algorithm performs better than other clustering approaches which they tested on some datasets. Wang *et al.* [38] propose an extension to k-means similar to k-POD which iteratively re-computes cluster centroids and fills in missing values from their corresponding centroid values. They report that their method outperformed other multi-stage approaches involving zero and mean imputation, Expectation Maximization (EM), and KNN filling.

Lithio and Maitra [39] defined a new distance measure for the k-means algorithm which ignores missing feature values, forming a new clustering algorithm called k_m-means. Their comprehensive simulation experiments showed that k_m-means was significantly faster than k-POD in almost all instances, while maintaining a high level of accuracy. Their algorithm also performed better than two-stage clustering methods where EM and MI were first used for imputation, followed by k-means for clustering. Datta *et al.* [40] propose a modification to the Euclidean distance measure for missing data called Feature Weighted Penalty Based Dissimilarity where a penalty weight is assigned to each missing feature. They then use the new distance measure both on k-means and hierarchical clustering, showing that their method outperforms existing two-stage imputation-based methods, especially in the sense that it performs well with all missingness mechanisms, whereas each two-stage method they tested was only able to perform well on some specific missingness mechanisms. Finally, AbdAllah and Shimshoni [41] modify the classical Euclidean distance for missing data to compute distances based on mean and variance for MCAR, and conditional mean and variance for MAR and MNAR. They also extended the centroid/mean computation required for k-means to accommodate missing values, and proposed a way of integrating their new distance measure into mean shift clustering. They reported that their algorithms outperformed all compared algorithms in most cases on six datasets from the Signal and Image Processing Unit.

From these reviewed studies, we note that in most, if not all cases, external validation, specifically the Adjusted Rand Index (ARI), was the method used to compare clusterings. Even algorithms which extended classical distance measures, and hence, would be able to evaluate their results using internal validation, chose to use external validation. This may simply be due to the need to compare with methods which do not make their distance measures externally available (where they are used for clustering) or with classical methods like k-means which do not accept distance matrices as input. It is also worthy of note that most direct clustering methods are compared with, and shown to outperform, multi-stage methods in terms of computational efficiency. This is an expected outcome considering that, generally, multi-stage methods inherently involve more computation than direct methods.

C. CLUSTERING MIXED AND MISSING DATA

As is evident from the preceding sections, a significant amount of research effort has been devoted to clustering data which is either solely mixed, or solely missing. However, at least not as much attention has been given to scenarios – which are fairly common – where the dataset to be clustered both has missing values and is made of a mixture of categorical and continuous variables. Most methods which can cluster mixed data do not accept missing data, and most methods which cluster missing data directly do not accept mixed datasets. This could be, in part, due to the fact that these two challenges can be handled individually and sequentially, with missingness being first dealt with, and the resulting complete dataset used as an input to methods which cluster mixed data. However, this multi-stage approach robs the researcher of the advantages associated with missingness clustering treatment methods which do not directly impute values into the dataset, while also subjecting them to the disadvantages associated with multi-stage methods, major among which is high computational complexity. At the implementation level, however, some of the more popular mixed data clustering methods earlier discussed have been made amenable to missing data. For example, in **R** [42], the *daisy* function of the ***cluster*** [43] package handles missing values in its Gower's distance implementation by assigning a weight of 0 to variables which are missing in the ith distance computation (a form of pairwise deletion or available case analysis). This approach is also used in the ***kproto*** function of the ***clustMixType*** [44] package which implements Huang's distance measure and the k-prototypes algorithm. As has been earlier pointed out, this approach has been shown in the literature to have the potential of introducing bias into the data analysis process.

From the previous sections, it will be noticed that one method which is common to both mixed and missing data clustering is EM. An algorithm which takes advantage of this phenomenon is MixAll [45], which applies mixture models to cluster categorical-only, continuous-only data, and mixed data, utilizing EM both for estimating the mixture models, as well as for imputing missing values. It provides support for Gaussian, Poisson, and Gamma distributions. The package vignette explains the algorithm in more detail. Similarly, the study by Revillon and Mohammad-Djafari [46], achieves the clustering of both mixed and missing data through the use of mixture models. They handle missing data "by taking advantage of properties of the multivariate normal distribution to obtain a distribution for missing values" [46, p. 3]. They applied their approach to the clustering and classification of radar emitters for electronic warfare, comparing its results with those of k-means, k-nearest neighbors (kNN), random forests (RF), and neural networks (NN). In addition, varying percentages of missingness were introduced into the dataset as part of the evaluation process. Specifically, the clustering results of their algorithm was compared with those of a multi-stage clustering process involving mean/mode imputation followed by k-means, and cluster number selection was done using a lower bound measure which was introduced by the authors, as well as Average Silhouette Width (ASW). However, they did not specify how they were able to apply k-means, which only classically works on continuous data, to the mixed dataset. Be that as it may, they report that their approach outperforms the multi-stage approach at higher levels of missingness in the dataset.

From the above review, it is clear that a plethora of clustering methods are available in the literature, each with its pros, cons, and applicability given the nature of the problem and dataset at hand. Selecting the suitable one will require evaluating the clustering results obtained in terms of desirable qualities such as computational efficiency, quality, and reliability. These considerations form the thrust of the next subsection.

D. CLUSTER EVALUATION FOR MIXED AND MISSING DATA

The quality of a clustering result essentially has to do with how well the underlying clusters in a dataset have been identified, and this is directly tied to the purpose and application of the clustering task at hand [15]. Hennig [5] posits that there are two approaches to clustering – a 'constructivist' approach and 'realist' approach. The realist seeks to identify the structures that are within the dataset in an objective manner, independent of any input from the researcher, while the constructivist argues that the underlying structures are only deciphered as they are constructed/seen by the researcher – they only make sense because the researcher attributes sense to them from his/her previous experience, expertise, and/or research requirements. It seems reasonable that a compromise between the two approaches is most appropriate. That is, to some extent, we seek to identify structures which truly underly the data, but these structures are only as good as they are applicable to the problem at hand, and as such, the researcher's input is indispensable in the clustering process. Clearly defining where one falls on this divide is crucial, as it largely forms the researcher's clustering goals, and hence, the methodology adopted in selecting the appropriate clustering results from those available through a plethora of clustering methods. It also informs the interpretation of the selected results are interpreted.

Broadly, cluster evaluation methods fall into two categories: external and internal. In external validation, the derived clustering is compared with an existing 'ground truth'. That is, there already exists an observable categorical variable in the dataset which divides the data into groups, and the clustering process aims to identify these groups exactly, and is evaluated on how well it is able to do so. As has been observed earlier, this is an approach which is widely used in the literature for comparing the performance of clustering algorithms. However, it could be argued that this is as far as the utility of external validation goes, because they require an "artificial situation" which does not apply in real world clustering problems. The major aim of clustering in the real world is to identify ***unknown*** groupings within a dataset, and as such, a variable with true cluster assignments is not available. Moreover, the 'trueness' of these labels where they are available in such artificial situations is contestable, given the constructivist-realist dilemma earlier alluded to – there may be other arguably valid underlying structures which are different from those reflected by the provided labels. Thus, external validation methods may not be sufficiently informative in selecting an appropriate clustering approach [5], [47]. External validation measures still have their place in cluster evaluation, however, especially in cases where internal validation is not feasible, which is the case with many clustering methods for mixed or missing data. They have been divided into three categories in the literature: counting pairs, set-matching, and information theory [48]. While counting pairs basically evaluates clusterings by counting the corresponding pairs of labels on which they agree, set-matching evaluates by comparing pairs of clusters, and information-theoretic validation methods use probability theory to express the amount of relative information contained in the clusterings [49]. These methods are discussed in greater detail in the referenced studies, and we provide further information on the rationale behind those selected for our study in the Materials and Methods section.

Internal validation methods provide insights into the intrinsic quality of a clustering, since they are able to variously provide a measure of its compactness and/or separation. Though these are the major criteria for evaluating a clustering, they are just two among the desirable characteristics of a clustering which are listed in [5], one or more of which each internal validation method is designed to evaluate. In addition, given that the majority of internal validation methods available in the literature perform at varied levels on clusterings depending on their shape, noisiness, density, skewness, and the presence of sub-clusters – Liu *et al.* [50] have grouped internal validation measures based on these features – they can be used to infer these characteristics for the clusters generated. In addition, they can be used to obtain insights on the performance of various clustering algorithms in light of these characteristics [51]. Some of the measures which evaluate cluster separation are the s-index, the Davies-Bouldin index (DB), Modified Hubert statistic (Γ), and the R-squared index (RS). Two measures which evaluate cluster compactness include the root-mean-square standard deviation (RMSSTD) and the widest within-cluster gap (*w*-gap). Some indices which measure the balance between cluster compactness and separation are the Calinski-Harabasz (CH), Dunn (D), and Xie-Beni (XB) indices, as well as the ASW and Clustering Validation index based on Nearest Neighbours (CVNN). Comprehensive discussions of these internal validation measures and others can be found in [50]–[52]. One important internal validation metric which is often overlooked in cluster analysis studies is cluster stability [53], which seeks to evaluate how similar the clusterings produced by an algorithm are when small perturbations are made to the dataset. Specifically, this measures the reliability and generalizability of the clusterings, as a clustering which changes significantly when a little perturbation is made to the dataset cannot be trusted as showing the 'true' underlying structure of the dataset [47], [51]. Generally, this involves resampling the dataset a number times, clustering the resulting datasets using a chosen methodology, comparing the resulting clusterings (either among themselves or to the original clustering) using the Jaccard index and evaluating instability as a function of the mean Jaccard distance between the clusterings.

Conducting internal validation on mixed and missing data clusterings poses a challenge due to the fact that most internal validation methods were designed for Euclidean data, and do not work on missing data [54]. A good number of them, however, work on dissimilarity data, making methods which accept mixed data and produce dissimilarity matrices using hybrid distance measures an available means through which the required dissimilarity matrices can be created. As a result, clustering methods which accept dissimilarity data can be evaluated by internal validation methods that also do so, handily solving the dilemma of internally evaluating mixed data. Some such hybrid distance measures like Gower's distance work for mixed and missing data, but for others, like Ahmad and Dey's distance which only work for mixed data, imputation can be performed as an initial step (though at a higher computational cost). Though Huang's distance also works with mixed and missing data, its use of centroids as a reference point for distance computation does not allow for pairwise dissimilarities. As was stated earlier, this could be trivially overcome by changing the centroid term in the distance equation to another candidate point, but the fact that weights are associated with clusters and automatically calculated as part of the *k*-prototypes clustering algorithm restricts our ability to derive a dissimilarity matrix which is purely data-derived, and not dependent on the clusters which are variable (in size, number, etc.), and thus, computation-driven. In addition, changing terms in the distance measure raises the question whether we could still refer to it as Huang's measure, given the peculiar rationale behind its development. Similar restrictions and considerations apply to the Modha-Spangler method. On the other hand, the Ahmad-Dey distance measure has no weights, and thus, can and has been adapted to produce pairwise dissimilarity matrices [55].

E. RELATED WORKS

In this section, we highlight a few studies which have applied cluster analysis specifically to EHR, providing some practical context for our study and the cluster analysis results we provide. Foguet-Boreu *et al.* [56] performed hierarchical cluster analysis on an EHR dataset with 322,328 multimorbidity patients' records who were over 64 years old. Ward's Linkage was used to combine clusters, which were first created by using the Jaccard index to compare patient diagnoses and determining the number of clusters using the CH and other indices. Patients were subsequently assigned to the clusters to which one or more of their diagnoses belonged. Some form of cluster stability testing was also done to access cluster quality. They were able to identify three clusters which were separated by age group. The first cluster had patients both in the 65-79 and ≥ 80 age group and was made of two diagnoses: hypertensive disease and metabolic disorders. The second cluster, consisting of patients aged 65-79 years, had three diagnoses, and the third, consisting of patients who were aged over 80, had five diagnoses. They concluded that some of the clusters identified were new in the literature and should guide clinical measures for the population. However, though the data was made of mixed variables, e.g. age and number of diagnoses (numerical), and sex and diagnoses (categorical), it was not clearly stated how clustering was done in light of the unique challenges posed by mixed data. Also, it was not stated if there were any missing data, and how that missingness was handled, if it existed. A similar methodology to that adopted in this study was used by Guisado-Clavero *et al.* [57] who reported the use of the tandem approach (Multiple Correspondence Analysis or MCA, for dimensionality reduction, and *k*-means for clustering). They also conducted cluster stability evaluation using the Jaccard index. However, missingness ratio and missingness handling methods were not discussed. Kneppers *et al.* [58] performed hierarchical clustering on data for patients with Chronic Obstructive Pulmonary Disease (COPD) whose dimensionality had been reduced using Principal Component Analysis (PCA). Median values were imputed in place of missing values, and ASW was used to select the number of clusters. Their clustering analysis produced two clusters, and *p*-values were used to evaluate the significance of differences in variable values between the clusters. They report that the first cluster showed more pronounced changes in autophagy, myogenesis, glucocorticoid signaling, oxidative metabolism regulation, etc. in reaction to pulmonary rehabilitation than the second one.

Yu *et al.* [59] used a heatmap and hierarchical clustering with Ward's Linkage to perform clustering analysis on a dataset of 2287 Chronic Kidney Disease (CKD) patients. Variables with $\geq 10\%$ missingness ratio were excluded from the analysis, and missingness in the remaining 23 variables was handled using EM. Their cluster analysis produced three clusters, with cluster one comprised of patients with CKD Stage 1, cluster two of patients with CKD Stage 2, and cluster three, of patients with CKD Stages 3-5. The methodology through which this number of clusters was arrived at is unclear, as no details were given on internal validation measures adopted. In addition, though it was stated that Euclidean distances were used in heatmap and hierarchical clustering, it is not clear how this was achieved, given the mixed nature of their dataset. A similar claim was made in Lenart *et al.* [60] who also performed cluster analysis on a mixed and missing longitudinal CKD dataset with 10 variables. Of 10,014 observations, missingness was handled by complete case analysis which resulted in the analysis being performed on 2,696 observations. K-means, k-medoids, and hierarchical clustering were explored, and k-medoids with four clusters finally selected. It is not clear how this selection was made, as no internal validation methods were mentioned. However, a Cluster Progression Score (CPS) was calculated and used to track patients' progression over time from one cluster to another, with a negative score indicating progress to a favorable cluster, and a positive score, the converse.

From the studies reviewed, it is clear that cluster analysis is of benefit in analyzing patterns in EHR data. However, a number of these studies failed to clearly indicate how clustering was done in light of the mixed nature of their datasets, with some indicating the use of Euclidean measures. The use of Euclidean dissimilarity on mixed data not practicable (as has been explained earlier in this study) barring some conversion from categorical to continuous or vice versa, which was not reported. It is also possible that a simple direct conversion of categorical values to numerical was done, in which case, the meaning behind those categorical variables would have been lost, raising questions about the validity of the results reported. Finally, a number of the studies used basic methods for missingness handling like complete-case analysis or mean, median, or mode imputation which have been shown in the literature to stand a significant chance of falsely altering the data distribution and introducing bias into the results [61].

III. MATERIALS AND METHODS

A. DATA

The dataset used for the study was derived from a longitudinal observational study conducted at the Inkosi Albert Luthuli Central Hospital, KwaZulu-Natal, South Africa over the course of three years (2007-2009) on a cohort of mixed-racial CKD patients at CKD stages ranging from 1 to 5. Demographic data such as age, sex, and race were recorded, and laboratory measurements were performed at six-month intervals. Variables measured include blood creatinine, proteinuria, uric acid, serum urea, Magnesium, and Phosphate. In addition, records were taken of interventions administered during patient visits including statins, carvedilol, angiotencin-converting enzyme inhibitor ACE(I) / angiotencin receptor blockers (ARBs), and non-dihydropyridone ca channel blocker (NDCCB). Estimated Glomerular Filtration Rate (eGFR) was computed using the Chronic Kidney Disease-Epidemiology Collaboration (CKD-EPI$_{\text{creat}}$) equation [62]:

$$\begin{aligned}\text{eGFR} = {} & 141 \times \min(\text{Scr}/\kappa, 1)^{\alpha} \times \max(\text{Scr}/\kappa, 1)^{-1.209} \\ & \times 0.993^{\text{Age}} \times 1.018\,[\text{if female}] \times 1.159\,[\text{if black}],\end{aligned} \tag{6}$$

where Scr is serum creatinine, κ is 0.7 for females and 0.9 for males, α is -0.329 for females and -0.411 for males, min indicates the minimum of Scr/κ or 1, and max indicates the maximum of Scr/κ or 1.

The CKD-EPI$_{\text{creat}}$ equation was used as implemented in the ***translplantr*** package [63]. CKD stage was then computed from the eGFR scores as specified in. Thus, the dataset consists of binary and nominal (categorical) as well as continuous variables. Each measurement/intervention was recorded five times with an interval of six months between measurements.

After preliminary data cleaning, variables with $\geq$ 80% missingness ratio were removed from the dataset, resulting in an overall missingness ratio of 35.2% for the dataset with 280 records. Missingness was then treated using both multiple and single imputation. Multiple imputation was carried out using the ***mice*** package [64], with the m imputed datasets aggregated into a single dataset by taking median for continuous values and mode for categorical variables. This aggregation was done as an approximation of the pooling stage of MI, since carrying out cluster analysis before pooling is largely impracticable for cluster analysis. Thus, we carried out pooling (in the form of data merging) before analysis (clustering). Single imputation was proxied by simply selecting one of the m datasets obtained from MI as the candidate for clustering (i.e., no pooling was done). Due to the fact that the clustering of longitudinal datasets is outside the scope of this study, only the measurements taken during the second time block formed part of the cluster analysis performed. This resulted in 32 variables – 12 categorical and 20 continuous.

B. EXPERIMENTAL SETUP

All experiments were carried out on a Core i7-7500U HP Zbook 14u G4 laptop which has four processors running at approximately 2.7GHz and 20GB RAM. Only one processor was used for the experiments, however. All experiments were conducted using R [42].

Given that there is no such thing as a generally best clustering algorithm [47], [65], we explored a number of clustering algorithms in our analysis in a bid to find the one which would best suit our dataset. Most of our experimentation was focused on mixed data clustering approaches, as missingness in the dataset had already been treated as earlier outlined. No direct conversion from categorical to continuous or vice versa was done due to the highlighted disadvantages associated with that approach, Approaches involving dimensionality reduction were also not applicable to our dataset due to the fact that the 'principal components' derived explained little of the variability in the dataset – the top five components explained, cumulatively, just 32.9% variability. For mixture

TABLE 1. Clustering algorithms used in the study.

Algorithm	Type	Parameters (default)
PAM	Partitional	-
DIANA	Heirarchical	Linkage: average
AGNES	Hierarchical	Linkage: average
Genie	Hierarchical	Gini Threshold: 0.3

models, KAMILA and ClustMD were tested through the packages ***kamila*** [19] and ***clustMD*** [66], respectively. Hybrid distance measures were also explored – Huang's distance, Gower's distance, and the Ahmad-Dey distance, as implemented in the ***clustMixType*** [44], ***cluster*** [43], and ***DisimForMixed*** [55] packages, respectively. The Ahmad-Dey distance was slightly modified in our experiments. As was earlier alluded to, Huang's distance could only be explored through the k-prototypes algorithm as implemented in the ***clustMixType*** package.

For the Gower and Ahmad-Dey distances, which could be accessed directly, their dissimilarity matrices were computed and used as input to four clustering algorithms: PAM, DIvisive ANAlysis clustering (DIANA), AGglomerative NESting (AGNES), and Genie clustering. PAM, DIANA, and AGNES are comprehensively described in [72], and are implemented in the ***cluster*** [43] package. The Genie algorithm is presented in [73], and implemented in the *genie* function of the ***geniclust*** [74] package. The need to find algorithms which accepted distance matrices, allowed for the specification of a k value for the number of clusters, and had a readily available implementation in **R** limited our choice of clustering algorithms for experimentation. These criteria were necessary for conducing comparisons using our internal validation methodology as has been discussed in the Literature Review section. A brief summary of these algorithms is presented in Table 1. The default values set for all algorithms parameters in their implementation were maintained. Thus, there was no parameter tuning conducted. Internal validation was not feasible for some of the methods above, and the reasons have been discussed in preceding sections of this study. In those cases, external validation was done, with CKD stage used as the true label, necessitating that the number of clusters be restricted to five. Meila [48] reports that the best counting pairs external validation criterion is the Adjusted Rand Index (ARI), which is widely used in the literature. Also reported to be widely used in the literature are the Normalized Mutual Information (NMI), and Variation of Information (VI), as well as the Misclassification Error (H) [48]. Following popular practice, thus, we adopted these metrics for external evaluation, but used Normalized Variation of Information (NVI) in place of VI because all indices were to be aggregated. Also, Meila [48] recommended that H be used only with cluster numbers less than 5-6. Thus, we replaced it with the F-measure. All external validation methods were used as implemented in the *external_validation* function of the ***clusterR*** [67] package.

Where internal validation was feasible, one index evaluating compactness (w-gap), one evaluating separation (s-index), and one evaluating both compactness and separation (ASW) were used. These three internal validation methods were used as implemented in the *cqcluster.stats* function of the ***fpc*** package [68]. We added a second and relatively new index, CVNN, because it has been shown to perform well on clusters that are skewed, arbitrarily shaped, noisy, and containing sub-clusters [50]. CVNN is implemented in the *cvnn* function of ***fpc*** with a correction explained in [54]. For each clustering approach, we tested two to five numbers of clusters, evaluated by all four internal validation criteria. Cluster stability was also evaluated using a bootstrap resampling scheme with 100 samples and the Jaccard index for comparison, implemented in the *clusterboot* function of ***fpc***.

For both internal and external validation, each clustering approach was tested on both the singly and multiply imputed datasets, and where a method accepted missing data (k-prototypes and Gower's distance), the unimputed dataset was also used. Thus, a synopsis of our methodology is as follows: data treatment is followed by clustering (done separately for externally and internally validated approaches). After clustering, external and internal validation are performed, and the clustering produced from both is then analyzed statistically analyzed. A graphical summary of this general methodology is provided in Figure 1.

IV. RESULTS AND DISCUSSION

We first present the results of KAMILA, ClustMD, and k-prototypes clustering on our dataset, as shown in Table 2, and summarized in Figure 2. As earlier stated, each algorithm was used to perform clustering on both multiply imputed and singly imputed datasets (represented with "_mi" and "_si" suffixes, respectively). The k-prototypes algorithm was also used cluster the missing dataset, since it accepts missing values. Overall, ClustMD performed best on three external validation indices (ARI, F-measure, and NMI), while KAMILA performed best on NVI. This was the case both on the SI and MI datasets, though it generally performed better on the MI than the SI datasets. This behavior can also be observed with KAMILA, which consistently performed better on the MI than SI dataset for all external validation measures. However, k-prototypes performed better on the SI and missing datasets than on the MI dataset. It is worthy of note that the highest ARI value across all measures is 0.46 (less than 0.5), for the F-measure, it is 0.60, and the same goes for NMI.

Though all clustering algorithms tend to perform better in terms of NVI, the best average value of all external validation measures was 0.56. This shows clearly that all the clustering results produced failed to track/align with the prescribed 5-class grouping that is espoused by the CKD stage.

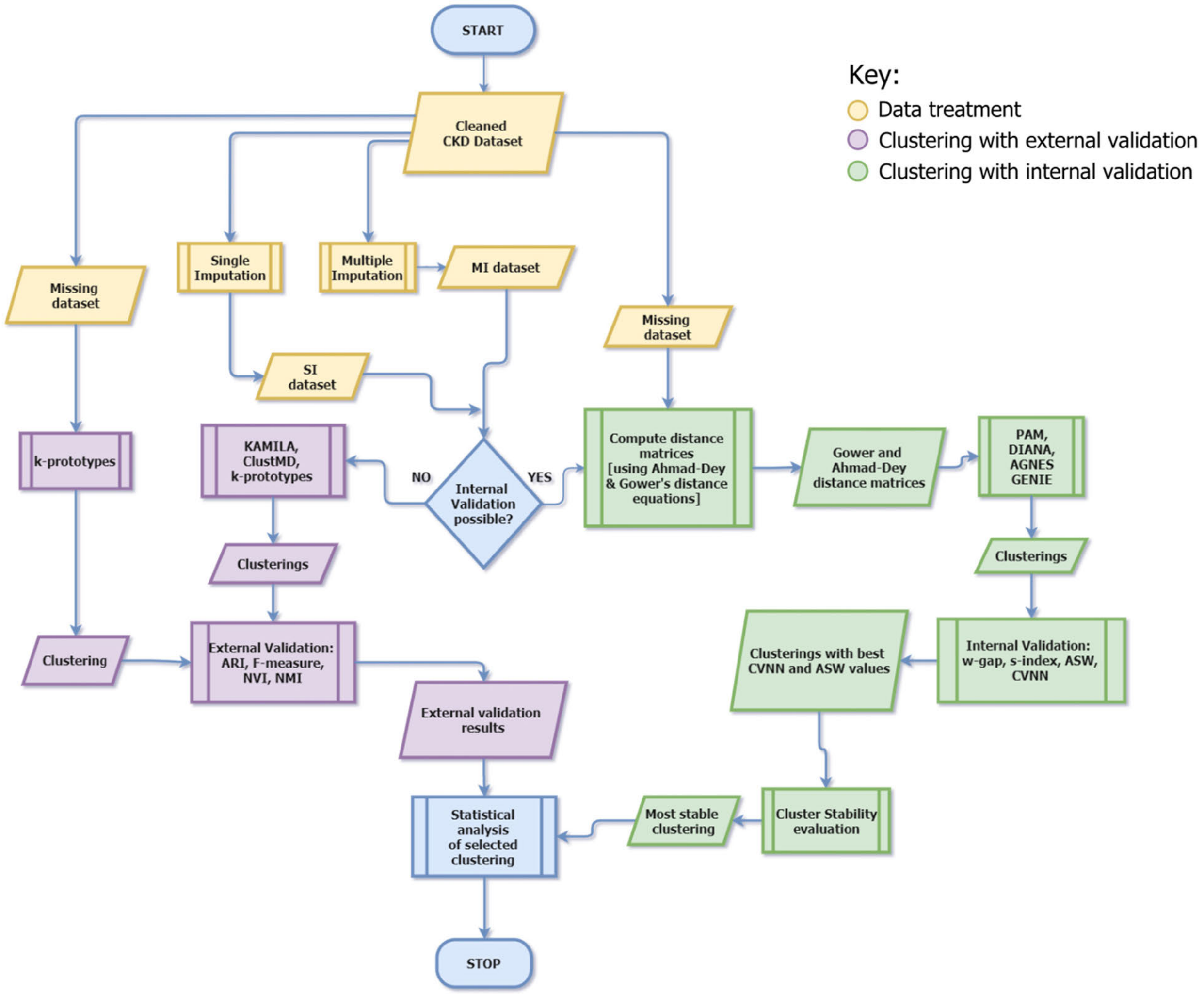

START
Key:
○ Data treatment
○ Clustering with external validation
○ Clustering with internal validation
Cleaned CKD Dataset
Single Imputation
Multiple Imputation
MI dataset
Missing dataset
SI dataset
Missing dataset
k-prototypes
KAMILA, ClustMD, k-prototypes
NO
Internal Validation possible?
YES
Compute distance matrices [using Ahmad-Dey & Gower's distance equations]
Gower and Ahmad-Dey distance matrices
PAM, DIANA, AGNES GENIE
Clusterings
Clusterings
Clustering
External Validation: ARI, F-measure, NVI, NMI
Clusterings with best CVNN and ASW values
Internal Validation: w-gap, s-index, ASW, CVNN
External validation results
Statistical analysis of selected clustering
Most stable clustering
Cluster Stability evaluation
STOP

FIGURE 1. Flowchart showing the general methodology adopted for cluster analysis of our mixed and missing CKD dataset.

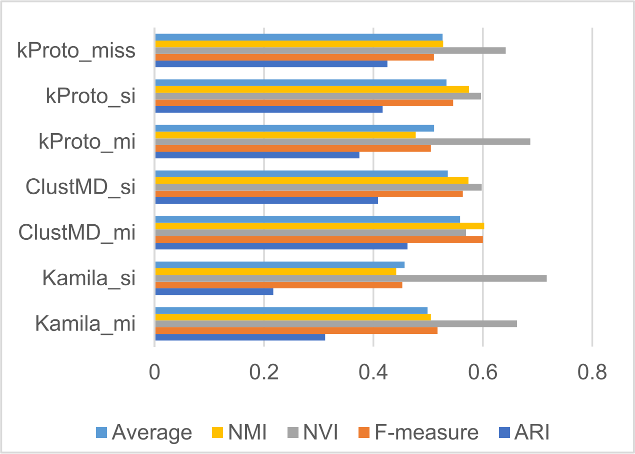


FIGURE 2. Clustered bar chart showing results of external validation conducted on missing and imputed datasets for k-prototypes, ClustMD, and KAMILA.

In other words, they give an indication that, taken together, all variables describing this dataset may form a latent structure which is not captured by the CKD stage. We now proceed to other algorithms where we were able to perform internal validation to obtain further insights on the latent structure of our dataset.

TABLE 2. External validation*.

	ARI	F-measure	NVI	NMI	Average
Kamila_mi	0.3117	0.5172	0.6623	0.5049	0.4990
Kamila_si	0.2169	0.4527	**0.7166**	0.4416	0.4570
ClustMD_mi	**0.4621**	**0.6001**	0.5690	**0.6024**	**0.5584**
ClustMD_si	0.4083	0.5632	0.5979	0.5736	0.5358
kProto_mi	0.3742	0.5049	0.6865	0.4773	0.5107
kProto_si	0.4167	0.5457	0.5966	0.5748	0.5335
kProto_miss	0.4255	0.5104	0.6417	0.5275	0.5263

*All indices are to be maximized, best values in bold

Table 3 shows the results of internal validation through the *w*-gap, *s*-index, ASW, and CVNN. Ordinarily, the *w*-gap index should be minimized because it reflects the maximum widest within-cluster gap across the three clusters, which we would want to be as small as possible to achieve compact and homogenous clusters. However, together with *s*-index and ASW, it's values were standardized towards maximization and into a [0,1] range.

TABLE 3. Internal Validation*.

			GOWER				AHMAD-DEY			
			PAM	DIANA	AGNES	GENIE	PAM	DIANA	AGNES	GENIE
MI	*w*-gap	2	0.5620	0.6136	**0.6300**	0.6297	0.7910	0.7910	0.7910	0.7910
		3	0.5620	0.6107	**0.6300**	0.6297	0.7910	0.9425	0.9425	0.7910
		4	0.5620	0.6060	**0.6300**	0.6297	0.9425	0.9425	0.9425	0.7910
		5	0.5738	0.6060	0.6128	0.6297	0.9425	**0.9547**	0.9736	0.7910
	s-index	2	0.1849	0.1995	**0.4483**	0.2314	0.0027	0.0030	0.1300	0.0027
		3	0.1502	0.1959	0.2715	0.2274	0.0011	0.0101	**0.1327**	0.0014
		4	0.1566	0.1947	0.2734	0.2235	0.0080	0.0101	0.0146	0.0010
		5	0.1460	0.1652	0.2658	0.2204	0.0083	0.0116	0.0159	0.0011
	ASW	2	0.1347	0.1499	**0.2392**	0.0243	0.8288	0.8319	0.8376	0.7963
		3	0.0860	0.1148	0.1401	0.0673	0.6459	**0.8427**	0.8199	0.4138
		4	0.0859	0.0993	0.0782	0.0645	0.6890	0.8256	0.8139	0.1138
		5	0.0805	0.0958	0.0707	0.0388	0.5860	0.8247	0.8055	0.1612
	CVNN	2	1.6465	**1.2760**	2.0000	1.6251	1.0100	1.0100	2.0000	1.1007
		3	1.8604	1.9901	1.9739	1.5725	**0.7958**	1.7686	1.9391	1.2229
		4	1.9403	1.9834	1.9711	1.7807	1.4318	1.7211	1.3034	2.0000
		5	1.8752	1.9330	1.9613	1.9254	1.3393	1.7082	1.3010	1.6460
SI	*w*-gap	2	**0.5852**	0.5698	0.4738	**0.5852**	0.7857	0.7857	0.9412	0.7857
		3	**0.5852**	0.5698	0.4738	**0.5852**	0.7857	0.9412	0.9412	0.7857
		4	0.5708	0.5698	0.4738	**0.5852**	0.7857	0.9412	0.9412	0.7857
		5	0.5708	0.5698	0.4738	**0.5852**	0.9412	0.9412	**0.9728**	0.7857
	s-index	2	0.1771	0.2096	**0.3767**	0.2354	0.0035	0.0032	**0.3410**	0.0024
		3	0.1651	0.2088	0.3323	0.2160	0.0017	0.0102	0.0117	0.0013
		4	0.1725	0.1835	0.3193	0.2154	0.0011	0.0104	0.0130	0.0011
		5	0.1719	0.1838	0.2504	0.2156	0.0079	0.0120	0.0145	0.0011
	ASW	2	0.1039	0.1440	**0.1696**	0.0732	0.8078	0.8211	**0.8935**	0.7898
		3	0.0872	0.1110	0.0949	0.0390	0.7207	0.8297	0.8262	0.4203
		4	0.0638	0.0901	0.0502	0.0389	0.5462	0.7776	0.7800	0.0679
		5	0.0649	0.0861	0.0731	0.0258	0.6245	0.6975	0.7718	0.1531
	CVNN	2	1.4669	**1.3779**	1.9630	1.4736	1.0098	1.0468	2.0000	1.0356
		3	1.6016	1.6891	1.9560	1.4872	**0.7634**	1.8069	1.3633	1.2024
		4	1.8037	1.6441	1.9907	1.7018	0.7833	1.7408	1.3568	2.0000
		5	1.9244	1.9429	1.9436	1.9495	1.3347	1.7180	1.3557	1.5736
Miss	*w*-gap	2	0.6650	0.6669	0.6520	**0.8250**	-	-	-	-
		3	0.6439	0.6466	0.6454	**0.8250**	-	-	-	-
		4	0.6521	0.6444	0.5861	**0.8250**	-	-	-	-
		5	0.6308	0.6444	0.5861	**0.8250**	-	-	-	-
	s-index	2	0.1761	0.1876	**0.2373**	0.0874	-	-	-	-
		3	0.1518	0.1851	0.2345	0.0813	-	-	-	-
		4	0.1163	0.1631	0.2284	0.0718	-	-	-	-
		5	0.1056	0.1581	0.2295	0.0618	-	-	-	-
	ASW	2	0.1225	0.1743	0.2206	0.2451	-	-	-	-
		3	0.0800	0.1536	0.1635	0.2670	-	-	-	-
		4	0.0811	0.1113	0.1197	**0.3105**	-	-	-	-
		5	0.0723	0.1036	0.0949	0.2615	-	-	-	-
	CVNN	2	1.3351	1.2963	1.7059	1.1455	-	-	-	-
		3	1.7720	1.9812	1.9190	1.1018	-	-	-	-
		4	1.8558	1.9420	1.9411	**0.9595**	-	-	-	-
		5	1.9137	1.9193	1.9743	1.7200	-	-	-	-

***All internal validation indices are to be maximized, except CVNN, which is to be minimized. Best values for each box in bold.**

These three indices have been used to find a good number of clusters in the literature [65], [68]. On the MI dataset, AGNES clustering using Gower's distance shows the best *w*-gap value of the three clustering algorithms, which is the same for cluster numbers 2-4 (0.6300). However, much higher values are achieved by all clustering approaches using the Ahmad-Dey distance, with DIANA achieving the highest with five clusters (0.9547). Similarly, on the SI dataset, the Ahmad- Dey distance produces much better *w*-gap values than Gower's distance, but in this case, AGNES posts the highest value with four and five clusters (0.9728). On the missing dataset, only results for Gower's distance are shown because the Ahmad-Dey distance does not accept missing values. However, Genie shows the highest *w*-gap value here (0.8250), as is the case with the SI dataset (0.5852). It is interesting to note that this is the highest *w*-gap value achieved across all three imputation and clustering approaches for Gower's distance. This suggests that the missingness handling method adopted by Gower's distance could be adopted and perform well for clustering. In conclusion, we note that,

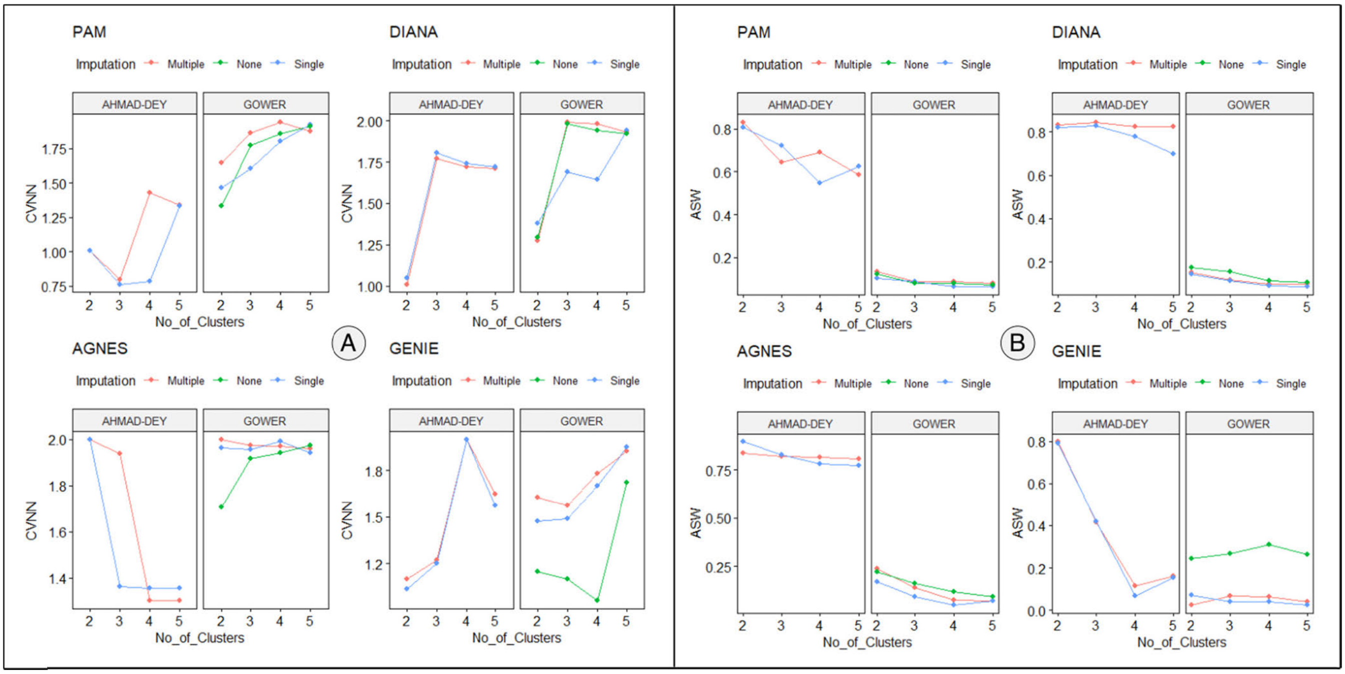


FIGURE 3. A comparison of CVNN (A) and ASW (B) values by imputation method, clustering algorithm, distance measure, and number of clusters ranging from 2 to 5. For CVNN, smaller is better, while for ASW, larger is better.

with respect to *w*-gap, five clusters have generally shown the best values in our experiments, the Ahmad-Dey distance outperformed Gower's distance in all cases, the SI dataset showed better values than the MI dataset, and hierarchical clustering methods performed better than PAM. On the MI dataset, AGNES with Gower's distance achieved the best *s*-index of 0.4483 (on two clusters), also achieving the best value with the Ahmad-Dey distance (on three clusters), though this was much lower at 0.1327. This is also the case on the SI dataset, where AGNES with the Ahmad-Dey distance on two clusters, being the best among the three clustering methods (at 0.3410), is outperformed by its Gower's distance counterpart which was the best of the three at 0.3767.

On the missing dataset, AGNES with two clusters outperformed the three other clustering approaches (with Gower's distance) at 0.2373. In summary, for the *s*-index, the MI dataset outperformed the two others, hierarchical clustering (AGNES in particular) outperformed PAM, Gower's distance outperformed the Ahmad-Dey distance, and a smaller number of clusters (2-3) showed better results. The two internal validation indices evaluated so far show a trend (observable from Table 2), that smaller cluster numbers lead to more well-separated clusters, while larger cluster numbers lead to more compact clusters. This trend has been similarly observed in the literature [54].

We now examine the results of the two internal validation indices which evaluate both separation and compactness. As Table 3 shows, for the MI dataset, AGNES with Gower's distance performs best of the three clustering approaches with an ASW of 0.2392 (two clusters).

TABLE 4. Stability testing*.

			GOWER				A-D			
			PAM	DIANA	AGNES	GENIE	PAM	DIANA	AGNES	GENIE
MI	2	1	0.68	**0.85**	0.99	0.26	**0.99**	0.94	0.98	0.88
		2	0.59	**0.91**	0.59	0.71	**0.96**	0.79	0.57	0.79
	3	1	0.48	0.22	0.92	0.38	**0.82**	0.99	0.91	0.65
		2	0.57	0.61	0.24	0.66	**0.93**	0.92	0.61	0.78
		3	0.53	0.86	0.68	0.41	**0.94**	0.70	0.63	0.93
	4	1	0.40	0.23	0.90	0.47	0.74	0.98	0.96	0.71
		2	0.56	0.28	0.33	0.30	0.90	0.82	0.86	0.70
		3	0.42	0.66	0.32	0.64	0.91	0.62	0.66	0.77
		4	0.46	0.50	0.75	0.48	0.69	0.58	0.58	0.42
	5	1	0.49	0.24	0.89	0.40	0.76	0.91	0.93	0.76
		2	0.47	0.28	0.37	0.36	0.66	0.76	0.88	0.71
		3	0.54	0.56	0.33	0.59	0.93	0.59	0.63	0.80
		4	0.50	0.49	0.48	0.26	0.81	0.59	0.65	0.49
		5	0.55	0.44	0.64	0.50	0.68	0.62	0.57	0.80
SI	2	1	0.71	**0.91**	0.97	0.31	**0.99**	0.97	0.98	0.89
		2	0.72	**0.82**	0.26	0.70	**0.97**	0.85	0.56	0.80
	3	1	0.64	0.78	0.94	0.43	**0.81**	0.98	0.93	0.68
		2	0.61	0.62	0.27	0.42	**0.94**	0.84	0.47	0.86
		3	0.57	0.34	0.34	0.46	**0.81**	0.68	0.65	0.88
	4	1	0.59	0.64	0.89	0.41	0.53	0.97	0.93	0.71
		2	0.58	0.62	0.29	0.40	0.65	0.81	0.54	0.80
		3	0.51	0.42	0.17	0.20	0.83	0.77	0.62	0.76
		4	0.36	0.45	0.48	0.43	0.62	0.58	0.74	0.46
	5	1	0.54	0.59	0.85	0.42	0.80	0.93	0.88	0.70
		2	0.59	0.59	0.39	0.39	0.71	0.79	0.59	0.70
		3	0.48	0.39	0.34	0.31	0.93	0.76	0.59	0.74
		4	0.39	0.49	0.33	0.42	0.61	0.63	0.80	0.51
		5	0.39	0.21	0.48	0.17	0.57	0.58	0.56	0.76

*Higher is better. Best results for each box in bold

However, with the Ahmad-Dey distance, DIANA performs best with a much higher value of 0.8427 (three clusters). For the SI dataset, with Gower's distance, AGNES outperforms others at an ASW value of 0.1696 (two clusters), but once again, its Ahmad-Dey counterpart, which also outperforms the two others, is much better at 0.8935 with the same number of clusters. On the missing dataset, Genie with four

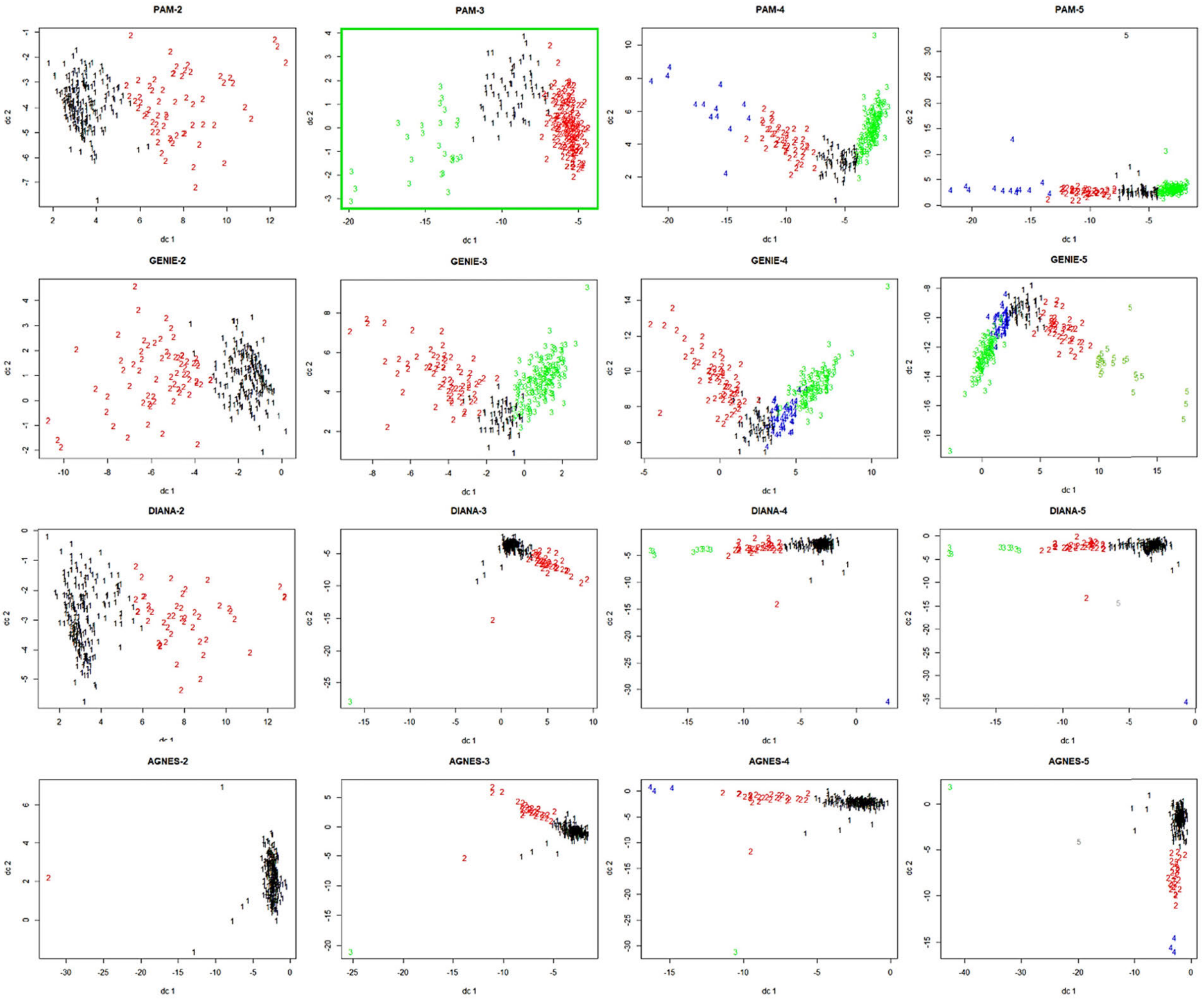


FIGURE 4. Scatterplots of all clustering results for SI with the Ahmad-Dey distance measure. On each row, from top to bottom: PAM, GENIE, DIANA, AGNES. Cluster numbers from 2 to 5, left to right. Selected clustering highlighted in green.

clusters performs best at 0.3105. Overall, we observe that the Ahmad-Dey distance significantly outperforms Gower's distance in all cases, AGNES with the Ahmad-Dey distance on two clusters achieves the highest overall ASW value, and the SI dataset produced better performance than the MI dataset. The CVNN index appears to tell a different story, though. On the MI dataset, PAM achieves the best value with two clusters and the Ahmad-Dey distance at 0.7361. The best result obtained from Gower's distance is from DIANA with two clusters, but at a significantly higher value of 1.2760 than that of PAM with Ahmad-Dey (lower is better for CVNN). On the SI dataset, we find an even better result with a three-cluster Ahmad-Dey and PAM (0.7634) than that obtained from MI with the same set up. Gower's distance performs worse with a two-cluster DIANA set up at 1.3779 (which outperforms both PAM and AGNES), also worse than both best results obtained from the SI and missing datasets. A four-cluster Genie set up performed best on the missing dataset with Gower's distance at 0.9595. In summary, for both ASW and CVNN, the Ahmad-Dey distance outperforms Gower's distance, and SI outperforms both MI and missingness. However, though ASW favours two clusters with AGNES, CVNN favours three with PAM. Figure 3 gives a picture of the results discussed here, showing more clearly, the observable disparities between CVNN and ASW, as well as between Gower and Ahmad-Dey. It also shows that, generally, the values of these internal validation indices decline with increasing cluster numbers, and that SI performed best of the three missing data treatment approaches. The two best results which can be conclusively selected from the internal validation carried out so far, are, therefore, the two-cluster AGNES, and three-cluster PAM, both with SI and the Ahmad-Dey distance measure. We now examine the results of cluster stability testing as presented in Table 4. The table shows the mean Jaccard similarities between each original cluster and the clusters obtained from the 100 resampled datasets for each number of clusters. That is, each cluster within the k-clustering obtained on a resampled dataset is compared to its corresponding cluster in the original clustering. This produces 100 Jaccard similarity values for each cluster, which are then averaged. Generally, a Jaccard similarity value, $J \geq 0.75$ is accepted as indicating a stable cluster, and $J \geq 0.85$ indicates a highly stable cluster. $J \leq 0.50$ indicates a clustering that is unreliable, and J between 0.60 and 0.75 indicates some pattern in

TABLE 5. Cluster baseline characteristics.

	Complete data	Cluster 1	Cluster 2	Cluster 3	
	n = 280	n = 61 (21.8%)	n = 193 (68.9%)	n = 26 (9.3%)	P value*
Demographics					
Age	44.00 [31.75-54.25]	46.00 [36.00-55.00]	42.00 [28.00-54.00]	52.00 [43.25-54.75]	ns
Race	Asian (58.57)	Asian (62.30)	Asian (57.51)	Asian (57.69)	ns
Gender	Female (61.43)	Female (54.10)	Female (67.88)	Male (69.23)	< 0.05
Weight	69.30 [59.15-82.85]	63.6 [57.20-77.00]	70.00 [59.80-84.00]	68.40 [61.20-82.95]	ns
Height	1.60 [1.52-1.68]	1.60 [1.54-1.68]	1.60 [1.51-1.67]	1.66 [1.60-1.70]	< 0.05
BMI	Obese (34.29)	Normal (40.98)	Obese (36.27)	Normal (46.15)	ns
Lab. measurements					
Creatinine	146.00 [76.00-335.25]	418.00 [328.00-505.00]	92.00 [68.00-162.00]	778.00 [723.00-987.20]	< 0.05
eGFR	43.43 [16.07-95.40]	12.34 [9.86-16.18]	75.76 [40.84-107.38]	5.96 [4.72-6.84]	< 0.05
Blood Pressure (sys/dias)	< 140/90 (75.00)	< 140/90 (78.69)	< 140/90 (75.65)	< 140/90 (61.54)	ns
Proteinuria	TRACE (25.71)	++ (26.23)	TRACE (30.05)	+ (26.92) & +++ (26.92)	< 0.05
Presence of Atherosclerotic disease	No (91.79)	No (90.16)	No (94.24)	No (80.77)	ns
Haemoglobin	11.95 [10.90-13.00]	11.20 [10.10-11.70]	12.40 [11.30-13.50]	10.60 [10.03-10.76]	< 0.05
Sodium	140.00 [139.00-142.00]	140.00 [139.00-142.00]	140.00 [139.00-142.00]	141.00 [140.00-142.00]	ns
Potassium	4.70 [4.30-5.32]	5.20 [4.60-5.90]	4.60 [4.30-5.10]	5.15 [4.50-5.68]	< 0.05
Chlorine	107.00 [105.00-110.00]	108.00 [104.00-112.00]	107.00 [105.00-109.00]	107.00 [105.00-109.80]	ns
HC03 - Bicarbonate	24.55 [21.40-26.43]	20.00 [17.40-24.60]	24.90 [22.90-27.10]	24.60 [19.80-25.40]	< 0.05
Urea	9.75 [5.60-17.93]	21.00 [14.60-24.90]	7.30 [4.90-12.40]	18.40 [13.90-25.90]	< 0.05
Calcium	2.27 [2.13-2.35]	2.26 [2.10-2.36]	2.27 [2.15-2.35]	2.24 [2.04-2.34]	ns
Phosphate	1.23 [1.08-1.52]	1.44 [1.23-1.90]	1.17 [1.02-1.40]	1.68 [1.40-1.99]	< 0.05
Magnesium	0.84 [0.78-0.94]	0.83 [0.76-0.96]	0.84 [0.78-0.91]	0.92 [0.81-1.08]	< 0.05
Uricacid	0.40 [0.30-0.50]	0.45 [0.40-0.50]	0.40 [0.30-0.50]	0.58 [0.40-0.60]	< 0.05
Ultrasound kidney size	Not taken (82.50)	Not taken (86.89)	Not taken (80.31)	Not taken (88.46)	ns
IGFR	Not taken (41.43)	30-60 (49.18)	Not taken (51.30)	30-60 (57.69)	< 0.05
Interventions					
Iron supplement	No (67.86)	No (59.02)	No (70.98)	No (65.38)	ns
Diuretics	Yes (75.36)	Yes (88.52)	Yes (68.39)	Yes (96.15)	< 0.05
ACEI	Yes (91.79)	Yes (91.80)	Yes (90.67)	Yes (100)	ns
NDCCB	No (54.64)	Yes (55.90)	No (59.59)	Yes (61.54)	< 0.05
Carvedilol	No (85.00)	No (75.41)	No (88.60)	No (80.77)	< 0.05
No. of Anti-hypertensives	2 [1-3]	2 [1-3]	2 [1-3]	3 [2-4]	< 0.05
Statin	Yes (66.07)	Yes (67.21)	Yes (66.32)	Yes (61.54)	ns

Data expressed as median [inter-quartile range] for continuous, and mode (%) for continuous variables.

***Significance of between-cluster (clusters 1, 2, and 3) differences assessed by the Kruskal-Wallis test at a 0.05 significance level.**

ns = not significant.

the data, but gives no assurance regarding cluster allocations. Thus, in interpreting the results in Table 3, we ruled out any clustering where one or more of the clusters had $J < 0.75$. Based on this criterion, and selecting the best where more than one configuration met the condition, the two-cluster DIANA + Gower and the two- and three-cluster PAM + Ahmad-Dey approaches were accepted for both the MI and SI datasets, and they are highlighted in bold in Table 3. This automatically eliminated the AGNES + Ahmad-Dey two-cluster approach favoured by ASW, leaving us with

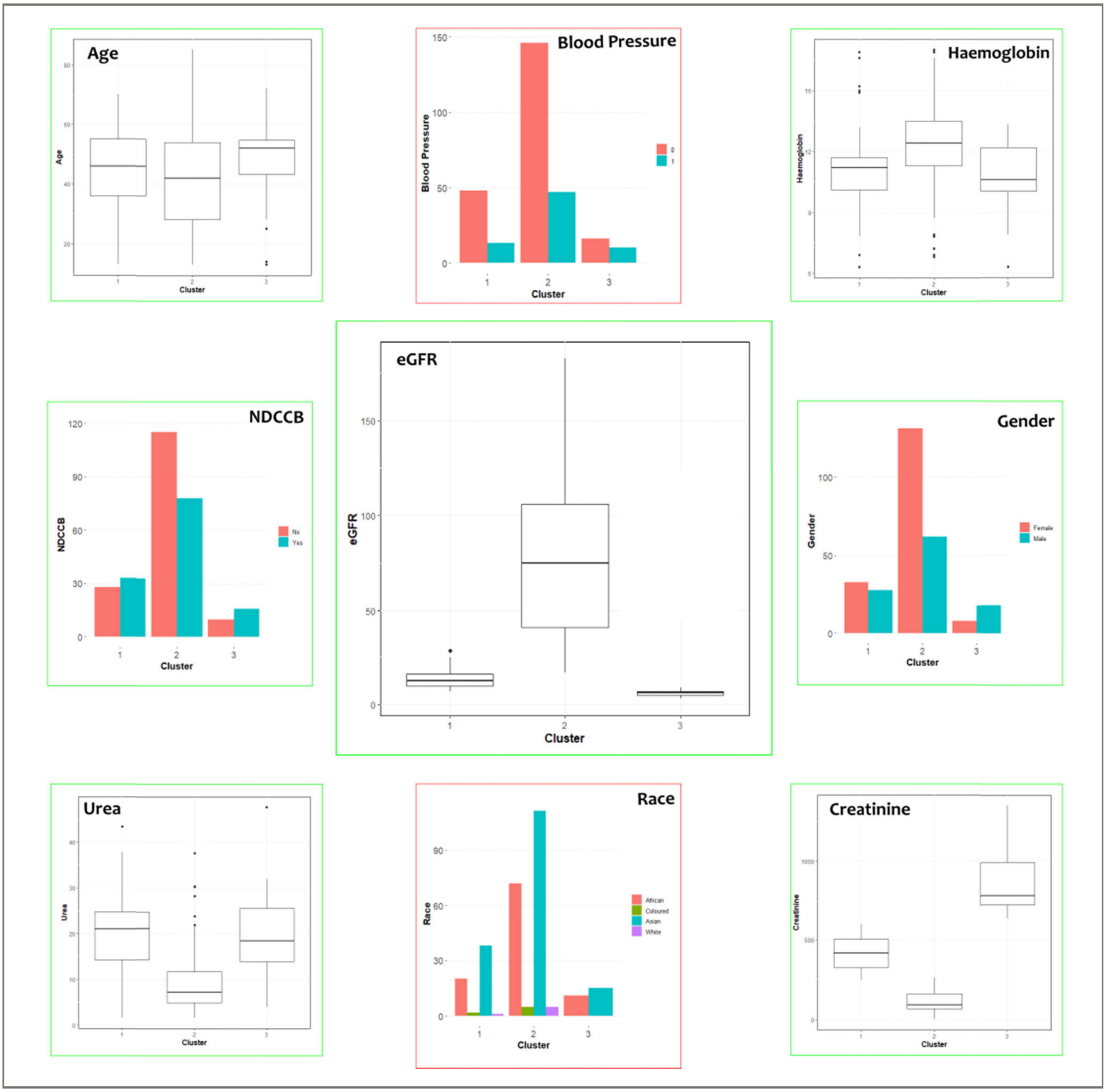


FIGURE 5. Box and bar plots showing variables by cluster (clusters 1 to 3, left to right). Plots enclosed in red reflect variables with no significant difference between clusters, while those enclosed in green reflect variables with significant difference between clusters.

the three-cluster Ahmad-Dey + PAM method on the SI dataset.

A graphical representation of the clusters produced for all four clustering algorithms on the SI dataset with the Ahmad-Dey distance measure is presented in Figure 4. From the graph, it can be observed that all clusterings produced by the AGNES clustering method are characterized by one or two singleton clusters. A similar behaviour can be observed with DIANA clustering. This might explain the relatively high w-gap and s-index values it showed, and casts a shadow on these high values, as generally, and more specifically in our case, singleton clusters are of little practical use, since we desire to obtain information on the latent structure of our dataset so as to provide insights which will be relevant to nephrologists managing CKD patients. The fact that our internal validation procedure led to the selection of a three-cluster solution also confirms the conclusion earlier reached from the results of external validation that there is a latent structure in the CKD dataset as described by its constituent variables which is different from the widely used five-stage CKD grouping. This is reinforced by the fact that the five-cluster PAM clustering also contains a singleton cluster.

We now present analysis of the clustering produced by the three-cluster approach. The baseline characteristics of these clusters are presented in Table 4, which also shows the result of a Kruskal-Wallis significance test conducted to identify the variables on which the three clusters differ significantly. Significance was tested at a 0.95 confidence level. The multi-racial nature of our dataset allows us to gain insights on whether race was a significant discriminating factor between the patients in our cohort which comprised largely of Africans and Asians, with a few White and Colored individuals. In Figure 4, we have attempted to show the variables which we believe would be of most interest to the reader, though all variables and their cluster partitionings are reflected in Table 5. The charts enclosed in red indicate variables with no significant difference between the three clusterings, and it can be immediately seen that the Race variable is one of them. This indicates that in our multi-racial CKD cohort, race was not a significantly distinguishing factor among individuals. The same can be said about blood pressure – there was no significant separation between the three obtained clusters based on blood pressure. Taking center stage in the chart is the eGFR plot, which shows that the clusters are significantly separated by eGFR values. Cluster 2 comprises of patients with relatively high eGFR values (IQR: 40.84-107.38).

This indicates that the majority of patients in cluster 2 fall within CKD Stages 1 to 3A (mild CKD). For Cluster 1, patients are at Stages 4 & 5 CKD (eGFR with IQR 9.86-16.18), while Cluster 3 comprises patients solely at Stage 5, but with clearly lower eGFR (IQR: 4.72-6.84). Stage 5 CKD patients have reached End-Stage Renal Disease (ESRD), also referred to as kidney failure. Of the remaining significant variables shown in Figure 4, Cluster 2 is characterized by relatively younger females with the lowest urea, lowest creatinine, and highest haemoglobin. Cluster 3 is made of relatively younger males with the highest creatinine, higher urea than those in Cluster 2, and lowest hemoglobin. Also of interest is the fact that relatively, individuals in cluster two are more likely to have been administered NDCCB. Individuals in Cluster 1 are mostly females younger than those in Cluster 3, but older than those in Cluster 2, with the highest urea, creatinine higher than those in Cluster 2 but lower than those in Cluster 3, and hemoglobin lower than those in Cluster 2, but higher than those in Cluster 3. It is worthy of note that hemoglobin and eGFR show a similar trend across the three clusters – they both go from highest to lowest from Cluster 2 to Cluster 1, and to Cluster 3. Other studies in the literature have also found an association between kidney function and hemoglobin levels [69], [70].

V. CONCLUSION AND RECOMMENDATIONS

In this study, the unique challenges associated with clustering datasets (particularly EHR) which are both mixed and missing in nature were discussed extensively, as were the approaches available in the literature for tackling them. Cluster analysis was then performed on a multi-racial CKD dataset obtained from the Inkosi Albert Luthuli Central Hospital, Durban, South Africa, the results evaluated using both external and internal validation indices, and theoretical and practical insights obtained.

The theoretical findings from our study are as follows. One, the Ahmad-Dey distance measure significantly outperformed Gower's distance on almost all internal validation indices and clustering algorithms (PAM, DIANA, AGNES, and GENIE). This could be attributed to its unique approach of computing distances through probabilities associated with variable co-occurrence, which removes the need for weights and their associated disadvantages. Secondly, in many cases, SI outperformed both MI and direct analysis on a missing dataset, though in some cases it was outperformed by the others. This indicates that though the pooling stage of MI may not be necessary for cluster analysis, it should be explored, and where clustering can be performed directly on missing data, that should also be done. The results can then be compared to find the most suitable approach on a case-by-case basis. More generally, and with respect to missingness treatment, a comparative analysis of both advanced and simple methods should be carried out to find out the most appropriate approach, as opposed to an off-hand adoption of simple missingness treatment methods like complete- or available-case analysis. Furthermore, it stands to be questioned if the appropriateness of such simple missingness treatment methods truly lies in how good their internal/external validation and cluster stability scores are, in light of the fact that they have been shown to introduce significant bias into a dataset, especially when the missingness ratio is more than 5% [61], [71]. By implication, they stand a chance of altering the latent structure which we seek to discover in our datasets. Thus, our ability to discover these altered structures may be irrelevant at best, or misleading at worst.

On the practical side, our results indicate a latent three-cluster structure in our CKD dataset, in opposition to the five-stage CKD categorization which is normally used for CKD patients. Though the methodology of this study does not allow for outright generalizations, this indicates that CKD datasets which are comprised of more than the four to six variables used for computing eGFR may have some underlying structure which deviates from the usual five CKD stages, and such structures should be explored, as they could provide valuable insights into the characteristics of the cohort being studied. That is, clustering which leverages internal validation should be used more as against simply performing external validation based on these five stages.

Given that the Ahmad-Dey distance measure does not accept missing values, it would be interesting to find ways of extending it to do so, as the methods which actually perform direct cluster analysis on mixed and missing datasets (Huang's k-prototypes [16] and Gower's distance [14]) didn't perform competitively on our dataset. In addition, we hope to perform cluster analysis on our dataset in its original longitudinal form, which we weren't able do in this study due to its constraints. It would be interesting to examine the challenges that this additional (longitudinal) constraint will pose to clustering, and how they can be addressed. Finally, it would be interesting to examine the relative performances of hierarchical and centroid-based clustering methods to identify any patterns in performance with respect to data structure.

REFERENCES

[1] F. Murtagh and M. J. Kurtz, "The classification society's bibliography over four decades: History and content analysis," *J. Classification*, vol. 33, no. 1, pp. 6–29, Apr. 2016, doi: 10.1007/S00357-016-9196-4.

[2] C. Hennig and M. Meilă, "Cluster analysis: An overview," in *Handbook Of Cluster Analysis*, vol. 9, C. M. Hennig, M. Meilă, F. Murtagh, and R. Rocci, Eds. Boca Raton, FL, USA: Chapman & Hall, 2016, pp. 1–19.

[3] M. van de Velden, A. Iodice D'Enza, and A. Markos, "Distance-based clustering of mixed data," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 11, no. 3, p. e1456, May 2019, doi: 10.1002/Wics.1456.

[4] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychol. Methods*, vol. 17, no. 2, pp. 228–243, 2012, doi: 10.1037/A0027127.

[5] C. Hennig, "What are the true clusters?" *Pattern Recognit. Lett.*, vol. 64, pp. 53–62, Oct. 2015, doi: 10.1016/J.Patrec.2015.04.009.

[6] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *IEEE Access*, vol. 7, pp. 31883–31902, 2019, doi: 10.1109/Access.2019.2903568.

[7] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS–clustering categorical data using summaries," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1999, pp. 73–83.

[8] Z. He, X. Xu, and S. Deng, "Squeezer: An efficient algorithm for clustering categorical data," *J. Comput. Sci. Technol.*, vol. 17, no. 5, pp. 611–624, Sep. 2002, doi: 10.1007/bf02948829.

[9] S. Amiri, B. S. Clarke, and J. L. Clarke, "Clustering categorical data via ensembling dissimilarity matrices," *J. Comput. Graph. Statist.*, vol. 27, no. 1, pp. 195–208, Jan. 2018, doi: 10.1080/10618600.2017.1305278.

[10] S. Naouali, S. Ben Salem, and Z. Chtourou, "Clustering categorical data: A survey," *Int. J. Inf. Technol. Decis. Making*, vol. 19, no. 1, pp. 49–96, Jan. 2020.

[11] A. Foss, M. Markatou, B. Ray, and A. Heching, "A semiparametric method for clustering mixed data," *Mach. Learn.*, vol. 105, no. 3, pp. 419–458, Dec. 2016, doi: 10.1007/s10994-016-5575-7.

[12] H. A. L. Kiers, "Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables," *Psychometrika*, vol. 56, no. 2, pp. 197–212, Jun. 1991, doi: 10.1007/Bf02294458.

[13] M. Vichi, D. Vicari, and H. A. L. Kiers, "Clustering and dimension reduction for mixed variables," *Behaviormetrika*, vol. 46, no. 2, pp. 243–269, Oct. 2019, doi: 10.1007/S41237-018-0068-6.

[14] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, p. 857, Dec. 1971, doi: 10.2307/2528823.

[15] A. H. Foss, M. Markatou, and B. Ray, "Distance metrics and clustering methods for mixed type data," *Int. Stat. Rev.*, vol. 87, no. 1, pp. 80–109, Apr. 2019, doi: 10.1111/Insr.12274.

[16] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific-Asia Conf. Knowl. Discovery Data Mining (PAKDM)*, 1997, pp. 21–34.

[17] A. Ahmad and L. Dey, "A K-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, Nov. 2007, doi: 10.1016/J.Datak.2007.03.016.

[18] D. S. Modha and S. W. Spangler, "Feature weighting in K-means clustering," *Mach. Learn.*, vol. 52, no. 3, pp. 217–237, 2003, doi: 10.1023/A:1024016609528.

[19] A. H. Foss and M. Markatou, "Kamila: Clustering mixed-type data in R and Hadoop," *J. Stat. Softw.*, vol. 83, no. 13, pp. 1–45, 2018, doi: 10.18637/Jss.V083.I13.

[20] G. Celeux and G. Govaert, "Latent class models for categorical data," in *Handbook Of Cluster Analysis*, vol. 9, C. M. Hennig, M. Meilă, F. Murtagh, and R. Rocci, Eds. Boca Raton, FL, USA: Chapman & Hall, 2016, pp. 173–193.

[21] L. Hunt and M. Jorgensen, "Clustering mixed data," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 4, pp. 352–361, Jul. 2011, doi: 10.1002/Widm.33.

[22] D. McParland and I. Claire Gormley, "Model based clustering for mixed data: ClustMD," 2015, *arXiv:1511.01720*. [Online]. Available: http://arxiv.org/abs/1511.01720

[23] L. Peng and L. Lei, "A review of missing data treatment methods," *Intell. Inf. Manage. Syst. Technol.*, vol. 1, no. 3, pp. 412–419, 2005.

[24] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*. Hoboken, NJ, USA: Wiley, 2019.

[25] K. J. Lee, G. Roberts, L. W. Doyle, P. J. Anderson, and J. B. Carlin, "Multiple imputation for missing data in a longitudinal cohort study: A tutorial based on a detailed case study involving imputation of missing outcome data," *Int. J. Social Res. Methodol.*, vol. 19, no. 5, pp. 575–591, Sep. 2016.

[26] T. D. Pigott, "A review of methods for missing data," *Educ. Res. Eval.*, vol. 7, no. 4, pp. 353–383, Dec. 2001, doi: 10.1076/edre.7.4.353.8937.

[27] B. J. Wells, K. M. Chagin, A. S. Nowacki, and M. W. Kattan, "Strategies for handling missing data in electronic health record derived data," *Egems*, vol. 1, no. 3, p. 1035, 2013, doi: 10.13063/2327-9214.1035.

[28] W. Young, G. Weckman, and W. Holland, "A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits," *Theor. Issues Ergonom. Sci.*, vol. 12, no. 1, pp. 15–43, Jan. 2011, doi: 10.1080/14639220903470205.

[29] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ, USA: Wiley, 2004.

[30] Z. Zhang and H. Fang, "Multiple-vs non-or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data," in *Proc. 1st Int. Conf. Connected Health, Appl., Syst. Eng. Technol.*, Jun. 2016, pp. 219–228.

[31] S. Goel and M. Tushir, "A new iterative fuzzy clustering approach for incomplete data," *J. Statist. Manage. Syst.*, vol. 23, no. 1, pp. 91–102, Jan. 2020, doi: 10.1080/09720510.2020.1714150.

[32] J. Tuikkala, L. L. Elo, O. S. Nevalainen, and T. Aittokallio, "Missing value imputation improves clustering and interpretation of gene expression microarray data," *BMC Bioinf.*, vol. 9, no. 1, p. 202, Dec. 2008, doi: 10.1186/1471-2105-9-202.

[33] M. C. de Souto, P. A. Jaskowiak, and I. G. Costa, "Impact of missing data imputation methods on gene expression clustering and classification," *BMC Bioinf.*, vol. 16, no. 1, p. 64, Dec. 2015, doi: 10.1186/S12859-015-0494-3.

[34] S. Løkse, F. Maria Bianchi, A.-B. Salberg, and R. Jenssen, "Spectral clustering using PCKID—A probabilistic cluster kernel for incomplete data," 2017, *arXiv:1702.07190*. [Online]. Available: http://arxiv.org/abs/1702.07190

[35] X. Yu, H. Li, Z. Zhang, and C. Gan, "The optimally designed variational autoencoder networks for clustering and recovery of incomplete multimedia data," *Sensors*, vol. 19, no. 4, p. 809, 2019, doi: 10.3390/S19040809.

[36] J. T. Chi, E. C. Chi, and R. G. Baraniuk, "K-POD: A method for K-means clustering of missing data," *Amer. Statistician*, vol. 70, no. 1, pp. 91–99, Jan. 2016, doi: 10.1080/00031305.2015.1086685.

[37] J. Li, S. Song, Y. Zhang, and Z. Zhou, "Robust K-median and K-means clustering algorithms for incomplete data," *Math. Problems Eng.*, vol. 2016, Dec. 2016, Art. no. 4321928, doi: 10.1155/2016/4321928.

[38] S. Wang, M. Li, N. Hu, E. Zhu, J. Hu, X. Liu, and J. Yin, "K-means clustering with incomplete data," *IEEE Access*, vol. 7, pp. 69162–69171, 2019, doi: 10.1109/Access.2019.2910287.

[39] A. Lithio and R. Maitra, "An efficientk-means-type algorithm for clustering datasets with incomplete records," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 11, no. 6, pp. 296–311, Dec. 2018, doi: 10.1002/Sam.11392.

[40] S. Datta, S. Bhattacharjee, and S. Das, "Clustering with missing features: A penalized dissimilarity measure based approach," *Mach. Learn.*, vol. 107, no. 12, pp. 1987–2025, Dec. 2018, doi: 10.1007/S10994-018-5722-4.

[41] L. Abdallah and I. Shimshoni, "Clustering algorithms for incomplete datasets," in *Recent Applications in Data Clustering*, H. Pirim, Ed. London, U.K.: Intech, 2018.

[42] R. Core and R. Team, *A Language and Environment For Statistical Computing*. Vienna, Austria: Foundation For Statistical Computing, 2019. [Online]. Available: https://www.r-project.org/

[43] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. (2021). *Cluster: Cluster Analysis Basics and Extensions: R Foundation for Statistical Computing*. [Online]. Available: https://cran.r-project.org/package=cluster

[44] G. Szepannek, "ClustMixType: User-friendly clustering of mixed-type data in R," *R J.*, vol. 10, pp. 200–208, Dec. 2018, doi: 10.32614/Rj-2018-048.

[45] S. Iovleff. (2019). *Mixall: Clustering and Classification Using Model-Based Mixture Models: R Foundation For Statistical Computing*. [Online]. Available: https://cran.r-project.org/package=mixall

[46] G. Revillon and A. Mohammad-Djafari, "A complete classification and clustering model to account for continuous and categorical data in presence of missing values and outliers," *Proceedings*, vol. 33, no. 1, p. 23, Dec. 2019, doi: 10.3390/Proceedings2019033023.

[47] C. Hennig, "Clustering strategy and method selection," in *Handbook of Cluster Analysis*, vol. 9, C. M. Hennig, M. Meilă, F. Murtagh, and R. Rocci, Eds. Boca Raton, FL, USA: Chapman & Hall, 2016, pp. 703–730.

[48] M. Meilă, "Criteria for comparing clusterings," in *Handbook Of Cluster Analysis*, vol. 9, C. M. Hennig, M. Meilă, F. Murtagh, and R. Rocci, Eds. Boca Raton, FL, USA: Chapman & Hall, 2016, pp. 619–635.

[49] M. Meilă, "Comparing clusterings—An information based distance," *J. Multivariate Anal.*, vol. 98, no. 5, pp. 873–895, May 2007.

[50] 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, doi: 10.1109/Tsmcb.2012.2220543.

[51] C. Hennig, "Cluster validation by measurement of clustering characteristics relevant to the user," 2017, *arXiv:1703.09282*. [Online]. Available: http://arxiv.org/abs/1703.09282

[52] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107–145, Dec. 2001.

[53] U. Von Luxburg, *Clustering Stability: An Overview*. New York, NY, USA: Now, 2010.

[54] M. Halkidi, M. Vazirgiannis, and C. Hennig, "Method-independent indices for cluster validation and estimating the number of clusters," in *Handbook Of Cluster Analysis*, vol. 9, C. M. Hennig, M. Meilă, F. Murtagh, and R. Rocci, Eds. Boca Raton, FL, USA: Chapman & Hall, 2016, pp. 595–618.

[55] H. A. Pathberiya. (2016). *Disimformixed: Calculate Dissimilarity Matrix For Dataset With Mixed Attributes [Computer Software Manual]: R Foundation For Statistical Computing*. [Online]. Available: https://cran.r-project.org/package=disimformixed

[56] Q. Foguet-Boreu, C. Violán, T. Rodriguez-Blanco, A. Roso-Llorach, M. Pons-Vigués, E. Pujol-Ribera, Y. Cossio Gil, and J. M. Valderas, "Multimorbidity patterns in elderly primary health care patients in a south mediterranean European region: A cluster analysis," *PLoS ONE*, vol. 10, no. 11, Nov. 2015, Art. no. e0141155, doi: 10.1371/Journal.Pone.0141155.

[57] M. Guisado-Clavero, A. Roso-Llorach, T. López-Jimenez, M. Pons-Vigués, Q. Foguet-Boreu, M. A. Muñoz, and C. Violán, "Multimorbidity patterns in the elderly: A prospective cohort study with cluster analysis," *BMC Geriatrics*, vol. 18, no. 1, p. 16, Dec. 2018, doi: 10.1186/S12877-018-0705-7.

[58] A. E. M. Kneppers, R. A. M. Haast, R. C. J. Langen, L. B. Verdijk, P. A. Leermakers, H. R. Gosker, L. J. C. Loon, M. Lainscak, and A. M. W. J. Schols, "Distinct skeletal muscle molecular responses to pulmonary rehabilitation in chronic obstructive pulmonary disease: A cluster analysis," *J. Cachexia, Sarcopenia Muscle*, vol. 10, no. 2, pp. 311–322, Apr. 2019, doi: 10.1002/Jcsm.12370.

[59] C.-S. Yu, C.-H. Lin, Y.-J. Lin, S.-Y. Lin, S.-T. Wang, J. L Wu, M.-H. Tsai, and S.-S. Chang, "Clustering heatmap for visualizing and exploring complex and high-dimensional data related to chronic kidney disease," *J. Clin. Med.*, vol. 9, no. 2, p. 403, Feb. 2020, doi: 10.3390/Jcm9020403.

[60] M. Lenart, N. Mascarenhas, R. Xiong, and A. Flower, "Identifying risk of progression for patients with Chronic Kidney Disease using clustering models," in *Proc. IEEE Syst. Inf. Eng. Design Symp. (SIEDS)*, Apr. 2016, pp. 221–226.

[61] K. M. Lang and T. D. Little, "Principled missing data treatments," *Prevention Sci.*, vol. 19, no. 3, pp. 284–294, Apr. 2018.

[62] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. Zhang, A. F. Castro, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene, and J. Coresh, "A new equation to estimate glomerular filtration rate," *Ann. Internal Med.*, vol. 150, no. 9, pp. 604–612, 2009, doi: 10.7326/0003-4819-150-9-200905050-00006.

[63] J. Asher. (2020). *Transplantr: Audit and Research Functions For Transplantation: R Foundation For Statistical Computing*. [Online]. Available: https://cran.r-project.org/package=transplantr

[64] S. Van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *J. Of Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011. [Online]. Available: https://www.jstatsoft.org/v45/i03/

[65] C. Hennig and T. F. Liao, "How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification," *J. Roy. Stat. Soc., Ser. C Appl. Statist.*, vol. 62, no. 3, pp. 309–369, May 2013, doi: 10.1111/J.1467-9876.2012.01066.X.

[66] D. Mcparland and I. C. Gormley. (2017). *Clustmd: Model Based Clustering For Mixed Data: R Foundation For Statistical Computing*. [Online]. Available: https://Cran.R-Project.Org/Package=Clustmd

[67] L. Mouselimis. (2020). *Clusterr: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids: R Foundation For Statistical Computing*. [Online]. Available: https://Cran.R-Project.Org/Package=Clusterr

[68] C. Hennig. (2020). *FPC: Flexible Procedures For Clustering: R Foundation For Statistical Computing*. [Online]. Available: https://Cran.R-Project.Org/Package=Fpc

[69] B. Zhu, W. Liu, D. Yu, Y. Lin, Q. Li, M. Tong, and Y. Li, "The association of low hemoglobin levels with IgA nephropathy progression: A two-center cohort study of 1,828 cases," *Amer. J. Nephrol.*, vol. 51, no. 8, pp. 624–634, 2020, doi: 10.1159/000508770.

[70] V. F. Feteh, S.-P. Choukem, A.-P. Kengne, D. N. Nebongo, and M. Ngowe-Ngowe, "Anemia in type 2 diabetic patients and correlation with kidney function in a tertiary care sub-saharan african hospital: A cross-sectional study," *BMC Nephrol.*, vol. 17, no. 1, p. 29, Dec. 2016, doi: 10.1186/S12882-016-0247-1.

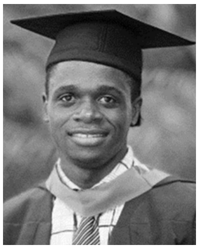
[71] J. R. Cheema, "A review of missing data handling methods in education research," *Rev. Educ. Res.*, vol. 84, no. 4, pp. 487–508, Dec. 2014.

[72] L. Kaufman and P.J. Rousseuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, 99th ed. New York, NY, USA: Wiley, 2009.

[73] M. Gagolewski, M. Bartoszuk, and A. Cena, "Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm," *Inf. Sci.*, vol. 363, pp. 8–23, Oct. 2016, doi: 10.1016/j.ins.2016.05.003.

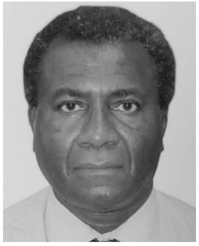
[74] M. Gagolewski. (2021). *Geniclust: The Genie++ Hierarchical Clustering Algorithm with Noise Points: R Foundation for Statistical Computing*. [Online]. Available: https://cran.r-project.org/package=genieclust

[75] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Commun. ACM*, vol. 29, no. 12, pp. 1213–1228, Dec. 1986, doi: 10.1145/7902.7906.



PETER A. POPOOLA received the B.Sc. degree in computer science from the University of Jos, Nigeria, in 2014, and the M.Sc. degree *(cum laude)* in computer science from the University of KwaZulu-Natal, Durban, in 2016, where he is currently pursuing the Ph.D. degree in computer science. His research interests include clustering, missing data treatment methods, soft computing, healthcare, and machine learning.

JULES-RAYMOND TAPAMO (Member, IEEE) is currently a Professor of computer science and engineering with the School of Engineering, University of KwaZulu-Natal, South Africa. His research interests include image processing, computer vision, machine learning, biometrics, intelligent monitoring, activity recognition, surface characterization, and formal methods. He is a member of the IEEE Computer Society, IEEE Signal Processing Society, IEEE Geoscience and Remote Sensing Society, IEEE Computational Intelligence Society, and the ACM.



ALAIN G. ASSOUNGA received the Ph.D. degree in immunology and molecular biology. He is the Head of the Department of Nephrology, Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, and the Chief Specialist at the Department of Health, Inkosi Albert Luthuli Central Hospital. Originally from Congo-Brazzaville, he received his medical training in Brazzaville, nephrology and immunology training in France and USA. He lectured and practiced medicine in the Congo and Botswana before moving to Durban, South Africa, where he teaches nephrology and immunology and leads a research team. He has trained over 20 nephrologists. He is the Editor-in-Chief of *African Journal of Nephrology*, and the Official *Journal of the African Association of Nephrology*.