Investigating Cluster Ensemble Methods to Develop Physician Phenotypes Based on Pathology Test Ordering Patterns

by

Noveenaa Pious

Submitted in partial fulfilment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
March 2021

To mom, dad and sister

Thank you for all the love and support!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Pathology laboratory testing is central to medical practice as most diagnostic and therapeutic decisions are guided by the patient's bloodwork results. Pathology laboratory tests are ordered by clinicians, and it has been observed that a significant number of tests ordered by physicians are *inappropriate*—i.e. the test is redundant, not clinically relevant, or not compliant with clinical guidelines. Inappropriate pathology test ordering not only affects laboratory utilization, but it also compromises patient safety by producing falsely abnormal results which may require unnecessary interventions. Recent laboratory utilization studies point to a discretionary behaviour in ordering tests which can be modified by providing physicians with peer comparisons, targeted education and an audit of physician's test ordering profile.

In this thesis, we aim to stratify physicians based on their patient case-mix as opposed to their order type and volume (which is circumstantial and inconclusive as comparator variables). The ensuing physician stratification will be used to generate physician phenotypes to both understand the physician's ordering behaviour and to perform peer comparisons with a similar patient case-mix.

Using pathology test ordering data spanning 6 years (2012-2017), we developed physician clusters for three temporal cohorts—i.e. 1-year, 2-year and 6-year—to track variations in the test ordering over time. We pursued a machine learning approach to investigate the phenotypical factors of physician ordering. We applied an ensemble clustering approach using three centroid models k-means, k-medoids and affinity propagation. We found the best physician clusters at k= 3 for 1st cohort, k= 4 for the 2nd cohort and k=3 for the 3rd cohort. We observed that ensemble clustering approach achieved the best results, compared to individual clustering algorithms in terms of cluster stability. We identified physician phenotypes, which interestingly change over time, which provides clear indications of underlying factors contributing to physicians test ordering pattern.

LIST OF ABBREVIATIONS AND SYMBOLS USED

PCA	Principal Component Analysis
DAE	Deep Auto-Encoder
LCMC	Local Continuity Meta Criterion
$AUC_{ln\_K}(R_{NX}(K))$	Area under the R_{NX} curve
CSPA	Cluster-based Partitioning Algorithm
LCE	Link-Based Cluster Ensemble
LCA	Latent Class Analysis
CDF	Cumulative Distribution Function
PAC	Proportion of Ambiguous Clustering
t-SNE	t-distributed Stochastic Neighbor Embedding
AP	Affinity Propagation
LDR	Linear Dimensionality Reduction
NLDR	Non-Linear Dimensionality Reduction
k-NN	k-nearest neighbours

ACKNOWLEDGEMENTS

I would like to thank God who has been the supreme protector and a constant helper throughout the entire phase of my research during the pandemic. I would like to express my gratitude and sincere thanks to supervisor Dr Syed Sibte Raza Abidi for his immense knowledge, valuable insights in formulating my research questions. He was patient in handling my writing style, technical presentation, practical feedbacks and trained to reason my results during virtual meetings. Without his persistent guidance this thesis would have not been possible.

I am extremely grateful to Dr Patrice Roy for his productive insights and opening my thoughts to diverse directions and experimentation. Thanks for answering my questions patiently, always been available for virtual meeting. A special thanks to all the members of the NICHE Research Group.

I would like to thank my parents Pious and Kingsly Pious and my sister Gladinaa Pious who have been the pillars of strength, for their constant love, moral support and sacrifice, encouraging me to achieve the best. A special thanks to all my friends and family members who have been in contact and expressing their support. A special thanks to my friend Subathra Vijayakumar who has been my constant.

CHAPTER 1 INTRODUCTION

Pathology laboratory testing is an essential part of health care. Laboratory test ordering involves the number of tests ordered by a physician on a consistent basis for various clinical purposes such as screening, monitoring of diseases, diagnosis, and management. Pathology tests ordered by the physician on a regular basis when reviewed, resulted in inappropriate, irrelevant order behaviour[1]. Nonetheless, the existence of inappropriate test orders occurs due to numerous reasons but still remains insignificant[2]. A 15-year meta-analysis stated about 4 – 5 billion of tests were estimated to be conducted every year due to laboratory testing, thereby increasing the volume of tests in the United States [2]. If inappropriate test orders are not checked properly, it may lead to diverse effects of downstream activities, which is assumed as Ulysses syndrome[3]. The syndrome occurs if the patient has undergone an incorrect diagnosis with a false positive results due to excessive testing that was not required[4].

Laboratory utilization due to inappropriate laboratory testing orders comprises of two factors namely underutilization and over utilization. Underutilization refers to the number of test orders to be recommended, but not ordered, whereas in over utilization represents the number of tests that are ordered without being specified [2]. Over utilization outcomes in the arise of unwanted blood tests for patients, that inherently increases the possibility of false positive results or misdiagnose a disease[4,5].Previous studies have adopted electronic medical records and computerized or paper orders and yielded conflict results [6][7]. A study was conducted in USA, where over-utilization counted for about 16 percent and under-utilization accelerated to 44 percent in bleeding and thrombotic disorders for laboratory tests[8]. The WHO introduced the ‘Choosing Wisely Canada’ campaign to create awareness, for the physicians and patients, introduced a conversation that consists of **“Five Things Physicians and Patients Should Question”** to determine the inappropriate use of laboratory testing[9]. Moreover, a list of impacts was observed and has improved the patient education during this health campaign. In order to reduce over utilization in laboratory testing, predictive analytics provides an effective solution.

Predictive analytics has emerged to plays a vital role to comprehend the test ordering pattern. Primarily, machine learning techniques are used to target the test ordering behaviour of physicians to tackle the problem of overutilization[1].Clustering analysis aids in reduction of unnecessary test orders and promotes better efficiency, quality by grouping the physicians ordering based on the patient case mix among the peers over a period of time.

1.1 Research Objectives

The project's overall objective is to develop the physician phenotypes to stratify the physician test ordering pattern. In this regard, this study aims to group physicians on the basis of their test orders and patient characteristics to provide a peer comparison of test ordering patterns. Peer review provides a progressive way of comparing similar physicians against all physicians [1]. Our intent is to generate physician clusters, based on their test ordering pattern, so that peer comparison is performed with inter-cluster physicians who are deemed to have a similar practice profile. To perform unbiased physician groups, we aim to use machine learning methods, particularly unsupervised clustering methods, applied to physician's test order data. Our focus in this research, is to use an ensemble approach to clustering techniques that enhance the precision and coherence in understanding the cluster stability and provide an improved quality of solution in diverse time cohorts which in turn, aims to identify the physician phenotypes that contribute to the physician test ordering pattern.

This thesis forms a set of research questions,

a) Is ensemble clustering more robust and effective than individual clustering for identifying physicians clusters in the different time-based physicians?
b) What is the most prominent dimensionality reduction algorithms in test ordering data?
c) What are the physician phenotypes found inside the physician clusters?
d) Can we identify the distinct clusters of physicians based on test ordering pattern?

The first research question is explained in Section 4.4.4 presents a detailed comparison between the two approaches across the three time-cohorts. For the second question, Section 4.2 provides the interpretation of the best dimensionality reduction technique. For the third and the fourth research question discovers the important physician phenotypes found in the physician clusters that contribute the test ordering pattern in Section 4.6.

1.2 Solution Approach

Our approach is to investigate an ensemble clustering method by combining multiple clustering methods and analysing multiple consensus functions. We take a data mining methodology that involves data preparation, feature extraction, clustering, and evaluation. The task of feature extraction involves the exploration of different dimensionality reduction algorithms to extract the useful features suitable for clustering. Moreover, this thesis examines the evaluation of embeddings produced by introducing evaluation metrics to assess the quality of different dimensionality reduction algorithms for an unsupervised approach.

1.3 Contribution

The contributions of the thesis outline is as follows:

- Investigation and application of ensemble clustering approach to cluster high-dimension temporal pathology test order data to discover data-driven physician peer groups. Our research proposed a new approach to clustering that addresses the effects of utilizing clustering ensemble technique to determine the stability of the discovered clusters and its robustness.
- Evaluation of linear and non-linear dimensionality reduction algorithms with respect to quality assessment criteria to recognize identify the best clustering model.
- Experiments to identify temporal variations inherent in physician's test order profiles, thus generating time-dependent peer groups, where physicians may be clustered with different peer groups in different time periods. Our clustering approach has yielded physicians clusters across three (3) different time periods.

1.4 Organization of the thesis

This thesis consists of five chapters. The first chapter proposes the introduction of the thesis, the second chapter presents the primary concepts of clustering ensemble, reviews the literature prevailing to different consensus functions used in the thesis, and explores the distinct feature extraction algorithms. The third chapter demonstrates the data mining methodology of analysing and extracting features from high dimensional space in three different time-cohorts. Building a heterogeneous cluster ensemble model from the features extracted.

The fourth chapter discusses about the results and analysis i) quality assessment of feature extraction models ii) comparison between the different base clustering methods and ensemble techniques in three time-cohorts.

The fifth chapter concludes with a summary of the study's contributions, limitations and suggestions for future research.

CHAPTER 2 BACKGROUND STUDY

In this chapter we provide the background study required to understand the rest of the thesis. We start discussing the clustering ensemble. This section involves a brief summary of cluster ensemble techniques and the context of consensus functions utilized on the ensemble approach and its evaluation metrics. Then, we discuss the different feature extraction algorithms used in reducing the features as well as the experimentation of quality assessment criteria implemented in different dimensions used in the literature.

2.1 Clustering ensemble

Cluster analysis is a preeminent technique implemented in various fields of research interests namely data mining, information retrieval, image recognition to understand the underlying structure of a dataset[10]. Clustering aids in uncovering the hidden pattern by grouping similar objects into homogeneous clusters [11] by increasing the heterogeneity over clusters[12]. Many applications have cultivated the use of clustering to identify groups of psychiatric patients based on the characteristics and symptoms experienced by the patients [13] clustering a group of genes that produce the similar biological functions[14], by recognizing medical patient groups who are in need of targeted interventions [15,16]. Over the years, there are numerous conventional approaches of clustering algorithms that have been designed and developed. Diverse clustering algorithms produce various clustering results for the same dataset by applying distinct structures on the data[17] i.e. arbitrary shaped-clusters and distance-based clusters. According to the popular "no-free-lunch" allegory, there is not an individual clustering algorithm which performs best for all data sets[18]. In addition, there is no definite clustering algorithm that could yield accurate results as well as there are no standard measures to follow to select the best individual algorithm for a given problem[17].

In order to improve and produce robust outcomes, the concept of cluster ensemble was introduced. The concept of combining different clustering algorithms appear to be a possible strategy for improving the quality and stability of the discovered clusters. Cluster ensemble otherwise known as consensus ensemble refers to the process of combining multiple clustering models to a single consolidated partition[19]. The cluster ensembles strategy was initially defined for integrating multiple clustering by outputting the labels by each individual clusterer to a consensus function which yields an universal clustering [20].Moreover, it relies on the successive collaboration of supervised classifiers[10]. Topchy et al has endorsed that intelligent combination of these clusters could lead to novel and significant cluster structures, even in the existence of noise[21].

Bollacker and Ghosh designed a cluster ensemble framework for knowledge reuse, where the merger can only access the cluster labels without approaching the original features. However, developing a clustering ensemble is a difficult task as it takes the cluster labels into account as well as it has to resolve the correspondence problem[19]. The label correspondence problem illustrates the set of labels from one clustering algorithm has no relation with the clustering labels of another clustering algorithm in an ensemble[10]. But if there prevails an association between the labels, the voting technique is the most appropriate method to be used in this case[10].There are various ways to create cluster ensemble depending upon the domain and the quality of the solution required.

2.2 Conventional Properties of Cluster Ensembles

There are a variety of reasons to use a clustering ensemble in different domains designed and developed that aids to solve a problem. Fred and Jain[22] and Topchy et al[23] framed the properties of a clustering ensemble namely robustness, stability, novelty and consistency[10]. The property of robustness and consistency indicate the performance of the ensemble should have better results and the combination used must be in accordance whereas stability and novelty represent clustering solutions with less prone to outliers, which is not feasible with the individual clustering algorithms[10].These properties indicate that the process of amalgamation acts more stable than an individual clustering.

2.2.1 Improved Quality

Ludmila and Stefan [18] designed a pairwise clustering ensemble to improve the quality of solution where the overproduced clusters is chosen as ensemble member, indicates the number of clusters produced is higher than the expected number of clusters. They introduced only the fundamental clustering algorithms that improved the accuracy of the ensemble members. In order to understand the quality of solution in the accuracy between the diverse cluster ensemble and non-diverse cluster ensemble, they observed the diverse cluster ensemble performed well. Hu et al [24] developed a clustering ensemble on gene expression data in which the individual clustering algorithms outcomes were combined in the form of distance matrix by using the similarity between the two data points and graph based partitioning was applied to retrieve the final clustering results. In this study, minkowski score was significant to identify the quality of the clusters. From these studies, the cluster ensemble was recognized to enhance the standard of solution by considering the average and bias of the individual solutions [25]. The generation of high accurate and quality results is difficult to obtain in gene expression data as it contains outliers in the experimental data and there is no stability across the different clustering algorithms. An efficient way of selecting the combination of several algorithms is the best way to improve the quality and the stability of the clustering ensemble[24].

2.2.2 Robust Clustering

Sevillano et al [26] conducted experiments on clustering of documents on diverse features to yield a global clustering for a collection of documents in an unsupervised way[8].The k-means was the clusterer consisting of four clusters supplied in parallel with these document representations and consensus clustering was introduced for ten cluster runs to reduce the random initialization of k-means. Furthermore, they investigated by taking into account supervised model order selection for each representation technique by computing the normalized mutual information (NMI) between each clustering and the documents original labeling. The NMI had better performance in terms of term based representations and was ranked with tf-idf weights. In the second experiment, they did not use the term based representation whereas continued with the extracted features and the consensus labeling was applied to the graph based partitioning algorithms such as CSPA, MCLA. Here, CSPA executed better with sub-optimal clustering and showed that the consensus function worked the best with the optimal order selection. The cluster ensemble implements robustly irrespective of data dimensionality as it adapts to produce better results and outcomes across wide variety of datasets with different dimensions.

2.2.3 Multi-view Clustering

Different views and multiple feature sets are available in different applications. In market basket analysis, the customer's preferences depend on various views and characteristics. The base clustering may be constructed on distinct views that involve non-identical sets of features or subsets of data points[25]. Strehl and Ghosh constructed two types of views for effectively combining the partitions in a cluster ensemble namely Feature distributed clustering and Objected distributed clustering[19]. In feature distributed clustering, multiple clustering is composed with different subsets of features but making use of all data points. Here, the clusterer forms clusters from the subspaces by utilizing the same clustering technique[19]. In the aggregation stage they are integrated using the consensus function.

While in the prediction of gene functions, separate gene clustering could be acquired from diverse sources such as gene sequence comparisons, and microarray data consisting the combinations of DNA sequences from many independent experiments, and mining of different biological literatures such as MEDLINE requires feature distributed clustering[19]. A study on Yahoo dataset was experimented by Strehl with 20 clustering with 128 dimensions[27]. The quality of results was better as it yields 0.20 score of NMI normalized mutual information was higher than the average of individual clustering. Consensus clustering performed better in quality than the individual input clustering. On the other hand, different clustering uses different subsets of data points but readily uses all the features. In ODC, the original features are not accessed as well as the labeling obtained are biased. The consensus function produces a meaningful clustering result by combining the overlaps between the labels. This scenario happens to people who have access to more than one store in market basket analysis, overlapping tends to result in such situations. These ensemble strategies are implemented in a distributed way and re-use the knowledge accomplished from hierarchies in order to preserve privacy related cases.

Fred and Jain[22] developed an ensemble with random initialisation of K-means algorithm for multiple cluster runs and mapped into a new co-association matrix. This matrix is partitioned into final clusters using hierarchical single link algorithm. Topchy et al designed an ensemble with plural voting and introduced a metric especially on the space of partitions[21].

The study of Asur et al [21] carried out clustering ensemble using reduced dimension of Principal component analysis (PCA). The reduced dimension aided the ensemble solution to compute the consensus function more credibly. A soft clustering ensemble for protein interaction was created in order to observe the views of multifaceted proteins.

2.3 Miscellaneous Generation Techniques

Clustering ensemble works in two phases: The first step is to generate the clustering in different ways in the generation step. Second, is to join the clustered labels by using a consensus function[10]. The study of Vega-Pons and J. Ruiz-Shulcloper proposed five ways to use the generation step that combines the different clustering algorithms to be used before the consensus functions. There different ways of generating the clustering namely different clustering algorithms, the same algorithm with initialization of different parameters, different objects representations, different subsets of objects or object projections on different subspaces[10]. Topchy et al[28] explains that the weak clustering algorithms have the ability to produce better consensus results in consensus clustering, well joint with a perfect consensus function [10]. Iam and Simon[29] divided the generation method in various techniques namely homogeneous ensemble, heterogeneous ensemble, data subsampling and selection of k. In homogenous ensemble, a single clustering is used for multiple cluster runs and base clustering are produced whereas in a heterogeneous ensemble model uses multiple clustering algorithms to output a final clustering result.

Figure 2.1 depicts the overview of cluster ensemble process discussed in Iam's study[10]. In the generation step, designing a clustering ensemble by selecting subsets of data from different clustering algorithms plays an effective role. But choosing the several subsets require better computation. Consensus clustering plays a vital role for running clustering algorithms multiple times for selecting subsets of data. Jain and Dubes emphasized the fact that the sub-populations could depict the members in each sub-population share common features and properties found within larger populations. This field of study was employed in the improvement of molecular based diagnosis, prediction and treatments in cancer.

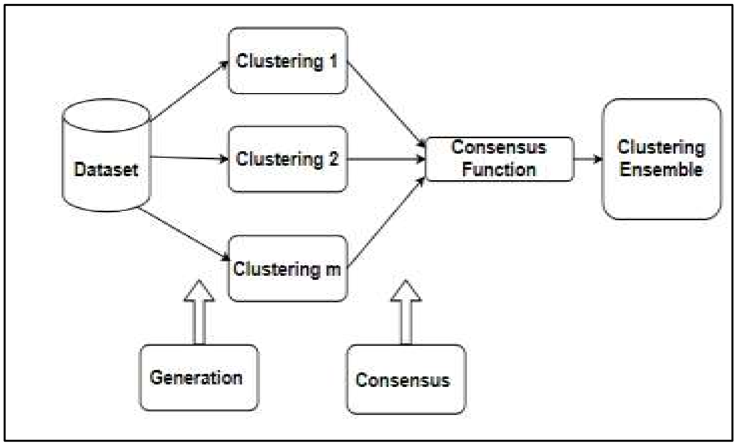

Figure 2.1 Overview of a cluster ensemble process[10]

Monti et al[30] introduced consensus clustering with the resampling techniques to evaluate the cluster stability over multiple runs of clustering algorithms with random initializations. The main reason was to determine the optimal number of clusters obtained while running the clustering algorithms multiple times, to examine the boundaries, cluster number and membership[30]. During sampling variability, if a greater number of clusters are obtained, depicts the robust nature of the clustering algorithm to the resampling technique. It represents the structure of the cluster existence.

The study was conducted based on the concept of consensus clustering with resampling for multiple cluster runs on three clustering algorithms such as Hierarchical clustering, Self-organized maps, and model-based Bayesian clustering. Consequently, there were defects in Hierarchical clustering because it was difficult to manifest the number of clusters as well as the boundaries. This behaviour of hierarchical clustering was due to deterministic nature of the agglomeration rule. The visualization of these clusters is univocal, irrespective the number of clusters given. With model-based clustering, chooses the number of clusters but results in complication due to distribution of the mixture produced. In order to understand their behaviour, cluster compactness was deployed by Dudoit and Fridlyand[31], Milligan and Cooper[32], Tibshirani, Walther and Hastie Yeung[33], Haynor and Ruzzo[34].

The cluster validation of the number of clusters formed and tracking the cluster assignments becomes an important criterion. Eventually, the absence of known class labels which occurs in supervised learning, the process of cluster evaluation becomes evasive. Many statistical procedures exist in identifying the number of significant clusters in low dimensional data, but it lacks clarity in high dimensional data (Bock 1985). Another possible approach in cluster validation for resampling based techniques is consensus matrix.

2.3.1 Generation Mechanism

The study of Monti et al [30] resampling consists of partitioning the dataset into set of clusters non-overlapped in nature. The resampling techniques takes place in two ways bootstrapping and subsampling. Efron and Tibshirani instituted bootstrapping, where the items are sampled with replacement from the original dataset, but this method results in defects as it takes identical replicates of the same item are chosen during every iteration leading to increase in the compactness of the clusters. The main focus of the study will be on subsampling, as it takes the subsets of data as it samples without replacements from the original dataset.

Suppose there exists a dataset D = {d_1, d_2, d_3.... d_N } where d_1, d_2 represent a set of items to be used for clustering[30]. The K-formation of cluster partition E of D is defined by[30],

E ≡ {E_1, E_2, E_3 .... E_k } such that $\cup_{k=1}^{K}$ E_k = D and $E_i \cap E_j \neq \emptyset\ \forall_{i,j}$ such that $i \neq j$. In the subsampling process, from the cluster partition of the items i and j are intersected if they occur in the same cluster, giving rise to the value 1 and otherwise its 0, because all samples will not be incorporated.

Here, the context of gene expression was taken as example. The number of genes to be clustered will be taken as items indicating ($d_i \in D$). The features are observations expressed. over many experiments.

Consensus matrix was utilized for assessing the correspondence in a dataset for multiple clustering runs. It is matrix ($N \times N$) that holds two clustered items that occur together during the distribution of cluster runs and the average of the connectivity matrices gives the result of consensus matrix.

After the application of clustering algorithm on the dataset, the formation of connectivity matrix is given by the condition in Equation 2.1,

Equation 2.1

$$\mathbf{M}^{(h)}\ (\mathbf{i}\ ,\mathbf{j}) = \left\{\begin{matrix} \mathbf{1}, \boldsymbol{if\ i\ and\ j\ belong\ to\ \ the\ same\ cluster} \\ \mathbf{0}\ \boldsymbol{otherwise} \end{matrix}\right\}$$

Here $I^{(h)}$ is the indicator matrix that keeps in track of the items i and j occurring together with respective to resampling techniques such that it takes the subsamples from the original dataset, not all samples are included in the analysis. The number of iterations should be taken into account to keep the track the of two items appear jointly to form resampled dataset. Consensus matrix is defined the number of times items i and j gets allocated to the same cluster during each cluster run divided by the total number of times both items are chosen given in the Equation 2.2,

Equation 2.2

$$\boldsymbol{M}\ (\mathbf{i},\mathbf{j}) = \frac{\boldsymbol{\Sigma_h}\mathbf{M}^{(\mathbf{h})}(\mathbf{i},\boldsymbol{j})}{\boldsymbol{\Sigma_h I^{(h)}}(\mathbf{i},\boldsymbol{j})}$$

Here $\boldsymbol{M}$ (i, j) is the normalized sum of all connectivity matrices of the dataset given. The consensus index was initiated to indicate the entries of i, j appearing in the consensus matrix. Monti et al developed the visualisation pattern of a perfect consensus matrix represented in the form of dendograms with non-overlapping blocks along the diagonal depicts 1 that belong to the same cluster whereas the block referring to different cluster as 0. In terms of perfect consensus matrix would represent either 1 or 0 only. They carried out the study on two different datasets for 500 iterations for K=3, in which one dataset had no structure with one block representing the only cluster formation. Conversely, the other dataset gave good results of a three-diagonal structure. The consensus matrix determines the best item order, and its visualization specifies the stability of the clusters found during multiple cluster iterations.

2.4 Classical Clustering Approaches

2.4.1 k-means

k-means is one of the most popular method from partitioning based clustering algorithms and widely used in health care[35] by grouping objects when the number of clusters are predefined[36].One of the advantages, the squared error difference between the mean of the cluster and the data points in that cluster is minimized in k-means[37]. Escudero et al utilized the k-means algorithm for detecting the onset of Alzheimer's Disease by clustering patients into pathology and non-pathology profile [38].The k-means algorithm is improvised to predict diseases from hemogram blood test samples by using weighted k-means algorithm in order increase the consistency and efficiency of the clusters[39]. Elbattah et al[40] in his study clustered elderly patients based on age characteristics to detect the inpatient LOS (length of stay)[40].

2.4.2 k-medoids

k-medoids or partitioning around medoids is a centroid based model, a variation of k-means, which chooses the members established on a minimum average cost within the cluster to be the centroid of the cluster in the next iterations[41]. Irwansyah et el [42] grouped the patients with cardiovascular disease to obtain the levels of complications inside the clusters using k-medoids that achieved a silhouette coefficient of 0.35[43]. Acharya [44] used k-medoids for the robust nature to noise and outliers, to obtain WBCs from a image[44].

2.4.3 Affinity Propagation

The affinity propagation method considers the data members in a network, by exchanging information during each iteration continues until the presence of good set of exemplars and clusters. Li et al grouped a set of brain images into different cluster partitions using affinity propagation to identify the similarity between the images. The affinity propagation has the potential to determine the number of clusters automatically[45]. Buch et al used affinity propagation to cluster bacteria for the robustness of the similarity measures with respect to the responsibility matrix and availability matrix used in the algorithm and observed that affinity propagation has the capacity to examine the humongous datasets with a greater speed[46].

2.5 Novel techniques in Consensus Functions

The final step in a clustering ensemble is the consensus function that presents the final clustering result from the combination of multiple clustering algorithms. Vega-Pons and J. Ruiz-Shulcloper introduced two approaches object co-occurrence and median partition.

The first approach consists of consensus partition that decides which cluster labels belong to each object. This approach follows the voting process, where each object will vote to the cluster that belongs to the consensus partition. Moreover, two methods are based on the voting process namely Relabeling and Voting method and Co-association matrix method. The second approach determines the median partition which maximizes the similarity measures in all the partitions of a clustering ensemble [10]. The median partition is given by the formula in Equation 2.3,

Equation 2.3

$$\boldsymbol{P * = argmax_{P \in P_j} \sum_{j=1}^{m} \Gamma(P, P_j)}$$

where $\boldsymbol{\Gamma}$ is the similarity measure. The median partition is NP-hard problem studied by Krivanek and Moravek and Wakabayashi. Topchy et al showed the validity of both the approaches, the outcome of consensus depends on the clustering solution produced by the clustering algorithm due to which the number of partitions in the ensemble expands. There are diverse consensus functions introduced over the recent years.

2.5.1 k-modes

The study of Zhexue et al devised a clustering algorithm that could perform well and faster in both categorical and numeric data in larger datasets. The k-modes were an extended version of k-means algorithm. The main objective was to minimise the cost function of the clustering process by utilizing a dissimilarity measure and replacing the means of a cluster by mode function and use a frequency-based method to update the modes. Initially, (Huang 1997) proposed k-prototypes which uses a dissimilarity measure for numeric attributes determined by the squared Euclidean distance denoted as s_n and for categorical attributes s_c decides the dissimilarity measure based on the number of mismatches of categories between two objects [47]. Moreover, a weight is assigned to circumvent favouring of the types of the attribute. Zhexue et al proposed a new method which updates the categorical attribute and selects the weight based on the use of average standard deviation. One of the greatest advantages of using k-modes is its scalable in large datasets.

The algorithm starts by taking k initial modes and assigning the objects to the nearest mode which is calculated by the distance d and reforming the modes for each assignment based on the set theory of modes. Following that, re-assessments take place if the current mode belongs to another cluster, re-allocation is made and continues until no changes are observed in the entire dataset. The dissimilarity measure is calculated by the number of mismatches between the two objects, the objects will be similar if the mismatch is less given by the measures in Equation 2.4 and Equation 2.5

Equation 2.4

$$\boldsymbol{d}(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^{m} \boldsymbol{\delta}(\boldsymbol{x_j}, \boldsymbol{y_j})$$

Equation 2.5

where, $$\boldsymbol{\delta}(\boldsymbol{x_j}, \boldsymbol{y_j}) = \begin{cases} \mathbf{0}, & (\boldsymbol{x_j} \neq \boldsymbol{y_j}) \\ \mathbf{1}, & (\boldsymbol{x_j} = \boldsymbol{y_j}) \end{cases}$$

d (X, Y) represents the dissimilarity measure between two categorical objects. By introducing two methods for initial mode selection, one that frequently takes the first distinct objects as initial mode and the other method by calculating the frequencies of all categories in descending order and allocates the most frequent ones as the initial mode. According to the study, the k-modes performs faster and clusters larger datasets containing millions of objects easily. Sun et al. (2002) applied k-modes and used subsamples depending on the size of the dataset, where subsamples were joined to a single set for a solution. Wu et al. (2007) introduced a new initialization method by using maximum probability by taking into account the product of density of each point[48].

2.5.2 Cluster-based Similarity Partitioning Algorithm (CSPA)

One of the consensus functions that comes under the category of co-occurrence of objects is Cluster based Similarity Partitioning algorithm. Strehl and Ghosh developed a clustering ensemble by modifying the clustering labels in the form hypergraphs. Hypergraphs consists of edges which can connect to any set of vertices. The entries in hypergraphs is a binary matrix $(n \times n)$ entered by checking the two objects in the same cluster gives a similarity measure 1 otherwise it outputs 0. The overall similarity matrix S is calculated by taking the average of the entries in the same cluster given by the formula in Equation 2.6,

Equation 2.6

$$\mathbf{S} = \frac{1}{r}\, HH^{T}$$

where S is the similarity matrix and r are the number of clustering and re-allocation of clusters takes place by similarity based clustering algorithm METIS. It is the simplest approach and quadratic in complexity and storage. The study of Faisal et al [49] used CSPA to improve the functionality of chemical structures in drug data report using consensus clustering. They used individual clustering algorithms with six distance measures for different threshold values to form a six-ensemble approach. The study concluded that CSPA works to produce stable clusters by reducing sensitivity to outliers in comparison with the other methods.

2.5.3 Link-Based Cluster Ensemble (LCE)

The study of Iam-on et al[29] designed link-based similarity measure to define a clustering ensemble. The link-based similarity matrix was used to improve the similarity values between the data points and will be created from the base clustering. Iam-on at al used various similarity measures namely Connected–Triple-based similarity (CTS), SimRank based similarity (SRS) and Approximate SimRank based similarity, an enhanced version of SRS. These similarity measures provide inherent relationships which is not possible with co-occurrence strategy. The connected triple (CTS) method was developed by Klink et al to evaluate the duplicates between the author names[50]. In the analysis of linking two authors, it consists of a graph (V, E) where each vertex coincides with an appropriate name and the edges join the two authors based on the information of the publication and calculated by the similarity measure in terms of connected triples. The SimRank (SRS) was initiated by Jeh and Widom and proposed as the standard method. The concept of similarity is based on the presumption that the neighbours are alike if their neighbours are similar. The bipartite graph includes a set of vertices V representing both the data points as well as the clusters in the ensemble, whereas E is a set of edges between the clusters and data points to which they are allocated to it.

The SRS matrix is the similarity matrix indicates the similarity between any two clusters or data points in the ensemble, given by the formula in Equation 2.7,

Equation 2.7

$$\mathbf{SRS\ (a,b)} = \frac{\mathrm{DC}\quad \sum_{a' \epsilon N_a} \sum_{b' \epsilon N_b} SRS(a',b')}{|N_a||N_b|}$$

where DC denotes constant decay factor ranging (0,1] constituting the data points connected to the set of vertices. Moreover, the recent studies of Iam-On & Boongoen, includes the data analysis of microarray experiments that works on feature transformation.

2.5.4 Latent Class Analysis (LCA)

Latent class analysis was designed for observing multivariate categorical data[51]. It is mainly used to observe the statistics of unobserved groups in the data [14]. The algorithm has been carried out in variety of use cases in categorizing responses in public opinions surveys, individual-level voting data, consumer behaviour and decision making by clustering similar use cases and understanding the examination of the distribution[51]. Niels in his study recognized that LCA was utilized to inspect the unobserved target-categories in a marketing firm by observing the diverse attitude structures of customers on the decision making in order to purchase any item [14]. The LCA represents a finite mixture model that considers the distribution of components as a multi-way cross classification in which all the variables are mutually independent[51]. The latent variables could be only inferred indirectly from other variables through a mathematical model that is observed. Dayton and Macready in their study, examined the responses received on a matrix algebra test, where the latent variables correspond to the knowledge of the matrix algebra of students and the latent classes indicate the masters and non-masters on matrix algebra. With respect to class membership, the conditional probabilities help identify various possibilities of certain answers are selected. The observed variables act statistically independent with each latent class. Drew et al introduced Polytomous Variable Latent

Class Analysis which utilizes the maximum likelihood estimates of the model parameters in the EM and Newton-Raphson algorithms given by the condition in Equation 2.8.

Equation 2.8

$$\mathbf{In\ L} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{ln} \sum_{\mathbf{r=1}}^{\mathbf{R}} \mathbf{p_r} \prod_{\mathbf{j=1}}^{\mathbf{J}} \prod_{\mathbf{k=1}}^{\mathbf{K_j}} (\boldsymbol{\pi}_{\mathbf{jrk}}) \mathbf{Y_{ijk}}$$

where J represents the polytomous categorical variables otherwise called as manifest variables containing K_j with j possible outcomes, Y_{ijk} constitutes the values observed in J manifest variables, π_{jrk}stands for class-conditional probability, p_r indicating the prior probabilities of latent class membership[51].Vidden, Vriens and Chen used latent class model to recover the identity of cluster members and stability to recognize the accurate number of clusters than the K-means clustering algorithm[52].

2.5.5 Majority Voting

Ayad and Kamel suggested a relabeling and voting technique for solving a correspondence problem that appears when the consensus is partitioned. One of the main issues faced is the label correspondence that constructs the unsupervised combination to be difficult[10]. By analysis, the voting strategy and bipartite scheme could be applied to hard or soft ensembles[53].In most of the cases, an iterative pairwise relabeling is employed in voting-based aggregation problem[54]. Dudoit and Fridlyand proposed a type of consensus which is similar to plurality voting in classifier ensembles[53]. Fischer and Buhmann [54]discussed the plurality voting for choosing the winning cluster for each object. Dimitriadou et al obtained a voting algorithm by reducing the squared-distance criteria between an ensemble of hard or fuzzy partitions and optimal fuzzy consensus [54].

The main idea is to permute the cluster labels such that best consensus between the cluster labels of two partitions is obtained and all the partitions should be relabeled according to a fixed reference partition from the ensemble[55].They introduced a new strategy in which a voting works as a multi-response regression problem and bipartite matching. The cumulative voting introduced in [54] works as a linear regression problem. The main function of cumulative voting follows two approaches relabeling and aggregation of the consensus partition. In relabeling, the most appropriate relabeled partitions is taken, the problem was viewed as supervised learning problem with continuous response variables leading to soft relabeled partition[53].

In aggregation, the ensemble partition is run multiple times with random ordering, in which Ayad and Kamel proposed two algorithms namely bVote and Ada-cVote. Moreover, Ada-cVote performs well and saves computational time than bVote. The study was applied on three artificial datasets and three real datasets namely Yahoo, E. coli proteins and LandSat in which Ada-cVote achieved greater accuracy when compared with other ensembles such as CSPA, MCLA and HGPA.

2.6 Performance Metrics

2.6.1 Cumulative Distribution Function (CDF)

Monti et al study initiated to evaluate the stability of clusters, as well as to obtain the optimal number of clusters in consensus clustering by taking a consensus distribution with the help of a histogram from consensus matrix. The histogram produces two bins 0 and 1 for good consensus clustering, which in turn plotted as an empirical cumulative distribution (CDF) ranging between 0 and 1 given by the formula in Equation 2.9,

Equation 2.9

$$\mathbf{CDF(c)} = \frac{\sum_{i<j} \mathbf{1}\{\boldsymbol{M(i,j)} \leq \boldsymbol{c}\}}{\boldsymbol{N(N-1)/\,2}}$$

where 1{. . .} represents the indicator function, ***M*** (i, j) illustrates the consensus matrix (i, j) and *N* delineates the number of rows and columns of *M*. The CDF graph curves illustrates a step function across 0 and a flat line passing between 0 and 1, and a second step function around 1. If the curve gradually climbs and constitutes a different a shape, then the clusters formed lacks the characteristics of stability. If the CDF curves forms a shape of bimodality, it estimates the presence of significant clusters.

2.6.2 Proportion of Ambiguous Clustering (PAC)

The study of Monti et al examined the optimal number of clusters in consensus clustering. But the method lacked sensitivity and specificity in assessing the optimality of clusters. Senbabaoglu extended the study by introducing a new metric called the Proportion of ambiguously clustered pairs (PAC) which evaluates the optimal number of clusters and the stability. The assessment of cluster strength is quite difficult in a hypothesis testing framework because of the unique covariance structure, and the other factors such as sensitivity and specificity that influence the cluster results[56]. According to the study, they introduced a condition in which all pairs of samples, the calculation of consensus rate defines the frequency in which a pair of samples is clustered together in multiple cluster runs with each run occurring with certain possibility of degree of permutation taken by random initialization.

The main hypothesis of this study is that there exist well separated and stable clusters for the true K value taken from the different subsamples and the visualization was illustrated in the form of consensus heat map. They introduced four methods for finding the optimal K value namely Cumulative Distribution function, the area of change under the CDF curve upon an increase of K, GAP statistic and CLEST. The GAP statistic method was used to determine the optimal number of clusters by taking the consideration of $\log(W_k)$ and differentiating with null distribution[33]. On the other hand, CLEST method estimates the number of clusters based on resampling technique, specifically designed for forecasting the cluster assignments[31]. Out of which, they chose the CDF curve method that predicts the true value K and performs well than the other metrics. In CDF curve of a consensus matrix, predicted the portions with samples of pairs that were hardly clustered in lower left side whereas the upper right side was always clustered together. The co-assignments appear in the middle portion for different cluster runs. The existence of a flat line in the middle segment stands for the true K, become ambiguous for a rare number of sample pairs observed. The PAC value is calculated using consensus index that refers to the fraction of sample pairs with values ranging in the intermediate sub-interval (x1, x2) g [0, 1] [56]. The lowest possible PAC value indicates the optimal value of K showing the flat line segment that appears in the middle proportion. Merely, the PAC outperformed well than the other methods when checked with simulated dataset. But no method is commonly the best according to the study.

2.6.3 Internal Evaluation Indices

Jain et el proposed the internal evaluation criteria for final data partition when the true cluster labels are not known by assessing the quality of a data partition based on the quantities and features assumed from the data[29].The evaluation is more reliable and provides efficiency in various scenarios as it does not use reference labels which is not possible to acquire in some cases. Assessing the quality of a clustering ensemble depends on the final data partition produced by the ensemble. The main factors that influence the internal evaluation indices are Compactness, Dunn index, Silhouette, Calinski Harabarz and Connectivity.

2.6.3.1 Compactness

In the study of Nguyen and Caruana[29] (2007) describes compactness as inherent property that measures the average distance between every pair of data points, which lie in the same cluster[29], the data points in the cluster will be close to each other which is given by the formula in Equation 2.10,

Equation 2.10

$$\mathbf{CP}\ (\pi\ *) = \frac{1}{N} \sum_{k=1}^{K} n_k \ (\frac{\sum_{x_i x_{j \in C_k}} d(x_{i,} x_j)}{n_k(n_k-1)/2})$$

where N represents the total number of data points in the data, K indicates the number of clusters, and n_k is the number of data points belonging to the k-th cluster and $d(x_{i,} x_j)$ denotes the distance between the data point $x_i \ and \ x_j$. The compactness value should be low as possible for better clustering results.

2.6.3.2 Dunn Index

Dunn introduced the Dunn index to recognize the compactness and well separated clusters[29] given by the condition Equation 2.11,

Equation 2.11

$$\mathbf{D} = \frac{\min_i min_j(min_{x \in Ci, y \in Cj} \ d(x,y))}{max\ _k(max_{x,y \in Ck} \ d(x,y))}$$

The larger the value of Dunn index indicates optimal clusters as it takes the inter-cluster separation and intra-cluster compactness[57].

2.6.3.3 Silhouette Index

Rousseeuw et al[41] proposed silhouette index that depicts the closeness of well separated clusters. Silhouette index indicates the average distance to objects in the same cluster as well as the distance to the objects in the alternate clusters. The silhouette index is calculated as in Equation 2.12 [58],

Equation 2.12

$$s(i) = \frac{b(i) - a(i)}{max\,\{a(i), b(i)\}}$$

The silhouette range lies between -1 and 1, high values represent good clustering with no overlaps between the clusters. In terms of convex clusters, silhouette score is low. The silhouette index is to be maximized for forming well separated clusters.

2.6.3.4 Calinski Harabasz

The Calinski Harabasz takes into account compactness and separation parallel within the cluster sum of squares[57]. Its ratio of separation by compactness where the degree of separation inspects the cluster centre extends and how the in-cluster objects are close to the cluster centre[57]. The CH index is calculated by the separation metric in the Equation 2.13[58].

Equation 2.13

$$\boldsymbol{CH} = \frac{\frac{\boldsymbol{SSBM}}{(\boldsymbol{M}-\boldsymbol{1})}}{\frac{\boldsymbol{SSEM}}{(\boldsymbol{M})}}$$

Higher the value of this index provides better compactness and separation. Calinski concludes that the index presents better execution and prohibits the common errors in centroid based clustering models.

2.6.3.5 Connectivity

Connectivity defines the neighbouring data items that share the same cluster[59]. Handl et el in his study proposes that the index is used for clustering algorithms of arbitrary shapes and provides a degree partitioning that captures the local densities and number of items grouped together in the data with the nearest neighbours. Connectivity is given by

Equation 2.14

$$\mathbf{C} = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nni_{(j)}}$$

The connectivity index should be minimum as possible having a value between 0 and 1.

2.7 Feature Extraction Techniques

Feature extraction is an essential step for handling high dimensional data. In a medical dataset, a large number of features are found in a gene expression data, clinical scores and medical imaging. Feature reduction aids in solving the curse-of-dimensionality problem by reducing the redundant features, noise and avoids over-fitting, improves the accuracy in a machine learning model[60]. The use of feature extraction in clustering is influential, because it's easier to find clusters in low-dimensional space, but it's quite challenging to discover the clusters in a high dimensional space due to existence of highly skewed and sparse data[61]. As the number of features increases, it's harder to understand the patterns[62].One of the primary purposes of dimensionality reduction is to decrease the computational complexity in data pre-processing and extract useful features with no prior loss of information[63].

2.7.1 Linear and Non-linear Dimensionality Reduction Algorithms

Dimensionality reduction is a extensively used approach to find interpretable representations of data in low dimensional space that are projected from high-dimensional spaces[64]. The use of dimensionality reduction algorithm is influential in medical applications, as the data is often dealt with high dimensional data consisting of proteomic data, raw hospital records and medical images [65]. Maaten et al[66] proposed a taxonomy and divided dimensionality algorithms into convex and non-convex dimensionality reduction algorithms based on local optima criteria. Friedman in his study, points that dimensionality reduction algorithms may help in predictive modelling draws on the bias and variance trade-off in predictive error. The latent structure in the data could be visualised by the algorithms [67].

Many dimensionality reduction algorithms have emerged with linear dimensionality algorithms with popular being principal component analysis and classical metric multidimensional scaling whereas non-dimensionality reduction algorithms with

Sammon's nonlinear mapping and auto-encoders. With non-linear dimensionality reduction algorithms, the concept of manifold learning tries to reduce the dimension without disrupting the topological properties of the data.

2.7.2 Principal Component Analysis (PCA)

PCA is a widely used techniques in feature reduction algorithms. Zhang et al in his clinical studies containing electronic healthcare records (EHR) consisting of large number of variables used PCA to solve the multi-collinearity problem[68]. Laura and Matthew utilized PCA for demonstrating the variance of a normal tissue cDVHs with the first principal component showing the large variation.[69]. PCA is utilized for reducing a huge set of correlated variables to a less number of uncorrelated components[70]. Rohan and Amarda suggested the performance of PCA was better when compared with other dimensionality reduction algorithms using the metrics such as Trustworthiness, LCMC and Continuity for the de novo protein structure[71].

2.7.3 Isomaps

In order to overcome the shortcomings of spatial metrics, geodesic distances were initiated[63]. The most important feature of Isomaps is it replace the conventional distance measurement, such as Euclidean distance between data points in the input space, with the geodesic distances[72]. Zhang et al utilized Isomaps in mining the gene expression data of lung cancer and pathological dataset of breast cancer data its performance was tested using residual variance. Though isomaps executed well, the projections of the non-linear axes produced is more complex and the clinical decision depends on medical experts[72].Isomaps have immensely helped in reducing dimensions of EEG signals which consists of recordings of electrical signals[73].

2.7.4 Deep Autoencoders

Auto-Encoders are unsupervised learning model comprises of single-layer neural network that transforms the input into a compressed representation by trying to minimize the reconstruction errors in the network between input and output values[74].Ahmad and Mehmet developed a deep sparse auto-encoder model to understand medical waveform datasets with differing dimensionality in Epilepsy Serious Detection, SPECTF Classification and Diagnosis of Cardiac Arrhythmias.[75]. Suk and Shen introduced a stacked auto-encoder with latent feature representation because of the existence of non-linear patterns in the relationship of features in the diagnosis of Alzheimer disease.[76]. In summary, deep auto-encoders play significant role in image analysis of medical applications.

2.8 Quality Assessment Metrics

There are variety of methods for assessing the quality and the evaluating the performance of dimensionality reduction algorithms. Two approaches have been followed in preserving the dimensions namely local and global approaches.

Though linear and non-linear dimensionality algorithms have emerged for the purpose of manifold learning but assessing the Non-linear dimensionality reduction (NLDR) techniques requires the preservation of topological properties. The reconstruction error acts as a universal measure, this property is possible with PCA and non-linear auto-encoders[77]. In order to evaluate the faithful embeddings as well as preserve the structure while reducing into low dimensional space from high dimensional space produced by LDR and NLDR techniques, a rank-based criteria was proposed.

2.8.1 co-Ranking

Lee and Verleysen introduced a metric called the co-Ranking, where the distance is calculated from every pair of data points in a low dimensional space from a high dimensional space. In the study, this criterion takes K –ary neighbourhoods are the outcomes from ranking the distance measures between every pair of data points found in high dimensional space and low dimensional space for different values of k [77].

The structure of the data is defined by the neighbours, the relationship between the neighbourhoods and their similarities[78]. The co-ranking is defined by Lee and Verleysen in Equation 2.15 and Equation 2.16,

Equation 2.15

$$\boldsymbol{q_{kl} = \left|\{(i,j): p_{ij} = k \text{ } and \text{ } r_{ij} = l\}\right|}$$

Equation 2.16

$$Q = [q_{kl}]_{1 \le k,l \le N-1}$$

where Q is a co-Ranking matrix, p_{ij} represents the rank matrix in the high dimensional space and r_{ij} indicates the rank matrix in low dimensional space from x_i and x_j. The errors are found in the non-diagonal entries of co-ranking matrix after the process of dimensionality reduction process and defines two terms in the matrix namely intrusion and extrusion. The intrusion indicates the positive rank error where the number of points enters a neighbourhood erroneously[54,60]. On the other hand, extrusion is called the negative rank error specifying the number of points exiting the neighbourhood wrongly[54,60].

co-Ranking has been useful in providing 3D representation of high dimensional feature space extracted from the data and provide the user the base of exploratory data analysis for visualization based approach in Earth observation data[78].

2.8.2 Area under R_{NX} Curve

Lee and Verleysen initiated a new criteria called R_{NX}, that constitutes the neighbourhood preservation of K-th nearest neighbours with random point distribution which depicts a value between 0 considered as random embedding and 1 as a perfect embedding. In the study, Lee introduced the metric the total area under the R_{NX} graph (AUC) to assess the average quality of DR on all scales with most appropriate weights higher the value of AUC, better the working of the DR algorithm. Lee et al tested the measure in different dimensionality reduction algorithms and classical Multidimensional Scaling(MDS), Non-metric Multidimensional Scaling(NMDS)[79], Curvilinear Component Analysis (CCA), Stochastic Neighbour Embedding (SNE), Neighbour retrieval and visualisation (NeRV), and Jensen-Shannon embedding (JSE) performed better in preserving mid- to small-size neighbourhoods[80].

2.8.3 Local Continuity Meta-Criteria (LCMC)

Chen et al proposed the Local Continuity Meta Criteria which is defined as the average size of the overlap of K-nearest neighbourhoods in the configuration of high-dimensional data and the low-dimensional data[81]. The LCMC reaches 100% when the number of nearest neighbours is high and higher the size of the expected overlap, in addition, higher the values of LCMC with less number of neighbours produce faithful embeddings in high dimensional space[71]. Rohan and Amarda utilized LCMC in five dimensionality algorithms such as PCA, Isomaps, LDFMap, KPCA and LLE for differentiating near-native from non-native structures predicted from de novo structure[71].

2.9 Summary

In summary, a complete background study and related conventional approaches for clustering ensemble is discussed. At first, an introduction of a cluster ensemble and its applications of various backgrounds they have been applied and explored. Moreover, an overview of clustering ensemble process i.e. different generation techniques is possible with clustering ensembles, in the overall studies, one of the significant ones analysed in the literature are homogeneous, heterogeneous ensemble models. Following that, the data has been subsampled or resampled and run multiple times for checking the cluster stability before feeding into the consensus functions. With respective to find an optimal k value in consensus clustering using cumulative distribution, and introduction to Proportion of ambiguous clustering and performance metrics for evaluating clustering ensembles.

Next, a detailed review of different consensus functions with their functionalities to handle enormous datasets and their implementation apparent in diverse fields. According to the studies, each consensus function has a unique role i.e. k-modes and LCA has the ability to manage humongous amount of categorical data and hypergraph based ensembles such as CSPA to utilize knowledge reuse framework for joining the consensus partition.

Feature extraction algorithms play vital role in reducing the dimensions effectively, before feeding the data to a clustering algorithm. From the studies, we summarize linear and non-linear dimensionality reduction algorithms such as PCA, Isomaps and Deep auto-encoders, where PCA execute superior in some medical applications whereas stacked deep auto-encoder perform well in the image analysis and depicting the behaviour of electric wave signals in EEG, Isomaps helps in understanding the manifolds of the data by preserving the topological structure.

Lastly, the quality assessment of feature extraction algorithms, we initiated a rank based criteria to evaluate the embedding formed in the low dimensional space from a high dimensional space. Moreover, co-Ranking aids in recognizing the data ranked and according to the study, they have been effective in visualizing 3D images when projected from high dimensional space.

CHAPTER 3 METHODOLOGY

This chapter presents our methodology followed to achieve our research objectives. Figure 3.1 illustrates our overall methodology. The methodology was divided into four phases, starting with the data preparation of the cohorts, extracting the meaningful features from the feature extraction process, assessing the best dimensions and comparing the different feature extraction algorithms, employing the reduced features to the formation of cluster ensembles, and the evaluation of the ensemble.

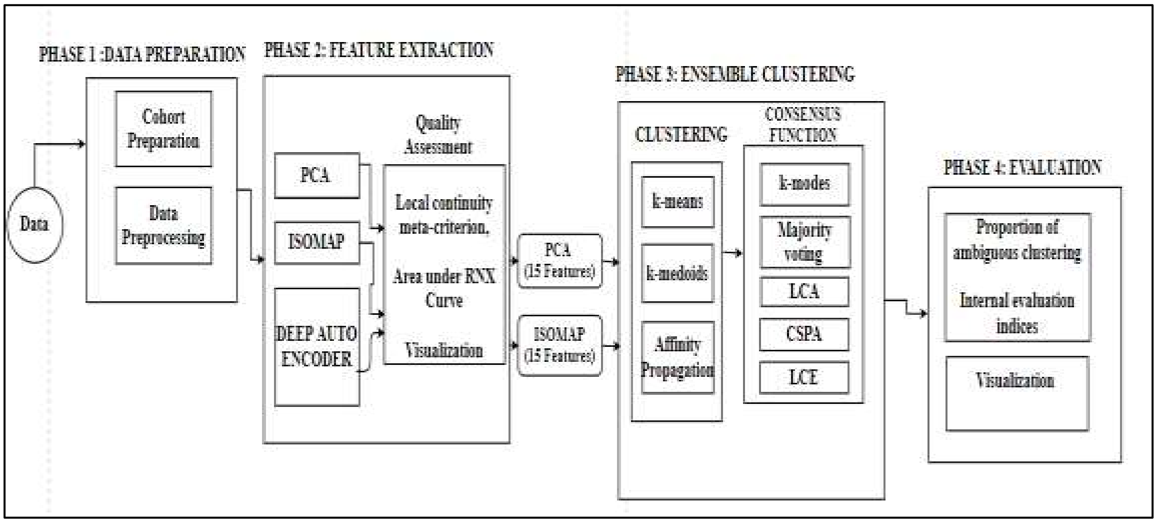

Figure 3.1. **An overview of the methodology**

PHASE 1—Data Preparation: The pathology dataset spans six years. We divided the data into three time cohorts namely 1-year from, 2-year and 6-year for observing the seasonal changes among the physicians. The pathology dataset encompasses of different age groups, laboratory tests and disease information of the patients. In this phase, we also filter the unwanted features and standardize the features in the dataset.

PHASE 2— Feature Extraction: In the feature extraction phase, we applied three feature extraction algorithms namely PCA, Isomaps and Deep Autoencoders, one to analyse the dataset with projection related to linear and non-linear dimensionality reduction with PCA acting as a baseline in linear and two non-linear dimensionality reduction algorithms such as Isomaps and Deep auto-encoders. Moreover, we assess the quality of the three algorithms using three quality measures in this phase and two preeminent techniques with the best feature set was taken to the clustering ensemble phase.

PHASE 3—Ensemble Clustering: This phase involves developing the ensemble clustering model using the features from the previous phase. We selected and used only centroid based clustering models namely k-means, k-medoids and affinity propagation as they are suited for the dataset. We applied five consensus functions to combine the three algorithms to build a robust cluster ensemble model.

PHASE 4—Cluster Evaluation: In this phase, we evaluate the ensemble clustering output, finding the optimal k- cluster value using internal evaluation indices and proportion of ambiguous clustering (PAC). We visualized the resulting cluster of physicians for all three time-cohorts using t-distributed stochastic neighbor embedding(t-SNE) as a two dimensional plot.

3.1 Cohort Preparation

The pathology dataset comprises of six years of data commencing from 2012-2017. It provides information about the laboratory test orders and patient's diagnosis information which is significant to determine the physician's case-mix. Abidi et al in their studies, took the specified test orders that are ordered to indicate the presence or absence of specific disease by considering that the physicians will consequently order more tests related to those diseases as well as the results of the test will affirm that patient suffering from the particular disease is consulted by the physician[1]. We performed a test-disease mapping between 15 pathology laboratory tests and 27 diseases over the span of 6 years. The main reasons for splitting the dataset into time cohorts was to determine the test ordering behaviour of the physicians over time, and whether this behaviour is consistent over time or subject to seasonal influences. We also were interested in examining whether the peer groups remain consistent across time or is there a pattern of change that is worth investigating further.

3.1.1 Dataset

In order to comprehend the similarities of the physician 's case mix among the peers over a period of time, the dataset was partitioned into three cohorts of 1-year, 2-year interval and 6 –year interval. This segregation of the data aids to recognize whether the test ordering behaviour of the physicians coincides with the peers across the given period of time. During this time period, the dataset comprises of 15 laboratory tests and 27 disease information.

There are about 15 laboratory tests such as CBC, PT, Electrolyte Panel, Glucose AC, Creatinine, Alkaline Phosphate, GGT, ALT, AST, Triglycerides, Cholesterol, HDL Cholesterol, TSH, Urea and Glucose Random. Table 3.1 illustrates the 27 disorders which were mapped from the test orders.

Table 3.1 Disorders

Hemolytic Anemia	Bleeding Disorder	Diabetes ketoacidosis
Hemorrhagic Anemia	Kidney Disorder	Hypoglycemia
Iron Deficiency Anemia	Liver Biliary Disorder	Hemodilution SIADH
Vitamin B12 Folate Deficiency	Addison Disease	Metabolic Bone Disease
Bone marrow Failure	Thyroid Disorder	Low Muscle Disease
Polycythemia Vera	Lung Disorder	Muscle Injury Hemolysis
Lymphoma	Dehydration	Parasitic Infection Allergy
Leukemia	Mineralocorticoid Excess Disorder	Inflammatory Conditions
Thrombocytopenia	Diabetes mellitus	Cardiovascular Disease

Initially, the number of test orders were split into four categories namely Normal, Abnormal, Unknown and ALL orders. A threshold value is set to decide if the order is a Normal test order or Abnormal test order. All the tests go to “Unknown” category if there is no threshold value to select if its normal or abnormal. Similarly, the diseases are derived from the test types where tests have a result i.e. (Normal or Abnormal) and were categorized into Normal, Abnormal and Unknown based on the test order category.

Consequently, the ALL category consists of the sum of Normal, Abnormal and Unknown order. Table 3.2 represents the other features present in the dataset apart from the test orders and disease information and provides the classification of different age groups and the gender attributes of the patient case-mix.

Table 3.2 List of features present in the pathology dataset

PMB ID	NB Order Female
NB Order	NB Order Male
NB Order 0-18	NB Order Ratio 0-18
NB Order 19-30	NB Order Ratio 19-30
NB Order 31-50	NB Order Ratio 31-50
NB Order 50-65	NB Order Ratio 50-65
NB Order 66 above	NB Order Ratio 66 above
NB Order Female Ratio	NB Order Age_min
NB Order Male Ratio	NB Order Age_max
NB Order Age_mean	NB Order Age_median

3.1.2 Cohort Separation

Table 3.3 provides the separation of time-cohorts. In the first cohort, the test orders were treated and separated as an individual year on a yearly basis. With respective to the second cohort, the data was segregated on a 2-year period merged together with 2012-2013 versus 2014-2015 versus 2016-2017 with no overlaps between them.

The main reason was to observe the behaviour of the physician’s ordering by considering the first two years together and then the next two years, which constitutes to analyse the changing effects in different time periods.

Table 3.3 Separation of time cohorts based on physician’s patient case-mix

Time-cohorts	**Year**
	2012

1st Cohort	2013
	2014
	2015
	2016
	2017
2nd Cohort	2012-2013
	2014-2015
	2016-2017
3rd Cohort	2012-2017

The third cohort represents the overall study of six years indicating the physicians occurring together for all the six years. The ensemble clustering was implemented for all the three time-cohorts i.e. on a yearly based, 2-year based and 6-year based partitions.

3.2 Data Pre-processing

3.2.1 Null column deletion

Firstly, the data was dealt with some unwanted rows and columns. Before the cohort separation, there were some universal columns to be removed indicating null values from all the tests and disease information observed under the three categories Normal, Abnormal and Unknown. Most of the columns, with the respect to the "Unknown" category with both test orders and disease information will be removed from the entire analysis as it never contributes nor declared as normal or abnormal. Likewise, the 'ALL' category will be filtered out since it's the sum of all three categories. Furthermore, the other features such as the pmb id, age min, age max, age mean will be removed, only age median will be used. The rest of the columns that were removed under Normal and Abnormal test orders from all the six years were Triglycerides, HDL Cholesterol, and Cholesterol. From the disease information, the Cardiovascular disease was refined from the both the test orders, indicating null values present in both columns. After the cohort separation, from the overall study of six years, there only 26 diseases and 12 tests orders were taken into analysis to the feature extraction process. In the remaining cohorts with respective to first and second cohort, there were some null columns still present in the test orders and disorders were filtered as it does not aid the analysis in the next phase. During the first two years of 2012 and 2013, the test order Glucose Random was removed. Following the next two years, the test ordering was carried out with 12 test orders. During the last two years, the column electrolyte panel test was discarded due to presence of null values that leads to 11 test orders in total, this test removal has concluded by taking only 25 diseases in the last two years, i.e. Mineralocorticoid Excess Disorder was also excluded because of null values present in it, in addition, age group 0-18 was only discarded, because of the presence of null values in the last two years of 2016 and 2017.

After the removal of null columns from the number of test orders in each time-cohort, the resultant test orders from (Normal and Abnormal) were left with 11 and 12 test orders in total. Table 3.4 shows the total number of laboratory test orders and disorders found in different time cohorts, the number of Normal and Abnormal test orders, Normal and Abnormal Disorders available in 1-year, 2-year and 6-year interval from 2012 to 2017.

Table 3.4 Laboratory test orders and disorders on different time-cohorts

Time-cohorts	Year	Normal Test Orders	Abnormal Test Orders	Normal Disorders	Abnormal Disorders
1st Cohort	2012	11	11	26	26
	2013	11	11	26	26
	2014	12	12	26	26
	2015	12	12	26	26
	2016	11	11	25	25
	2017	11	11	25	25
2nd Cohort	2012-2013	11	11	26	26
	2014-2015	12	12	26	26

	2016-2017	11	11	25	25
3rd Cohort	2012-2017	12	12	26	26

3.2.2 Removal by Criteria

In the first cohort, there were many physicians with less number of test orders, hence this confers that they received less number of patient case mix with disease information. We cleared the number of physicians who had placed less than or equal to five orders, which aids the next phase. Specifically, this condition occurs only in the first cohort, but not in the second or third cohort because we are considered two or six years together.

3.2.3 Variance Threshold

This technique was used to filter the features of low variance to improve the performance of the model in the next phase. In this dataset, the ratio feature of different age groups and gender information was present namely NB Order Ratio 0-18, NB Order Ratio 19-30, NB Order Ratio 31-50, NB Order Ratio 50-65, NB Order Ratio 66 above, NB Order Male Ratio, NB Order Female Ratio. All these features had a low variance score of less than 1. In order to discard these features, we use the scikit-learn package Variance Threshold library. After experimenting with different threshold values, we finalized the cut-off value to be 0.6 in the three time-cohorts**.** The features that fall under the particular threshold value were removed. As the feature selection filter method was applied on all the three cohorts, it filtered the seven attributes that holds ratio information of variance less than 1. Finally, after removing the unwanted features and rows by criteria, Table 3.5 Provides the feature set to be used in the feature extraction process. All the features found were continuous in nature. Table 3.6 illustrates the number of features present in each cohort after pre-processing the data.

Table 3.5 Main Feature Set

Features	**Description**
NB Order	Continuous
NB Order Patient Age Median	Continuous

NB Order Age group 0-18	Continuous
NB Order Age group 19-30	Continuous
NB Order Age group 31-50	Continuous
NB Order Age group 50-65	Continuous
NB Order Age group 66+	Continuous
NB Order Sex Male	Continuous
NB Order Sex Female	Continuous
NB Order Normal Test Orders	Continuous
NB Order Abnormal Test Orders	Continuous
NB Order Normal Disorders	Continuous
NB Order Abnormal Disorders	Continuous

Table 3.6 Feature set in three time-cohorts

Time-cohorts	**Year**	**Number of features**
1st Cohort	2012	83
	2013	83
	2014	85
	2015	85
	2016	80
	2017	80
2nd Cohort	2012-2013	83
	2014-2015	85
	2016-2017	80
3rd Cohort	2012-2017	85

3.2.4 Standardization

The next pre-processing step was to standardize the features with zero mean and standard deviation 1 which is a suitable scaler to be applied for dimensionality reduction algorithms. The scikit-learn Standard Scaler library was implemented, and the features were standardized with the distribution of mean value around 0 and standard deviation as 1.

3.3 Feature Extraction

Feature extraction has the potential to increment the performance of the learned models by extracting features from the input data[82]. With the huge amount of data which primarily leads to the problem called the curse of dimensionality that occurs due to the increase in demand for processing and storage requirements[83]. The general framework of this phase is to reduce the dimensionality of data by eliminating the redundant data and with minimal information loss after the transformation[82,83]For the purpose of understanding the features of the data, and to project the data from a high dimensional space, there are two ways to present them by ensuring that the original features fall under the category of linear combinations or if there exists a non-linear relationship among the variables. In this dataset, there are many relevant features that include test orders, disease information, and other features related to age and gender. Moreover, it is difficult to obtain whether the features form linear or non-linear relation. So, we experimented three feature extraction algorithms namely PCA, Isomaps and Auto-encoders, one deals with linear dimension and other two algorithms represent non-linear dimensions. John A. Lee et al. provided a different taxonomy of DR-FE techniques [62] of evaluating the reduced dimension by keeping the shape of the geometry, the local properties and neighbourhood information of the data.

3.3.1 Principal Component Analysis (PCA)

PCA-based feature extraction authorizes to reduce the dimension into limited number of components from a large number of features[84]. Interestingly, PCA-based feature extraction has been implemented in many medical applications that involves the diagnosis of valvular heart diseases[85], a medical segmentation technique using 3D Discrete Wavelet Transform [86].

We implemented PCA-based feature extraction as a baseline method and compared with the other non-linear dimensionality reduction methods to conclude the results. Table 3.7 presents the parameters used for PCA.

Table 3.7 Parameter for PCA

Parameters	Default values	Optimized values
n_components	2	10, 15, 20
whiten	false	false
svd_solver	auto	auto

We utilized the scikit-learn implementation as it provides various parameters for checking the dimensionality and explains the cumulative variance of the principal components. A scree plot provides the information about the variability of each principal component. For selecting the best number of components, we plot the cumulative variance on a graph, which gives the information of the explained variance ratio. We start with the cut-off value 95 % of the variance as it covers 10 principal components. At last, three dimensions were chosen till it reaches a maximum variance of 99% resulting in 15 components and with 99.8 %, the PCA was able to capture 20 principal components. The parameter svd_solver is set to ‘auto’ based on the shape and number of components to reduce to the lowest dimension. We used the svd_solver in the scikit-learn package that illustrates the linear dimensionality in PCA. The parameter whiten was set to false because it removes some information from the transformation.

3.3.2 Isomaps

Isomaps is one of the non-linear dimensionality reductions method that operates on by preserving the distances of the underlying data. Furthermore, the data present may contain essential multiple nonlinear relationships between features that cannot explained by linear models[63]. Antonio and Santiago described that Isomaps were proposed in order to overcome the shortcomings of spatial distances[63] and the use of Euclidean distances could not indicate their intrinsic similarity as well as its not appropriate for obtaining the embedding[87].

From a diverse set of NLDR techniques present in the literature that depend on less geometric concepts and prefer to use other distance measures. We selected Isomaps as it uses geodesic distances and extends the classical Multidimensional scaling method. Table 3.8 presents the parameters for Isomaps.

Table 3.8 Parameter for Isomaps

Parameter	Default values	Optimized values
n_components	2	10,15,20,32
neighbours	5	20,25,30,35
metric	Minkowski	Minkowski
max_iter	None	30
p	2	2

We used the scikit-learn package of Isomaps as it advances numerous choices of selecting distances. From the package, in unsupervised non-linear dimensionality reduction, the two optimal parameters that provide faithful embeddings of Isomaps depends on the number of components and number of neighbours. The optimal parameter required perform better is to choose the right number of neighbours with respect to the minimization of the reconstruction error as low as possible. Samko et al in the study, implemented by choosing the neighbours manually but this resulted in removing the small structures of data in the manifold[88].We explored with different metrics by using Minkowski and precomputed which uses distance matrix as input. The parameters for the Minkowski and precomputed metrics were evaluated with trials of Euclidean distance, Manhattan distance and arbitrary distance.

3.3.3 Deep Autoencoders

We used deep auto-encoders which works by flattening the input to a reduced dimension, and then reconstructs the output from the reduced one. We executed the keras framework for building a five-layered deep auto-encoder with two encoder layers, one bottleneck or code layer and two decoder layers. In this case, we employed an under complete auto-encoder where the bottleneck layer has the lowest dimension than the input dimension and introduced some sparsity. Table 3.9 presents the layers used in the deep auto-encoders.

We deployed different activation functions such relu, sigmoid and softplus but finalized the encoded layers and decoded layers with sigmoid and softplus. In the code layer, we initiated a tanh activation that tends to preserve the manifold and topological properties in the reduced dimension.

Table 3.9 Layer for Deep Auto-encoders

Layer Type	**Layers**	**Activation functions**
encoder layer_1	64	sigmoid
encoder layer_2	40	softplus
code_layer	32, 20,10	tanh
decoder layer_1	40	softplus
decoder layer_2	64	sigmoid

Table 3.10 illustrates the parameters used in the layers. The reconstruction error was calculated and appears to be reduce the RE when the tanh activation function was utilized. We employed the penalty term L1 regularizer and tried different threshold values, that helps retrieving a meaningful feature during feature extraction process and provides regularized outcome. The mean squared error loss was implemented while training the unsupervised input data. The adaptive moment estimation (ADAM) optimizer was utilized as the learning rate for the parameters in the neural net.

Table 3.10 Parameter for Deep Auto-encoders

Hyperparameters	**Default values**	**Optimized values**
Loss	mse	mse
regularizer	L1, L2	L1

regularization penalties	-	kernel
threshold	0.001,0.000001	1e-05,1e-04
optimizers	0.99	0.9

3.3.4 Quality Assessment of Feature Extraction Algorithms

There are a variety of quality assessment criteria to evaluate dimensionality reduction algorithms. Interestingly, there are both local and global approaches for the assessment. One of the primary technique was to minimize the reconstruction error followed by PCA and nonlinear auto-encoders[77]. In order to assess the more complex quality criteria of NLDR and LDR techniques, is to preserve the structure of data. The objective is to evaluate dimensionality reduction algorithms specifically linear dimensionality reduction namely PCA and non-linear dimensionality reduction techniques such as Isomaps and Deep Autoencoders using ranking strategy when the features are reduced from a high dimensional space.

3.3.4.1 co-Ranking

Here co-Ranking provides an assessment of embeddings of the data points in low dimensional space projection from high dimensional space. Initially, we implemented the scikit-learn package for co-Ranking that calculates the Trustworthiness, Continuity and LCMC but it lacked important details of how much neighbours were covered. We utilized the co-Ranking package in R focused on the data projected on a low dimensional space. It provides many parameters to measure LCMC, R_{NX} and Area under the R_{NX} Curve. At first, the co-ranking matrix is formed from the inputs of the reduced dimension and original dimension. Mainly focuses on K-ary neighbourhood for different values of K, where the neighbourhood values are determined from ranking the distance measures[77]. The visualization of the co-ranking matrix was implemented by using image plot function.

3.3.4.2 Local Continuity Meta Criteria (LCMC)

Local continuity meta-criteria represent the maximum overlap of the neighbours by tracing the k - nearest neighbours in high dimensional and low dimensional space[71]. We evaluated the LCMC score from the co-ranking matrix. We visualized the line plot which reflects the information about the embeddings such as Q_{local} and Q_{global} proposed by Lee and Verleysen which describes the left and right mean values of k neighbours, especially Q_{local} is observed more than Q_{global}.

3.3.4.3 Area under R_{NX} Curve

The R_{NX} (K) Curve was executed using the co-Ranking matrix to check the overall performance level of K [89]. The curve indicated the refinement of the embedding over a random embedding for the size K of the neighbourhood by Lee and Verlysen[80]. The Area under R_{NX} curve was carried out and was the best metric to check the overall performance of DR on all scales.

3.4 Cluster Ensemble

This phase involves the clustering ensemble process. We built heterogeneous clustering ensemble model that entails to use three or more different clustering algorithms resulting in a final clustering. We utilized a package in R called diceR (diverse cluster ensemble in R) to construct ensemble model. Derek and Aline[90] the authors who developed the diceR framework to understand the behaviour of clustering the patients into sub-populations that helps in detection, prevention of cancer in response to drugs and analysis of genomics data.

The framework provides extensive functions such as producing diverse clustering, ensemble partition from the consensus functions, and selection of algorithms that is best suitable for the data. Firstly, it presents diverse clustering algorithms where each clustering algorithm has its unique function namely k-means, hierarchical clustering, divisive analysis clustering, k-medoids, affinity propagation, self-organizing map, spectral clustering, Gaussian mixture model, bi-clustering, fuzzy c-means clustering, hdbscan and nonnegative matrix factorization. Interestingly, there are a variety of consensus functions offered in the package such as k-modes, LCA, LCE, majority voting and CSPA. Furthermore, the final clustering is evaluated by using the internal evaluation indices and visualization of the evaluation is also processed.

3.4.1 Clustering Algorithm Selection

After the execution of the methodology, we selected k-means, k-medoids and affinity propagation for the analysis of clustering ensemble process. The selection was based on validation of internal evaluation indices i.e., by taking metrics of compactness, how well the clusters are separated and cluster stability for multiple cluster runs, visualization of the clusters produced by the algorithms to produce homogeneous clusters and existence of consistent clusters. Primarily, for some algorithms the stability of the cluster becomes insignificant and leads to the formation of univocally defined clusters and cluster boundaries[30]. We tested various clustering algorithms agglomerative clustering, divisive clustering which formed univocal clusters indicating the existence of single cluster. Following that, we experimented on gaussian mixture model, spectral clustering, fuzzy clustering was not selected for this dataset based on the low validation scores obtained from the internal evaluation indices.

3.4.2 Centroid models

Centroid based algorithm constitutes a set of objects [91], that are assigned to the clusters, based on the distance between the cluster and central vector is minimized as possible as one of the primary objectives and are suitable to handle spherical based clusters. k-means computes the mean of the objects and choose its initial cluster centres, whereas k-medoids uses the medoids and assigns the objects to the nearest medoid and employs the dissimilarity measures such as Euclidean distance, whereas Manhattan distance provides robust solutions, because its uses the sum of the absolute distances and the most appropriate measure to handle outliers if its resides in the data. With affinity propagation acts similar to that of k-medoids, where the initialization is not required, and takes the advantage of exemplars to find the similarity between data points. At first, the data points are treated as possible exemplars, the message-passing process occurs with all the data points exchanging information, the process is continued until a good set of clusters with the best exemplars have reached a consensus[92]. We selected the three centroid models for the ensemble process.

3.4.3 Cluster Ensemble Generation

We constructed the clustering ensemble process in five steps. Interestingly, the five steps involve cluster generation from the three algorithms, imputation of the missing NA values using k-nearest neighbours, joining the partition using the five consensus functions and evaluation of the final clustering. We implemented the pipeline function of dice function in diceR, that executes the process sequentially of the three functions consensus_cluster (), impute_knn () and consensus_evaluate (). Table 3.11 demonstrates the parameter description used in the ensemble process of diceR.

Table 3.11 Parameter description of diceR

Parameters	**Description**	**Values**
data	data matrix	data matrix
nk	number of clusters	k =2 to 7
reps	number of subsamples	10,15,20,25
prep.data	performed on a raw data or bootstrap samples.	raw
nmf. method	non-negative matrix factorization with lee and brunet methods.	lee
distance	distance measures such as Euclidean, Manhattan, spearman, minkowski etc.	Euclidean and Manhattan
algorithms	number of algorithms	k-means, k-medoids and affinity propagation
cons functions	consensus functions – majority voting, k-modes, CSPA, LCA and LCE.	majority voting, k-modes, CSPA, LCA and LCE
sim.mat	similarity matrix for LCE	asrs, cts, srs
seed	imputation seed for k-nn.	1
n	ranks the algorithm based on performance.	n=1
evaluate	Internal evaluation indices	TRUE
trim	Trimming the poor algorithms.	TRUE
reweigh	Re-assigning the weights after trimming	TRUE
plot	visualization of CDF graphs	TRUE

3.4.4 Cluster Generation Mechanism

The cluster generation process involves the generation of the clustering obtained from the three clustering algorithms such as k-means, k-medoids and affinity propagation. We initialized the best 15 features of PCA and Isomaps was given as a data matrix, the cluster size was carried out from k = 2 to k = 7, the number of subsamples, the clustering algorithms, two distance measures namely Euclidean and Manhattan distance was applied for k-medoids, the non-negative matrix factorization was also initiated. During this process, Monti et al subsampling technique with multiple cluster runs was used, in which each algorithm is executed by engaging various subsets of data and about 80 percent of the actual observation was taken into account for the analysis[30][90]. We experimented with reps parameter that indicates the subsamples, and the cluster runs for the algorithms applied. Since several subsets of data was utilized, there will be some missing values due to subsampling which are rectified in the next step. The prep.data was checked between raw and bootstrap samples.

3.4.5 Imputation

The missing values occurs in the data matrix, since 80 percent of the data is taken from every clustering algorithm with cluster runs maintained at reps =20, with random subset chosen at each cluster run with respect to the subsampling method[90]. In order to remove the missing NA entries, the imputation of the k-nearest neighbour was set to the random seed. After imputation, the resultant matrix is a clustering matrix, with resample data obtained from the number of columns equivalent to the clustering [90].

3.4.6 Implementation of Consensus Functions

We experimented all the five consensus functions to examine which consensus function is best suitable for this data. Table 3.12 illustrates the parameters for the consensus functions. Each consensus function has a unique methodology of joining the ensemble partition. We implemented k-modes, majority voting, LCA, CSPA and LCE by taking the matrix of clustering obtained from imputation. For LCE, the similarity matrix was checked and finalized with approximated similarity rank matrix. The ASRS matrix was the improved version of sim-based similarity matrix[29] and the decay constant was also set.

Table 3.12 Parameter for consensus functions

Parameters	Default values	Optimized values
reps	10	15, 20, 25
prep.data	raw	raw
distance	Euclidean distance	Euclidean and Manhattan
sim.mat	cts	asrs

3.5 Internal Evaluation Indices

We evaluated the final clustering with the internal evaluation indices. The internal evaluation indices examine the cluster labels itself without any reference labels. The diceR package provided 16 internal evaluation indices from the imported packages such as clValid, clusterCrit and LinkClue. The five internal evaluation indices were chosen based on the importance measures of compactness, connectivity, and separation [59] of the clusters formed. We used five internal validation indices namely Calinski Harabasz, Compactness, Dunn index, Silhouette index, Connectivity into the analysis. The Proportion of ambiguous clustering was acquired for finding the optimal value of K for the three clustering algorithms. The validation indices were assessed for both individual clustering algorithms obtained after multiple cluster runs and the consensus functions.

3.6 Visualization

We visualized the clusters on a two-dimensional space using t-distributed stochastic neighbor embedding(t-SNE). The t-SNE is a statistical technique developed to visualize the data points projected on a high dimensional space[93]. Moreover, t-SNE was able to determine the well separated clusters over different values of K. For obtaining the optimal value of K, the cluster range values were plotted in the cumulative distribution function graphs with the consensus index values for the three clustering algorithms. For cluster stability of the clusters for cluster runs, the consensus matrix is displayed in a heat map which depicts how well the clusters are connected and the number of clusters to be found and significant.

CHAPTER 4 RESULTS AND DISCUSSION

The study was conducted on a total of 997 physicians who appear in the overall study of six years from 2012 to 2017. We developed heterogeneous cluster ensemble approaches for the 1-year, 2-year and 6-year intervals. We evaluated the consensus individual clustering and ensemble clustering in the different cohorts (as mentioned in the Chapter 3) with the internal evaluation indices. Based on the evaluation, we compared the two techniques to conclude and analyse which method appears to be a suitable technique for the three different cohorts. Moreover, the comparative analysis aids to understand the changing effects and significance of clustering towards individual clustering and ensemble clustering. Though the analysis was carried out for various cluster ranges starting from k = 2 to k = 7 in the three time cohorts, the comparison was made between the two techniques by identifying the optimal k-value in consensus clustering. In a centroid based models, the optimal k-value is an important criterion to determine the best number of homogeneous clusters with the extracted features. The cluster ranges were checked to recognize how well the clusters are separated, but only the best k-value was included in the analysis. Secondly, we differentiated the homogenous clusters obtained from two feature extraction methods such as PCA and Isomaps were assessed, to check whether the linear or non-linear dimensionality reduction algorithm works the best for this pathology dataset.

Firstly, the analysis was carried out on all physicians occurring from 2012 to 2017 with their patient case mix's tests and disease information. From the beginning, there were no missing values in the dataset, but the existence of numerous unwanted columns was removed because of null values and low frequency threshold of some features, in addition, with respect to first cohort, physicians with few test orders were filtered. So, the complete cases with the number of physicians having the patient case mix from 1-year interval, 2-year interval and 6-year interval varies. The total number of features taken into the analysis were about 85 features in the overall study of six years. With the respect to the first cohort and second cohort, there was a count of 83 features appearing in the first two years, with 85 features occurring in the next two years. Finally, there were less number of features of

80 in the last two years. Most of the features present in the entire analysis were only continuous features.

As for the feature extraction models, we utilized one linear dimensionality reduction algorithm (i.e., PCA) and two non-linear dimensionality reduction algorithms one which concerns with geodesic distance and topology preservation (i.e., Isomaps) and the latter that deals with neural networks to extract features in the three cohorts. Mainly, PCA was set as a baseline method among the feature extraction methods. The unsupervised feature extraction algorithms were assessed and compared based on the rankings of the embedding produced. During this validation, there is a quite variation and difference of the resultant embeddings produced between the linear and non-linear dimensionality reduction algorithms. With the assessment of quality, two feature extraction methods were chosen for carrying out the clustering tasks.

As mentioned in the Clustering ensembles, we selected only the centroid based partitioning algorithms for the ensemble formation by examining and eliminating the other models based on the evaluation and visualization of the consensus matrix in the consensus clustering. Interestingly, k-means performed well when compared with the other two algorithms namely k-medoids and affinity propagation for multiple cluster runs. The performance of the five cluster ensembles, majority that that works on relabeling and voting method achieves better clustering based on the evaluation of the internal validity indices for all the cluster ranges from k = 2 to k = 7. On the other hand, k-modes and LCA executed equally well in comparison with LCE and CSPA. Lastly, the visualization of the t-SNE enables to visualize the highly separated clusters more effectively.

The first phase of the analysis, all the computations were carried out using Python, especially the feature extraction methods utilizing the scikit-learn library[43]. The entire phase of clustering analysis is conducted using R packages, specifically the clustering ensemble computation is carried out using diceR package. We ran the simulations on an Intel Core i7 -4770 CPU 3.4GHz PC, equipped with 12.00 GB of RAM Windows 10 64-bit machine.

In the next sections 4.1, we will discuss about the feature extraction results from the experiments in the three time cohorts then in section 4.4 we will discuss the clustering ensemble and individual clustering results of the two feature extraction methods in the three-time cohorts and finally in section 4.5, we will present the visualization of the clusters.

4.1 Results of Feature Extraction Models

The three feature extraction models namely PCA, Isomaps and Deep Autoencoders were assessed based on the embeddings produced on a low dimensional space when projected from a high dimensional space. The quality of the embedding is evaluated by applying co-Ranking matrix, Local Continuity Meta Criterion, R_{NX} and Area under R_{NX} curve metrics.

The co-Ranking matrix indicates whether the data points are intact and assessing the loss of quality of the reduced data from a high dimensional space[77]. The image function plot of co-Ranking matrix shows of how the projection is reflected in the values of the upper triangle and the lower triangle of the diagonal matrix representing the rank errors[77].

The Local Continuity Meta Criterion select the best dimension suitable for the three feature extraction models. The evaluation depends on the number of K neighbours. Moreover, higher LCMC scores with less number of neighbours indicates perfect embedding in a high dimensional data[71]. A perfect embedding of the data produces a score 1 whereas the value 0.5 indicate random embedding. In LCMC, we defined two criteria namely Q_{local} and Q_{global}, indicating a higher Q_{local} maintains the local neighbourhood when compared to others and considered superior method than Q_{global}[94]. The number of K neighbours could be extended as much as possible. The overall performance of the Area under the R_{NX} curve reaches 1 that determines the best quality of the embedding produced.

4.1.1 Baseline Results

PCA was set as a baseline because of its classical approach used in the literature as to other non-linear dimensionality reduction algorithms. Initially, we examined Isomaps and Deep Autoencoders with quality assessment metrics but the performance of baseline algorithm namely PCA was superior and consistent at all the three dimensions namely 10, 15 and 20 and produced a better quality in a low dimensional space.

We implemented Principal Component Analysis (PCA) using the scikit-learn package. The quality was assessed using co-Ranking R package. Table 12 illustrates the LCMC criteria with Q_{local} score obtained by the line plot. The Qlocal score is considered in all the three dimensions. The three dimensions were selected based on the cumulative variance explained ratio which reaches 99 percent at the 20 dimension and further dimensions were not considered as it will result in higher complexity. Table 4.1 provides the Area under R_{NX} curve score for the three different dimensions for the three-time cohorts. The PCA produces a perfect embedding of LCMC score of 0.8-0.9 in all three dimensions in the three different time-cohorts when the number of neighbours by default was set to the 1000 neighbours in the line plot. The AUC_ln_K[95] performs well in PCA and yields better scores at two dimensions namely 15 and 20 respectively.

Table 4.1 Results of LCMC score of PCA

Time-cohorts	**LCMC score**		
	Dimensions		
	10	15	20
1st Cohort	0.8	0.9	0.9
2nd Cohort	0.8	0.9	0.9
3rd Cohort	0.8	0.86	0.9

Table 4.2 Results of AUC_ln_K score of PCA

	AUC_ln_K score			
Time-cohorts	**Year**	**Dimensions**		
		10	**15**	**20**
1st Cohort	2012	0.81	0.90	0.94
	2013	0.81	0.90	0.94
	2014	0.80	0.93	0.88

	2015	0.76	0.86	0.93
	2016	0.81	0.90	0.95
	2017	0.81	0.90	0.95
2nd Cohort	2012-2013	0.82	0.91	0.95
	2014-2015	0.81	0.89	0.94
	2016-2017	0.83	0.95	0.90
3rd Cohort	2012-2017	0.83	0.91	0.95

4.1.2 Isomaps

We executed Isomaps in two different metrics namely pairwise and minkowski, Table 4.3 demonstrates the minkowski metric was appropriate with Euclidean distance in the three time cohorts, as resulted in better LCMC score and reduced reconstruction error when compared with the other distances. The pairwise metric was examined with three distances namely euclidean, manhattan and arbitrary distances but was not included in the analysis because of poor embedding produced. Moreover, we checked for four dimensions, but the best dimension was evident at dimension 15. Interestingly, the outcome of the embeddings was slightly above the random embedding threshold of 0.5. Mainly, the AUC_ln_K and LCMC scores in the three cohorts were around 0.6. Table 4.4 describes the outcomes of AUC_ln_K in Isomaps.

Table 4.3 Results of LCMC score of Isomaps

Time-cohorts	**LCMC score**			
	Dimensions			
	10	15	20	32
1st Cohort	0.65	0.65	0.65	0.65
2nd Cohort	0.65	0.65	0.65	0.65
3rd Cohort	0.60	0.60	0.60	0.60

Table 4.4 Results of AUC_ln_K score of Isomaps

		AUC_ln_K score			
		Dimensions			
Time-cohorts	Year	10	15	20	32
1st Cohort	2012	0.65	0.66	0.66	0.65
	2013	0.64	0.65	0.65	0.65
	2014	0.62	0.63	0.64	0.64
	2015	0.61	0.62	0.62	0.62
	2016	0.64	0.65	0.65	0.64
	2017	0.65	0.66	0.66	0.65
2nd Cohort	2012-2013	0.65	0.669	0.667	0.668
	2014-2015	0.65	0.66	0.668	0.660
	2016-2017	0.66	0.669	0.67	0.660
3rd Cohort	2012-2017	0.683	0.694	0.699	0.696

4.1.3 Deep Autoencoders

We implemented Deep Auto-encoders using keras framework. The reconstruction error was trained and minimized to the lowest value of 0.45. Different activation functions were tried in the bottle neck layer, but tanh activation function and the kernel regularizer L1 used were able to achieve the lowest reconstruction error as possible. The encoded dimension was evaluated for three dimensions 10, 20 and 32 respectively. The unlabeled data was trained at different epochs and the batch size was maintained to 64. With quality assessment of deep auto-encoders, the resultant embeddings when projected from a high dimensional space into a low dimensional space were random at 0.5 to 0.6. From Table 4.5 demonstrates the random embeddings as the Q_{local} score of LCMC lies between 0.5 and 0.6. Similarly, Table 4.6 indicates the AUC_ln_K scores of deep auto-encoders. The random embeddings were less than 0.5 in the first cohort indicating the poor performance of deep auto-encoders for this dataset.

Table 4.5 Results of LCMC score of DAE

Time-cohorts	**LCMC score**		
	Dimensions		
	10	20	32
1st Cohort	0.5	0.6	0.6
2nd Cohort	0.6	0.6	0.6
3rd Cohort	0.5	0.5	0.6

Table 4.6 Results of AUC_ln_K score of DAE

	AUC_ln_K score			
Time-cohorts	**Year**	**Dimensions**		
		10	**20**	**32**
1st Cohort	2012	0.52	0.57	0.56
	2013	0.52	0.57	0.56
	2014	0.45	0.36	0.55
	2015	0.50	0.46	0.51
	2016	0.494	0.55	0.58
	2017	0.5	0.5	0.54
2nd Cohort	2012-2013	0.54	0.56	0.61
	2014-2015	0.52	0.57	0.61
	2016-2017	0.586	0.592	0.60
3rd Cohort	2012-2017	0.528	0.538	0.57

4.2 Analysis of Feature Extraction models

From the results of the three feature extraction models, PCA performs the best based on the metrics of AUC_ln_K and LCMC metrics. The PCA presents a near perfect embeddings of the data in the three time cohorts with respect to the two metrics as illustrated in the Table 4.7 and Table 4.8.

Isomaps exhibit embeddings that are slightly random—i.e. above the threshold of 0.5, but it performed better than Deep Auto-encoders. The embeddings and the rankings of the neighbourhood values produced by Deep auto-encoders were random and poor at two dimensions 10 and 20, so we selected only PCA and Isomaps for the analysis for clustering.

For the next phase, we utilized the best dimension 15 from PCA, by considering the 99% of variance and for Isomaps, the best dimension occurs at 15 with the help of the quality assessment metrics. We considered the reduced dimension 15 by feature extraction process as it does not increase the complexity of the two models.

Table 4.7 Results of AUC_ln_K score for the three time-cohorts

Feature Extraction Models	**Dimensions**		
	10	15	20
3rd Cohort			
PCA	0.83	0.91	0.95
ISOMAP	0.68	0.69	0.69
DAE	0.52	-	0.53
2nd Cohort 2012-2013			
PCA	0.82	0.91	0.95
ISOMAP	0.65	0.66	0.66
DAE	0.54	-	0.56
2nd Cohort 2014-2015			
PCA	0.81	0.89	0.94
ISOMAP	0.65	0.66	0.66

DAE	0.52	-	0.57
2nd Cohort 2016-2017			
PCA	0.83	0.95	0.90
ISOMAP	0.66	0.67	0.66
DAE	0.58	-	0.59
1st Cohort 2012			
PCA	0.81	0.90	0.94
ISOMAP	0.65	0.66	0.66
DAE	0.52	-	0.57
1st Cohort 2013			
PCA	0.81	0.90	0.94
ISOMAP	0.64	0.65	0.65
DAE	0.52	-	0.57
1st Cohort 2014			
PCA	0.80	0.93	0.88
ISOMAP	0.62	0.63	0.64
DAE	0.45	-	0.36
1st Cohort 2015			
PCA	0.76	0.86	0.93
ISOMAP	0.61	0.62	0.62
DAE	0.50	-	0.46
1st Cohort 2016			
PCA	0.81	0.90	0.95
ISOMAP	0.64	0.65	0.65
DAE	0.49	-	0.55
1st Cohort 2017			
PCA	0.81	0.90	0.95
ISOMAP	0.65	0.66	0.66

DAE	0.50	-	0.50

Table 4.8 Results of LCMC score for the three time-cohorts

Time-cohorts	FE MODELS	Dimensions		
		10	**15**	**20**
1st Cohort	PCA	0.8	0.9	0.9
	Isomaps	0.65	0.65	0.65
	DAE	0.5	-	0.6
2nd Cohort	PCA	0.8	0.9	0.9
	Isomaps	0.65	0.65	0.65
	DAE	0.6	-	0.6
3rd Cohort	PCA	0.8	0.86	0.9
	Isomaps	0.60	0.60	0.60
	DAE	0.5	-	0.5

The LCMC score was used for selection of the dimension, where the best dimension was observed at 15 for the feature extraction process (illustrated in Table 4.8). We examined the Q_{local} score for the three feature extraction models in the three time-cohorts. PCA performed better at all the three dimensions compared to the other models whereas the latter Deep Autoencoders produced random embeddings at the two dimensions.

4.3 Visualization of Quality Assessment Metrics

We implemented the line plots to visualize the dimensions of the reduced features and to observe the difference of projection between a linear and non-linear dimensionality reduction algorithm when reduced from a high dimensional space. The visualization of the metrics namely co-Ranking, Local Continuity Meta Criterion and R_{NX} curve provide the visual examination of the data in a low dimensional space and the quality of the embeddings produced by PCA and Isomaps.

4.3.1 co-Ranking

Figure 4.1 depicts the imageplot function of co-Ranking matrix of the third cohort under dimension 15. The co-Ranking matrix of PCA indicating the reduced dimension illustrates a straight diagonal, as the points of the data remain intact and uniform throughout the dataset. On the other hand, the Isomaps compresses the points and diminishes in the middle and reflects in more values on the lower right part of the co-Ranking matrix. The co-Ranking matrix are log scaled for better visualization purposes[95].

Moreover, the embeddings in the PCA preserves the smaller distances, where the points close to the diagonal are higher in the upper part whereas the Isomaps detains to keep larger distances where more values are populated in the upper and the middle part of the diagonal.

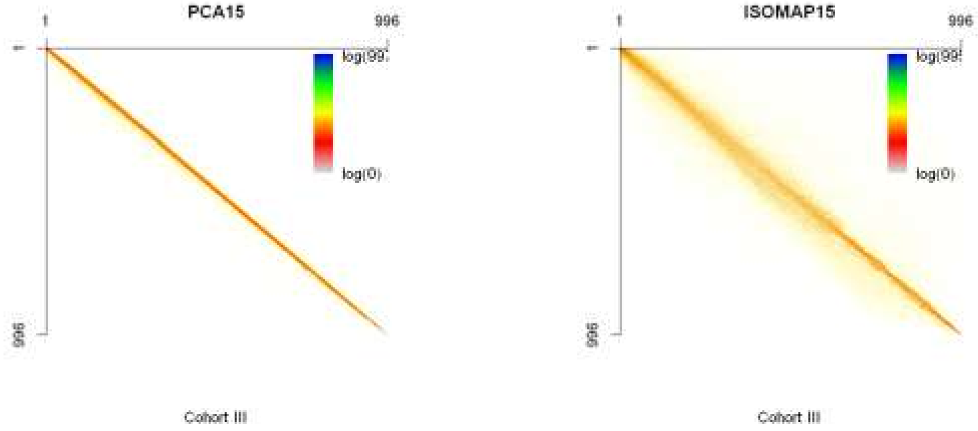

Figure 4.1 Plot of co-Ranking of PCA and Isomaps for dimension 15 of Cohort III

4.3.2 Local Continuity Meta Criterion (LCMC)

We executed the line plots for PCA and Isomaps to examine the difference between the three different dimensions namely 10,15, 20 in linear and non-linear dimensionality reduction algorithms. Figure 4.2 depicts the line plot of LCMC criteria for different dimensions of PCA and Isomaps in the third cohort. By checking the different dimensions of PCA, the Q_{local} criteria of LCMC increases steadily for the different dimensions and reaches the maximum of 0.95 at the dimension 20 with number of neighbours set to 1000.

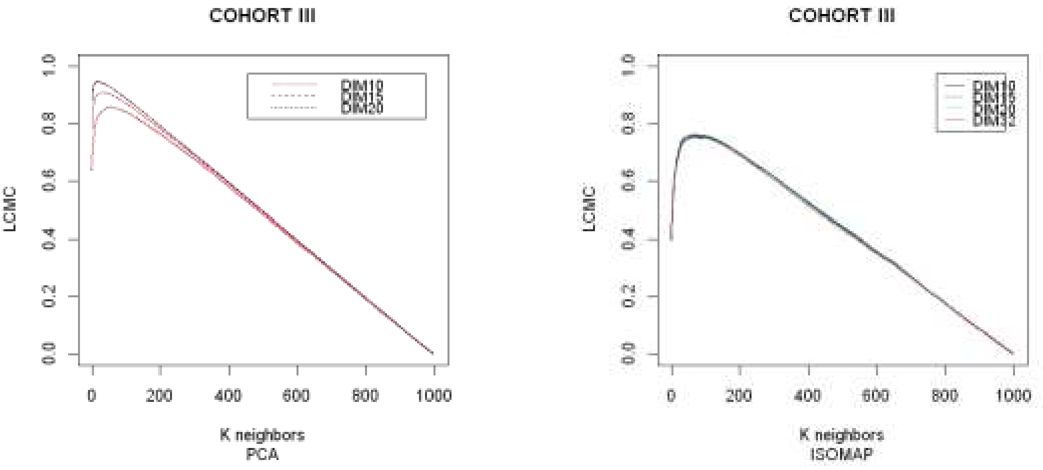

Figure 4.2 Line plot of LCMC of PCA and Isomaps for different dimensions in Cohort III

The Isomaps (in Figure 4.2) remains equivalent for all the four dimensions with Q_{local} score at 0.66 for 1000 neighbours indicating that the overlaps of the neighbours in the neighbourhood values were minimum.

Moreover, PCA achieves better overall performance in all the three dimensions for the K neighbours when reduced to a low dimensional space, the LCMC score increases as the dimension increases. On the contrary, in Isomaps the LCMC score remains the same in all three dimensions.

4.3.3 R_{NX} Curve

The R_{NX} curve reflects the neighbourhood values and ranks along the diagonal[95]. Figure 4.3 illustrates the R_{NX} curve of PCA under dimension 15 where the points of the curve start from 0.80 and reaches to 1, thus indicating better neighbour values on a low dimensional space. On the other hand, the curve begins from 0.4 by increasing steadily and extends to 1. The difference between the two R_{NX} curves is that there are slight bumps in the projection of data of Isomaps, whereas in PCA, the curve is flat and remains stable.

Our experiments conclude that the data is stable with the projection of PCA. The R_{NX} curve of a perfect embeddings is 1 and the randomized curve indicates 0. The data in both curves does not fall steeply down, but the steepness of the data is observed in Deep auto-encoders, where the projection of data starts from the lowest point and falls down after a point indicating random embeddings formed in the reduce dimension.

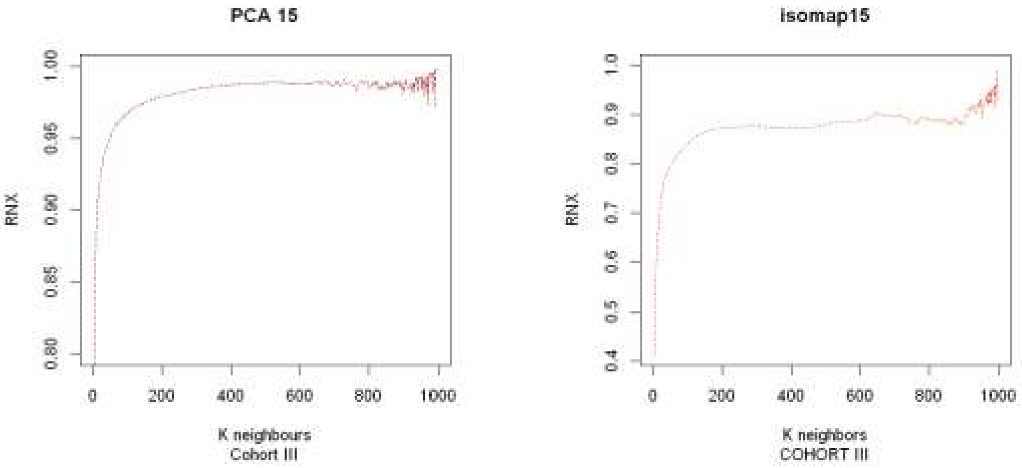

Figure 4.3 Line plot of R_{NX} Curve of PCA and Isomaps for dimension 15 of Cohort III

The outcome of the feature extraction phase, as per the above experiments with the three feature extractions models, we conclude the following: For the three dimensions (shown in Table 4.7 and Table 4.8) as evaluated by the two metrics of Area under the R_{NX} curve and Local continuity meta-criteria, and the line plots of co-Ranking matrix, LCMC and R_{NX} curve, the features from PCA and Isomaps are to be used for the clustering ensemble. It may be further noted that, PCA performs better and produces a perfect embedding than the non-linear dimensionality reduction algorithms.

From the quality assessment metrics used co-Ranking, Local Continuity Meta Criterion, and Area under R_{NX} curve indicate that linear dimensionality reduction algorithm PCA is the most effective and suitable algorithm in the pathology test ordering data because of the linear dependencies found in the variables when the data is reduced from a high dimensional space. The features extracted by PCA did not suffer the loss of quality when experimented at various dimensions. The visualization of the three metrics provided a clear coherence and interpretability of the reduced dimensions.

4.4 Results of Clustering

We applied the k-means, k-medoids and affinity propagation clustering algorithms to generate the clustering ensemble using the diceR package. We experimented with cluster ranges from k = 2 to k = 7 and determined the optimal k-value by applying the proportion of ambiguous clustering which provides the best k-value for multiple cluster runs. We selected five evaluation indices for both the cluster ensemble and the individual clustering algorithms—the metrics take into account the structure and behaviour of clustering algorithms—i.e., k-modes, CSPA, majority, LCA and LCE. Our evaluation intent is also to ascertain whether we can get better clusters using a cluster ensemble as opposed to the use of single clustering algorithms.

4.4.1 Results of Clustering Ensemble and Individual Clustering

Table 4.9 and Table 4.10 demonstrates the individual clustering results and clustering ensemble results obtained by using 15- dimensional space of PCA with optimal values of k obtained in the three time-cohorts. From Table 4.9 and Table 4.10 provides the details of PAC and the internal evaluation indices to check the Compactness, Connectivity, Calinski Harabasz, Dunn index, Silhouette coefficient denoted by the notations CP, C, CH, D and S [29,66,96].The outcome for the best performing clustering ensembles was presented among the five ensembles used. Similarly, the best individual clustering algorithm was taken among the three clustering algorithms. The cluster k-values were chosen from the range of k = 2 to k = 7 respectively.

Table 4.9 Results of internal evaluation indices of PCA

Individual Clustering					**Clustering Ensemble**			
Algorithms	**k**	**CH**	**D**	**S**	**Cons.function**	**CH**	**D**	**S**
3rd Cohort 2012-2017								
k-means	2	1070	0.020	0.48	Majority vote	1070	0.025	0.49
k-means	3	859	0.018	0.342	Majority vote	859	0.018	0.342
2nd Cohort 2012-2013								
k-means	2	1518	0.021	0.495	Majority vote	1518	0.021	0.50
k-medoid	4	977	0.008	0.25	LCA	969	0.009	0.25
2nd Cohort 2014-2015								
k-means	2	1695	0.024	0.504	Majority vote	1697	0.025	0.51

k-means	3	1372	0.019	0.352	Majority vote	1372.1	0.017	0.353
2nd Cohort 2016-2017								
k-medoid	2	1654	0.013	0.46	Majority vote	1781	0.029	0.496
k-means	3	1390	0.018	0.341	Majority vote	1390	0.0181	0.341
1st Cohort 2012								
k-means	2	1462	0.016	0.484	Majority vote	1462	0.016	0.484
k-means	3	1221	0.013	0.359	k-modes	1220	0.0133	0.361
1st Cohort 2013								
k-means	2	1576	0.015	0.492	Majority vote	1576	0.015	0.492
k-means	3	1274	0.015	0.354	k-modes	1273	0.015	0.354
1st Cohort 2014								
k-means	2	1726	0.028	0.506	LCE	1726	0.028	0.506
k-medoid	3	1056	0.008	0.227	LCE	1057	0.009	0.228
1st Cohort 2015								
k-means	2	1645	0.027	0.493	Majority vote	1645	0.028	0.50
k-medoid	3	1134	0.012	0.306	Majority vote	1336	0.026	0.342
1st Cohort 2016								
k-means	2	1754	0.027	0.495	Majority vote	1754	0.027	0.495
k-medoid	4	1047	0.008	0.227	Majority vote	1047	0.009	0.228
1st Cohort 2017								
k-means	2	1846	0.033	0.498	Majority vote	1847	0.033	0.497
ap	3	1305	0.017	0.30	Majority vote	1305	0.018	0.31

Table 4.9 illustrates the three internal evaluation indices such as Calinski Harabasz, Dunn index and Silhouette coefficient, as all three indices possess both the properties of

compactness and separation together. So, these three indices were used as comparison for individual clustering and clustering ensemble.

Table 4.10 represents the indices the compactness and connectivity indicate the intra-cluster homogeneity and connectedness among the clusters. These two indices were compared on the two approaches. The proportion of ambiguous clustering (PAC) is shown for the individual clustering algorithms such as k-means, k-medoids and affinity propagation.

Table 4.10 Results of internal evaluation indices of PCA

Individual Clustering					**Clustering ensemble**		
Algorithms	**k**	**PAC**	**CP**	**C**	**Cons.function**	**CP**	**C**
3rd Cohort 2012-2017							
k-means	2	0.03	6.04	57	Majority vote	6.05	53
k-means	3	0.04	5.22	101	Majority vote	5.23	104
2nd Cohort 2012-2013							
k-means	2	0.02	5.57	57	Majority vote	5.57	57
k-medoid	4	0.09	4.27	175	LCA	4.26	167
2nd Cohort 2014-2015							
k-means	2	0.03	5.58	68.8	Majority vote	5.60	62
k-means	3	0.02	4.80	112	Majority vote	4.80	112
2nd Cohort 2016-2017							
k-medoid	2	0.007	5.23	71	Majority vote	5.44	76
k-means	3	0.03	4.64	140	Majority vote	4.64	140
1st Cohort 2012							
k-means	2	0.03	5.617	75	Majority vote	5.617	75

k-means	3	0.08	4.87	94	k-modes	4.88	93
1st Cohort 2013							
k-means	2	0.04	5.51	72	Majority vote	5.52	73
k-means	3	0.09	4.87	109	k-modes	4.87	101
1st Cohort 2014							
k-means	2	0.03	5.60	68	LCE	5.60	68
k-medoid	3	0.11	4.30	203	LCE	4.30	201
1st Cohort 2015							
k-means	2	0.01	5.61	84	Majority vote	5.63	75
k-medoid	3	0.11	4.72	134	Majority vote	4.83	110
1st Cohort 2016							
k-means	2	0.02	5.41	77	Majority vote	5.40	71
k-medoid	4	0.19	4.2	210	Majority vote	4.2	203
1st Cohort 2017							
k-means	2	0.03	5.44	63	Majority vote	5.43	65
ap	3	0.13	4.59	170	Majority vote	4.59	170

Table 4.11 and Table 4.12 illustrates the individual clustering results and clustering ensemble results produced by using 15- dimensional space of Isomaps with optimal values of k produced by the three cohorts.

Table 4.11 Results of internal evaluation indices of Isomaps

Individual Clustering					**Clustering ensemble**			
Algorithms	**k**	**CH**	**D**	**S**	**Cons.function**	**CH**	**D**	**S**
3rd Cohort 2012-2017								

k-means	2	969	0.021	0.474	LCE	969	0.22	0.474
k-means	3	735	0.018	0.33	Majority vote	735	0.0184	0.33
2nd Cohort 2012-2013								
k-means	2	1307	0.023	0.469	Majority vote	1307	0.0234	0.47
k-medoid	4	754	0.010	0.223	Majority vote	757	0.010	0.224
2nd Cohort 2014-2015								
k-means	2	1485	0.0223	0.48	Majority vote	1485	0.0223	0.49
k-medoid	3	1015	0.011	0.30	LCA	1121	0.0133	0.34
2nd Cohort 2016-2017								
k-means	2	1485	0.015	0.475	LCE	1476	0.020	0.476
ap	3	1017	0.0137	0.303	LCA	1017	0.0137	0.304
1st Cohort 2012								
k-means	2	1251	0.016	0.462	Majority vote	1251	0.016	0.462
k-means	3	1009	0.016	0.351	LCA	1009	0.016	0.351
1st Cohort 2013								
k-means	2	1351	0.016	0.462	Majority vote	1351	0.021	0.463
k-medoids	4	680	0.0013	0.23	LCE	704	0.0009	0.25
1st Cohort 2014								
k-means	2	1513	0.020	0.484	Majority vote	1513	0.020	0.484
k-medoid	3	1142	0.011	0.331	LCA	1170	0.015	0.340
1st Cohort 2015								
k-means	2	1413	0.018	0.48	LCE	1414	0.029	0.48
k-means	7	351	0.0009	0.127	Majority vote	570	0.001	0.16
1st Cohort 2016								

k-means	2	1476	0.025	0.480	k-modes	1476.8	0.0253	0.480
k-medoid	7	380	0.0008	0.101	Majority vote	613	0.0013	0.15
1st Cohort 2017								
k-means	2	1494	0.028	0.474	Majority vote	1494	0.025	0.470
k-medoids	5	659	0.0011	0.220	Majority vote	667	0.0015	0.22

Table 4.12 Results of internal evaluation indices of Isomaps

Individual Clustering					**Clustering ensemble**		
Algorithms	**k**	**PAC**	**CP**	**C**	**Cons.function**	**CP**	**C**
3rd Cohort 2012-2017							
k-means	2	0.01	8.01	43.4	LCE	8.01	43
k-means	3	0.03	7.11	91	Majority vote	7.11	89
2nd Cohort 2012-2013							
k-means	2	0.01	7.72	59	Majority vote	7.71	59
k-medoid	4	0.05	6.17	212	Majority vote	6.17	209
2nd Cohort 2014-2015							
k-means	2	0.01	7.410	58	Majority vote	7.410	53
k-medoids	3	0.07	6.29	141	LCA	6.49	113
2nd Cohort 2016-2017							
k-means	2	0.04	7.53	63.1	Majority vote	7.54	61.1
ap	3	0.01	6.44	142.84	LCA	6.44	142
1st Cohort 2012							
k-means	2	0.01	7.94	66	Majority vote	7.94	66
k-means	3	0.04	7.10	86	k-modes	7.10	87

1st Cohort 2013							
k-means	2	0.03	7.92	65	Majority vote	7.9	62
k-medoids	4	0.07	6.393	205	LCE	6.392	202
1st Cohort 2014							
k-means	2	0.03	8.01	69.9	Majority vote	8.01	69.9
k-medoid	3	0.11	7.0	119	LCA	7.05	128
1st Cohort 2015							
k-means	2	0.01	8.21	51.68	LCE	8.24	46
k-means	7	0.22	6.3	266	Majority vote	5.8	212
1st Cohort 2016							
k-means	2	0.02	7.64	65	k-modes	7.64	65
k-medoid	7	0.11	5.66	338	Majority vote	5.44	284
1st Cohort 2017							
k-means	2	0.02	7.72	73.2	Majority vote	7.68	61.12
k-medoids	5	0.04	5.858	270	Majority vote	5.8	252

The clustering ensemble results obtained from the Table 4.9, 4.10, 4.11 and 4.12, were based on multiple cluster runs. The cluster run was set to 20 for all the time cohorts which gave better results. The cluster runs used optimizes the proportion of ambiguous clustering value and aids to obtain the optimal cluster k-value and emphasize the cluster stability which are visualized in the form of heat maps namely consensus matrix in section 4.5.1. The comparison between the two approaches namely individual clustering and clustering ensemble were based on the PAC value and the five internal evaluation indices.

4.4.2 Determination of optimal k-value

To determine the optimal k-value (i.e., the number of clusters), we utilized the Cumulative Distribution Function (CDF) graph for the individual clustering algorithms. Senbabaog˘lu et al [32] extended the idea of understanding the PAC value using CDF graphs. The PAC value was calculated for each clustering algorithm for k-means, k-medoids and affinity propagation to determine the optimal number of clusters in consensus clustering. Figure 4.4 illustrates the cumulative distribution for the cluster ranges from k = 2 to k =7 with the optimal k-value occurring at k = 2 and k = 3 for the clustering algorithms.

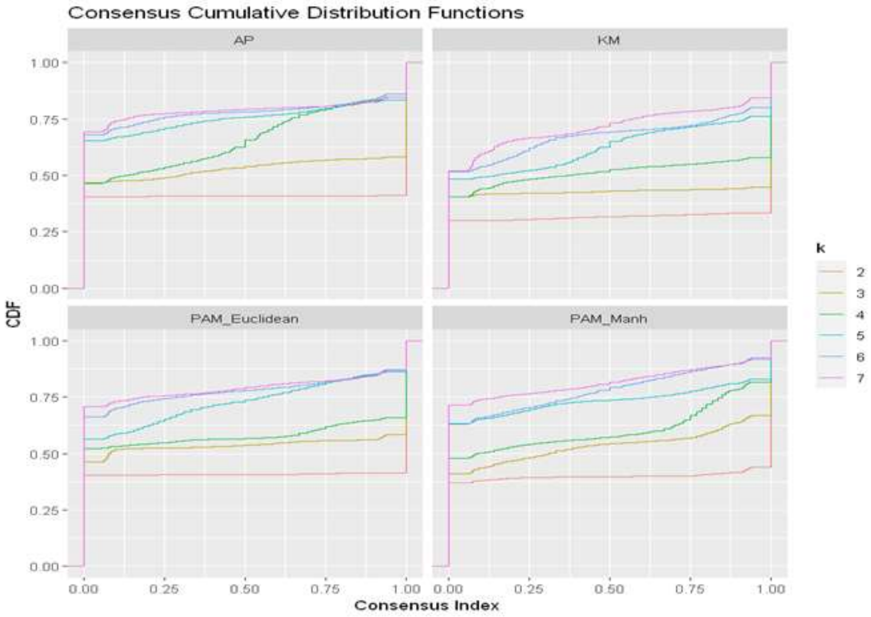

Figure 4.4 CDF graph of k-means, k-medoids and affinity propagation in Cohort III

The lowest value of PAC specifies a flat line segment in the middle is shown by the CDF graphs and lies between intervals [0,1] with 0 indicating a perfect clustering for understanding the cluster stability [62] indicating the optimal K value [32]. The optimal k-value occurs at k=2, k =3, k =4 in all the three time-cohorts, where the PAC value remains the lowest. For all the three time-cohorts, using both PCA and Isomaps based features, the flat line segment appears at k = 2, k = 3 and k = 4 which was consistent on all the three clustering algorithms—i.e., k-means, k-medoids, and affinity propagation. The other cluster ranges i.e., k = 7 and k = 5 were retrieved in rare cases in Isomaps, because there is a possibility of the lowest PAC value in the three time cohorts. Mainly, the optimal k-value of the clustering results were evaluated at cluster runs reps = 20. Table 4.13 demonstrates the optimal k-values for all the time-cohorts in PCA. Though the cluster value k = 2, produces better internal evaluation indices scores, the centroid models such k-means are convex and isotropic because of inertia, always converges and remains optimal at lower cluster values. Thus, the optimal cluster k values were deemed to be k = 3 for the 3rd cohort, and k = 3 or 4 for the 1st and 2nd cohort in PCA.

Table 4.13 Optimal k-values

Time-cohorts	Optimal k-value
1st Cohort 2012, 2013, 2015, 2017	3
1st Cohort 2014, 2016	4
2nd Cohort 2012-2013	4
2nd Cohort 2014-2015, 2016-2017	3
3rd Cohort 2012-2017	3

4.4.3 Performance of Clustering based on Feature Extraction

In order to analyse the difference between linear and non-linear dimensionality reduction algorithms and its effect of changes in clustering was mainly determined by the factor termed Compactness. The measure of compactness is an important criterion which assess the intra-cluster homogeneity and illustrates how well the data points of the cluster are closed to each other[29].

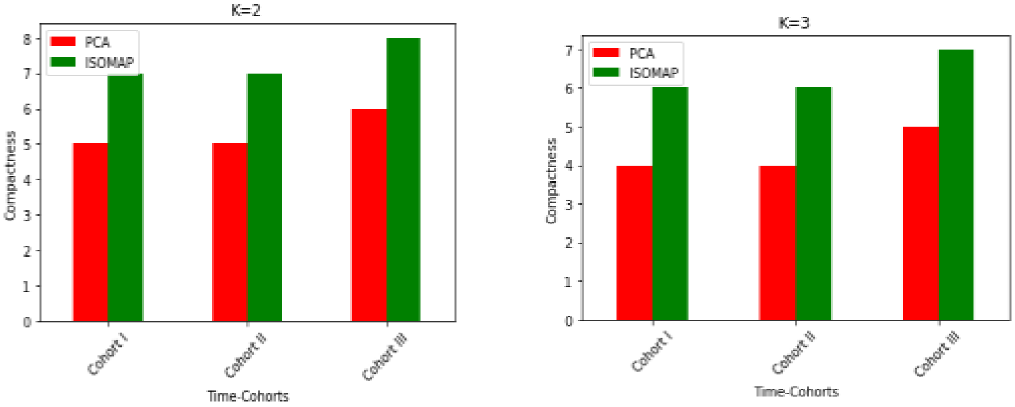

Figure 4.5 Compactness score of PCA and Isomaps at k =2 and k=3.

The clustered bar charts as in the Figure 4.5 represent the compactness score on three time-cohorts between PCA and Isomaps at cluster value k =2, k = 3. The overall value of internal evaluation index Compactness in PCA achieves lower compactness by providing better clustering profile than Isomaps. In the first cohort and the second cohort of PCA at optimal cluster values 2,3 and 4, the compactness value lies around 5 and 4 as well as reaches a maximum score of 6 in the third cohort. Conversely, the compactness score is high in Isomaps was reaching a compactness value of around 8 and 7 at the two cluster values. Moreover, the internal evaluation index Calinski Harabasz causes a difference between two dimensionality reduction algorithms. The Calinski Harabasz index should be maximum and apprehends the combination of compactness and separation together. This value is higher in PCA than Isomaps at optimal k-values.

4.4.4 Comparison between Individual Clustering and Clustering Ensemble

We compared the two clustering approaches namely the individual clustering and clustering ensemble approaches in a 15 dimensional space data of PCA and Isomaps. We assessed the comparison based on internal evaluation indices at the optimal cluster k values. The clustering results were calculated at reps = 20 for each cluster value that determines the cluster runs and the number of subsamples considered.

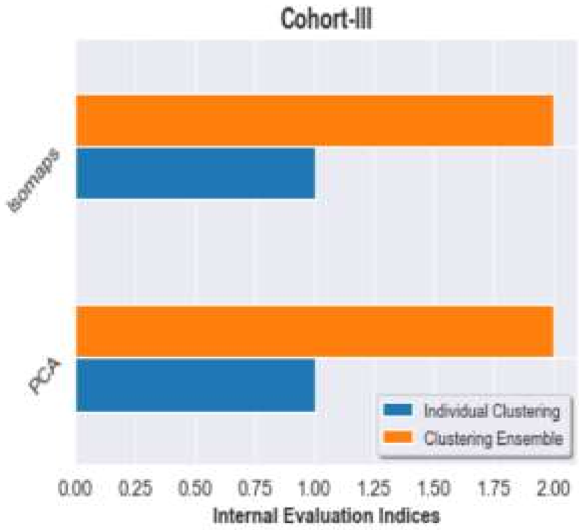

Figure 4.6 Internal evaluation indices of PCA and Isomaps at optimal clusters of Cohort III

Figure 4.6 depicts the comparison between the two approaches in the third cohort of PCA and Isomaps for the optimal cluster ranges. When using the PCA features, for k = 2 and k=3, the Dunn index, Silhouette score was significantly higher and connectivity score was minimal for the clustering ensemble approach at k = 2 as compared to all the individual clustering algorithms, with k =3 , the internal evaluation indices were comparably similar to only k-means and not the latter algorithms k-medoids and affinity propagation —the majority voting ensemble technique provided the best clusters when using the ensemble clustering approach. For comparison purposes, the k-means was used as it provided the lowest PAC value score when compared with the other clustering algorithms. Similarly, for clustering results generated using the Isomaps features, the clustering ensemble performed better than the individual clustering algorithms.

Figure 4.7 illustrates that comparison between the ensemble clusters and the individual clusters for k= 2, 3 and 4 of Cohort II. The cluster ensemble technique performs significantly better for all evaluation metrics based on the PCA features. However, the clustering results are inconclusive based on Isomaps features. The indices of Dunn, silhouette and Calinski Harabasz scores and Connectivity show a huge difference between the ensembles and k-means, k-medoids and affinity propagation.

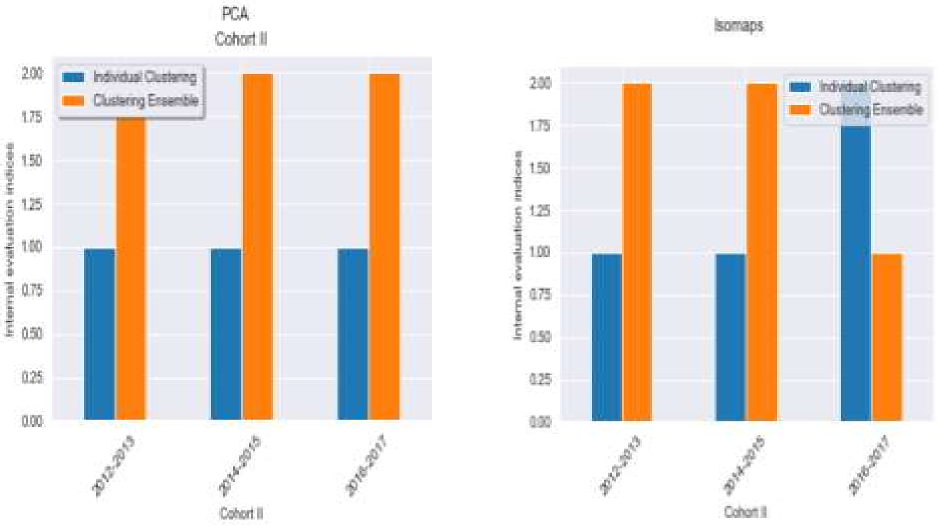

Figure 4.7 Internal evaluation indices of PCA and Isomaps at optimal clusters of Cohort II

Lastly, the first cohort results in Figure 4.8 using PCA, for k = 2, 3 and 4 follow a similar pattern as cohorts 3 and 2, where the cluster ensemble performs better than the individual algorithms. whereas, for Isomaps the comparison is indecisive to draw a meaningful interpretation.

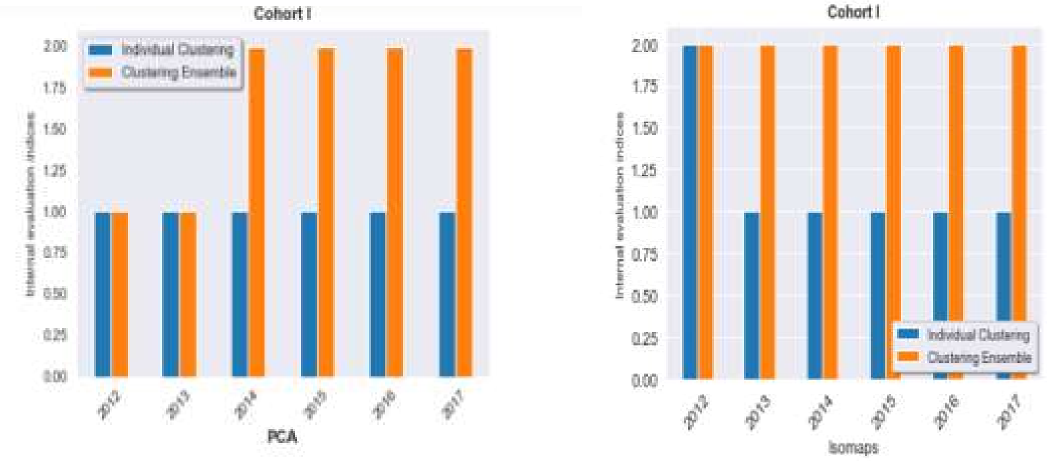

Figure 4.8 Internal evaluation indices of PCA and Isomaps at optimal clusters of Cohort I

In conclusion, our evaluation results confirm that the clustering ensemble approach performs significantly better—i.e., provides more robust clusters—compared to the individual clustering algorithms throughout all the three cohorts of the 15 dimensional data of PCA. In most of the cases, the Majority voting consensus function performs the best amongst all other cluster ensemble models. With the reduced dimensions of PCA applied to both the clustering approaches, there is a small degree of difference between the two approaches. This was due to multiple cluster runs and subsampling technique applied to individual clustering algorithms which make them perform equally well with the clustering ensemble techniques only in some cases.

We also noted that the performance of clustering ensembles is further enhanced with higher k values, in comparison to individual clustering approach. Interestingly, the k-means and k-medoids algorithms were prominent in most of the individual clustering experiments. Topchy et al proposed that the weaker clustering performed effectively when utilized with proper consensus function [10]. Moreover, PCA provides better clustering configuration by achieving the main criteria of compactness, separation, and connectivity, as compared to Isomaps. The compactness and connectivity indices were also greater in non-linear dimensions.

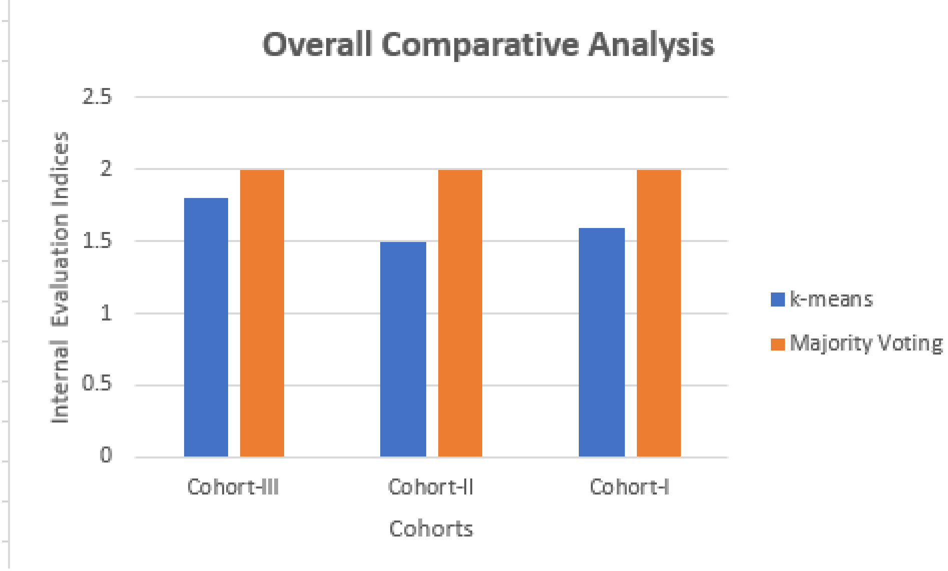

Figure 4.9 Overall Comparison

From diverse experiments of the two clustering approaches as in the Figure 4.9 of overall analysis, i.e. ensemble clustering executed is superior across the three time-cohorts than the three individual clustering algorithms k-means, k-medoids and affinity propagation by comparing the internal validity indices with respective criteria to the structure of the cluster i.e. Compactness and Connectivity and well-separation of the cluster i.e Silhouette index, Dunn index, and Calinski Harabasz. Out of the five consensus functions experimented, the relabeling and voting strategy – majority voting performed better as it solves the label correspondence problem and appears to be the most accurate ensemble approach for the existence of relation[10] between the three clustering algorithms. The ensemble of the three centroid models fits the best for the pathology test ordering data.

4.5 Cluster Visualization

The clusters formed by the clustering ensemble was visualized using the CDF and consensus matrix as it provides insights to the cluster stability. Finally, we visualized the clusters using t-SNE in order to retrieve a well-defined clusters.

4.5.1 Consensus Matrices

Consensus matrix determines the cluster stability for each cluster run of the individual clustering k-means, k-medoids and affinity propagation before applying the consensus functions. Monti et al proposed the consensus matrix, otherwise called the connectivity matrix to determine the cluster stability for multiple cluster runs using resampling technique[30]. Figure 4.10 indicates the clustering stability of k-means, k-medoid or PAM (Partitioning around medoids) and affinity propagation for multiple cluster runs at the optimal cluster k = 3.

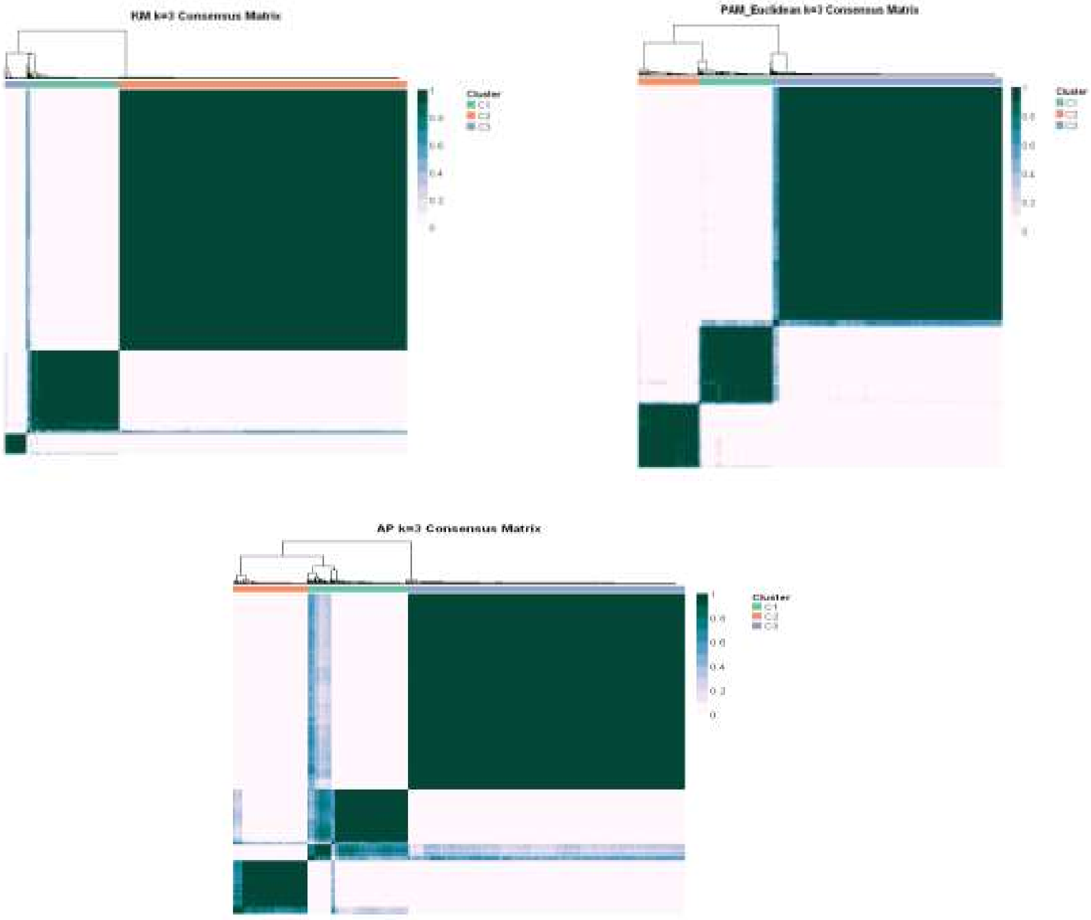

Figure 4.10 Consensus matrix for k-means, k-medoids and affinity propagation at k = 3

The consensus matrix is calculated only for the individual clustering algorithms such as k-means, k-medoid and affinity propagation. The consensus heat map shows lucid clusters at the three ranges without any distortion. If the distortion occurs, the clustering stability starts to fade and turns into a unimodal distribution without the ability to form clusters.

4.5.2 Visualization of Clusters

We visualized the clusters generated by the clustering ensemble approach by applying the t-distributed stochastic neighbor embedding visualization, using Rtsne package in R.

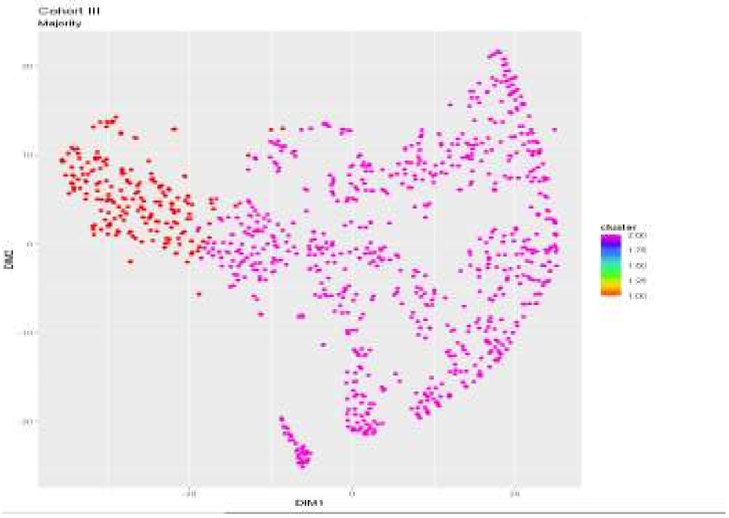

Figure 4.11 Visualization of cluster at k = 2 of Cohort III

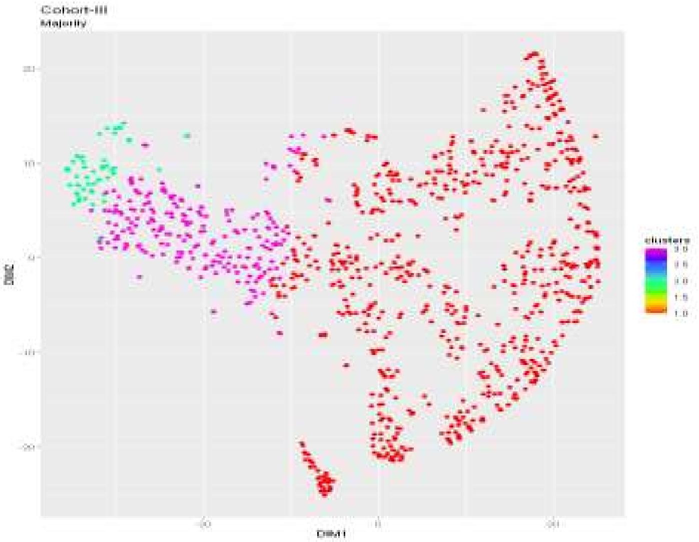

Figure 4.12 Visualization of cluster at k = 3 of Cohort III

Figure 4.11 and Figure 4.12 illustrates the results of the clustering ensemble (using the majority voting consensus metric) for the optimal k-values of k = 2 and k =3. We experimented different cluster values from k =2 to k =7 to visualize the clarity of the clusters, but only the important optimal cluster k- values with the best consensus function was chosen in Figure 4.11 and Figure 4.12 t-SNE provided well-separated and accurate clusters in a two dimensional plot.

4.6 Interpretation of Clusters

The analysis of the clusters was performed to interpret the characteristics of physicians and patient case-mix. This analysis is done to understand the test ordering behaviour of the physicians remains similar or varies in the three time cohorts namely 1st cohort, 2nd cohort and 3rd cohorts. We analysed the clusters generated by the clustering ensemble approach as follows:

4.6.1 Feature Importance

We retrieved the clusters for the three time cohorts. The data was labeled for the entire dataset, we included all the features such as the number of orders, pmb id, different age groups, sex (male and female), ratio of age groups and gender information, normal and abnormal test orders, normal and abnormal disorders. We used the optimal cluster k-value for each cohort as in the Table 4.13.

In order to interpret the clusters, we scaled all the features using z-score normalization and calculated the variance of means between the clusters for each feature, finally choosing the best features based on high variance.

We classified the 26 disorders into four groups such as anemia, hematology, biochemical1 biochemical2 based on disease ontology i.e., the hematology group deals with the blood disorders such as lymphoma, leukaemia, thrombocytopenia, and bleeding disorder whereas anemia group represents the blood deficiency disorders namely hemolytic anemia, hemorrhagic anemia, iron deficiency, vitamin B12 folate deficiency, bone marrow failure and polycythemia vera. Moreover, the disorders that is genetically present and affects the biochemical process of the body results in biochemical disorders. The first set of biochemical disorders consists of liver biliary, lung, kidney, thyroid, dehydration and addison disease and mineralocorticoid excess disorder. The second set of biochemical disorders consists of diabetics mellitus, diabetics ketoacidosis, hemodilution SIADH, metabolic bone disease, low muscle mass, hypoglycemia, muscle injury hemolysis, inflammatory conditions, parasitic infection allergy.

We conducted the variance analysis by taking the mean of the z-scores and grouping the features based on the clusters for all the three time cohorts. The range of mean score falls between 0 and 3.5 in the time-cohorts. The mean score of less than 0 represent a negative mean score which does not contribute the importance of the feature in the analysis.

- **Age Groups and Gender Information**

Firstly, we analysed the different age groups of 0-18, group 19-30, group 31-50, group 50-65 and group 66+ in the three time -cohorts. The features containing ratio of the age groups and gender information was taken for the feature importance analysis, as it represents the proportion. Figure 4.13 illustrates the bar chart of different ratio of age groups present in the three clusters of Cohort III reaching a mean score of 0.40 for age group ratio 66 in cluster 2. The most important feature found was age group ratio 66 and group ratio 50-65 in all the three time cohorts because of high mean variance in the clusters, whereas the age groups 31, 19 and 0 were important and in some cases, provided a variance score of less than 0. So, the features that provided less variance score below 0, were clipped off due to low feature importance as well as it does not contribute in the feature analysis in the clusters.

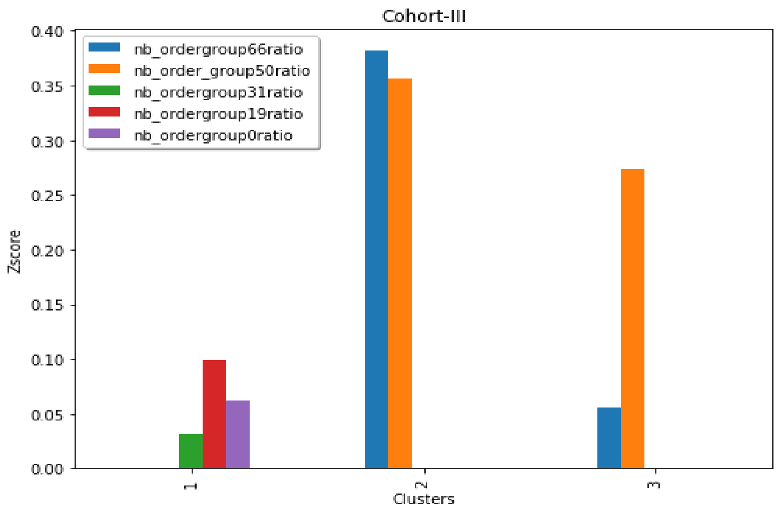

Figure 4.13 Different age groups in Cohort III

The age group ratio 50-65 was high in the last two years (2016-2017) of the laboratory utilization. Following that, the gender information consisting of male and female, variance of male was higher than the female in all the three time cohorts whereas the female ratio was high only in the last year of 2017.We visualized the features using bar charts which illustrates the important features in each cluster and removes the unimportant features which falls below variance 0.

Normal Test Order and Abnormal Test Order

The physician test ordering behaviour of normal and abnormal test order remains similar throughout the 1st cohort, 2nd cohort and 3rd cohort. By using the z-score normalization and grouping the clusters by taking the mean of all the normal and abnormal test orders, the most important features were chosen. The normal test orders such as CBC, Urea, Creatinine, Electrolyte panel, Glucose AC, ALT, AST, TSH have higher importance because of high mean score were found in the clusters. On the other hand, the PT, GGT, Glucose Random and Alkaline Phosphatase were present low in all the three time-cohorts. Figure 4.14 illustrates the normal test orders present in the clusters of Cohort III.

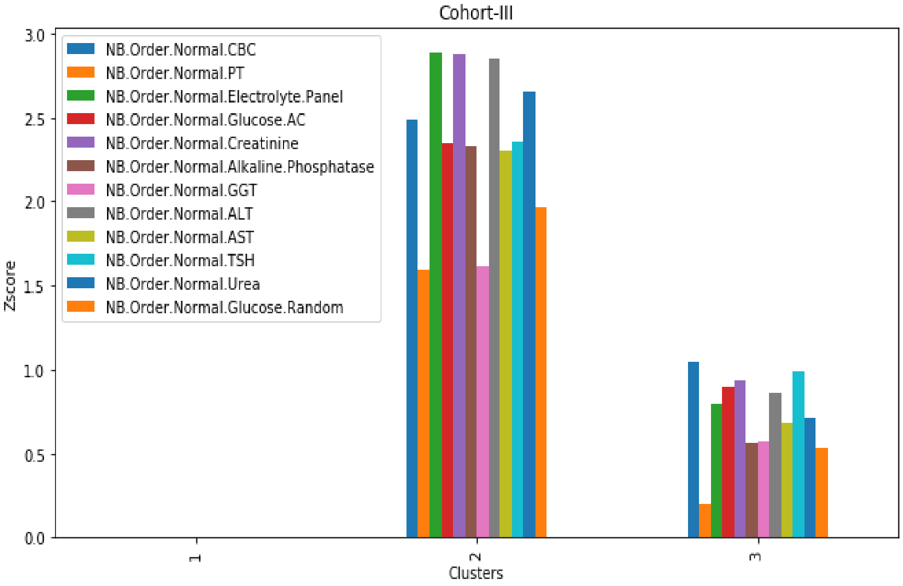

Figure 4.14 Normal test orders present in the clusters of Cohort III

With the abnormal test orders namely CBC, ALT, Electrolyte Panel, Glucose AC, Creatinine, TSH, AST and PT were recurrent test orders in the three time-cohorts whereas

Urea, Alkaline Phosphatase, GGT and Glucose Random represent low mean in the clusters of 1st, 2nd and 3rd Cohort.

- **Normal Disorder and Abnormal Disorder**

The normal and abnormal disorders were categorised into four groups in all the time cohorts. Figure 4.15 indicates all the features in the anemia group. Firstly, the anemia group consisting of the features namely hemolytic anemia, hemorrhagic anemia, iron deficiency, vitamin B12 folate deficiency, bone marrow failure and polycythemia vera were highly significant because of high mean seen in the clusters from both the normal and abnormal disorder section.

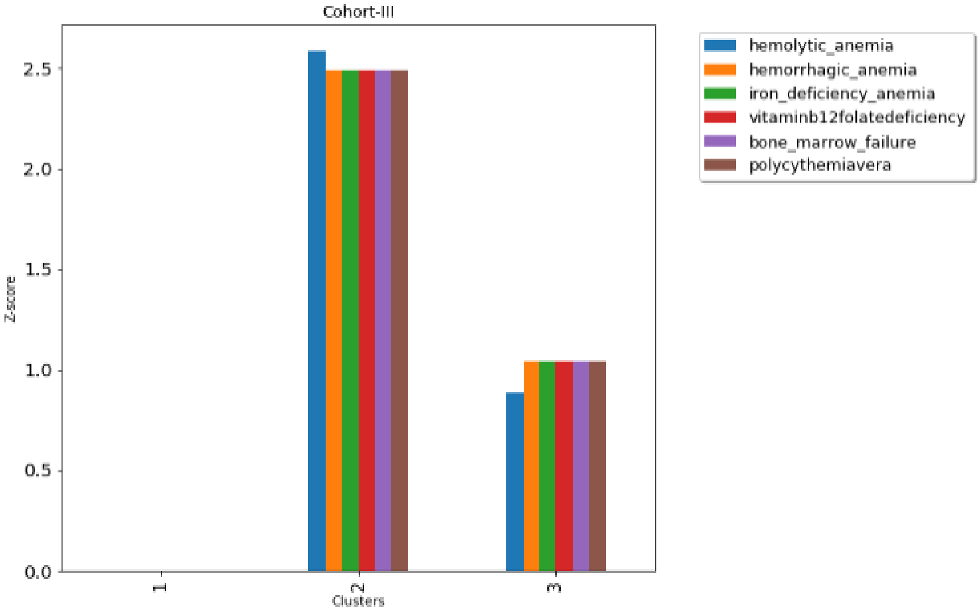

Figure 4.15 Features present in the clusters of anemia group of Cohort III

Figure 4.16 represents the features from the hematology group. From the group of haematology, the bleeding order remains less important feature among the other disorders namely lymphoma, leukemia and thrombocytopenia having a high score of variance in the three time-cohorts in all the three clusters.

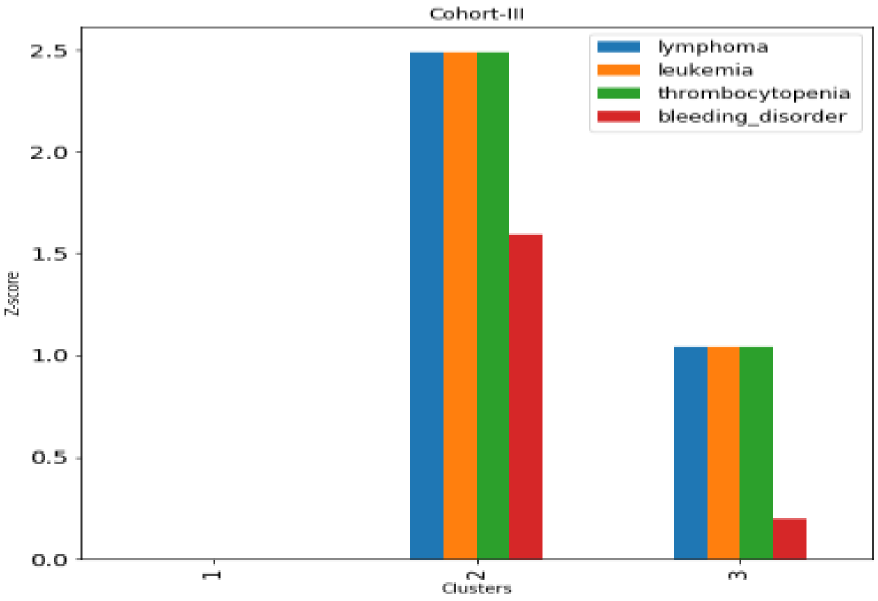

Figure 4.16 Features present in the clusters of hematology group of Cohort III

From the first set of biochemical disorders, the most important features that occur in the normal disorder section were kidney disorder, lung disorder, mineral excess corticoid and dehydration, liver biliary disorder, addison disease and thyroid disorder. All the features were considered important.

Figure 4.17 represents the features present the clusters of biochemical 1 group from the normal disorder section. In the abnormal section, we found some disorders were highly significant such as lung disorder, dehydration, kidney disorder, addison disease, mineral excess corticoid, liver biliary disorder whereas thyroid disorder feature present is relatively low.

From the second set of biochemical disorders namely hemodilution SIADH, low muscle mass, diabetics mellitus, diabetics ketoacidosis, inflammatory conditions, parasitic infection allergy are the important features in the clusters from both the normal disorder and abnormal disorder section as in the Figure 4.18. On the other, metabolic bone disease, muscle injury hemolysis and hypoglycemia are relatively low. The features such as diabetic mellitus, and diabetics ketoacidosis, parasitic infection allergy and inflammatory conditions remain in the same level of variance in the three time cohorts.

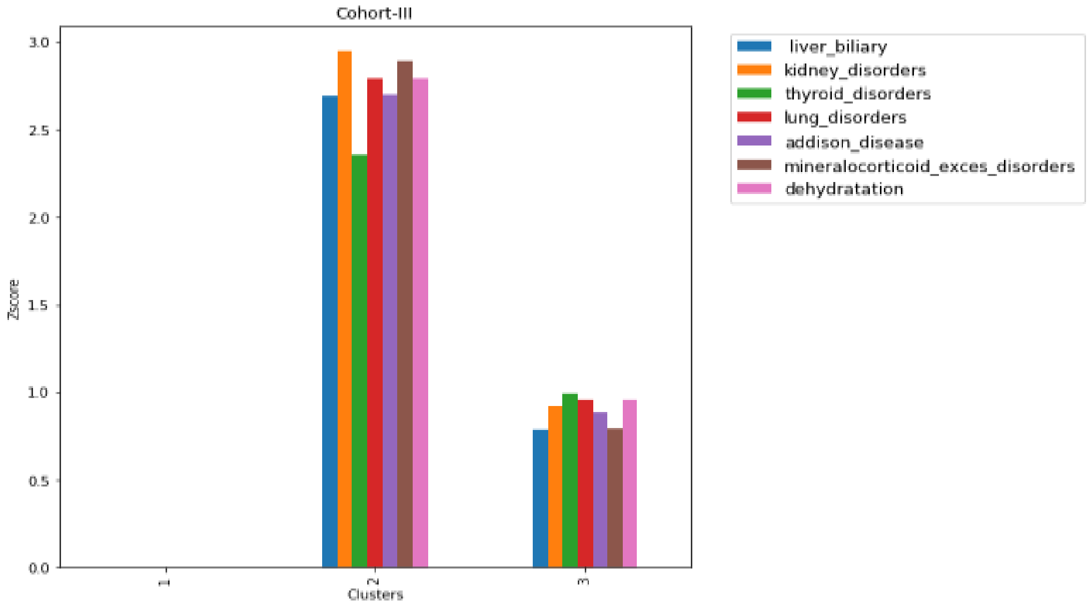

Figure 4.17 Features present in the clusters of biochemical 1 group of Cohort III

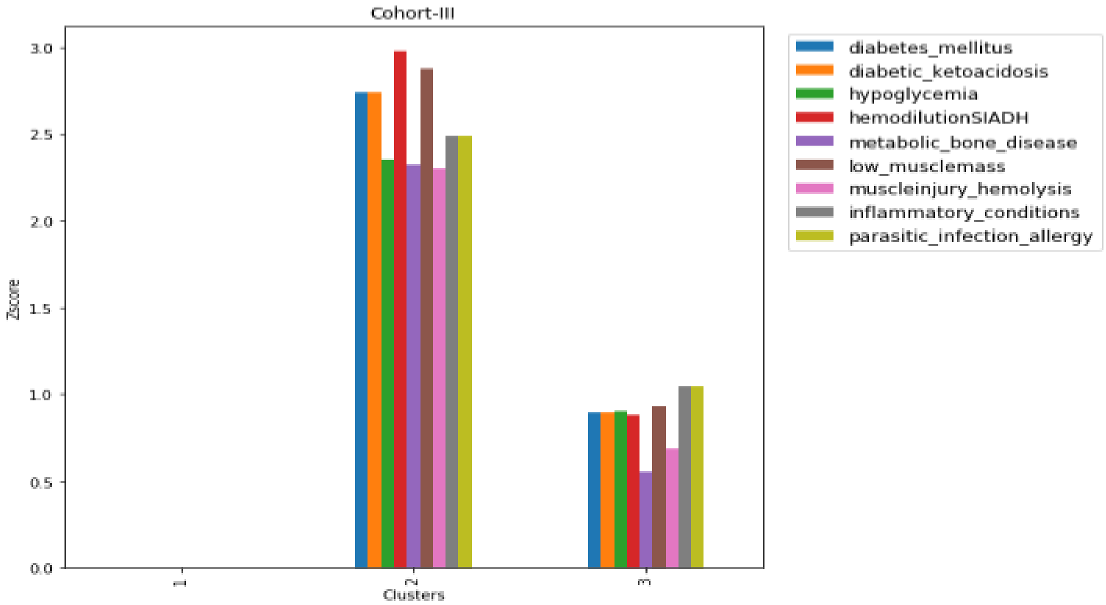

Figure 4.18 Features present in the clusters of biochemical 2 group of Cohort III

Figure 4.19, Figure 4.20 and Figure 4.21 indicates the importance features obtained from the overall analysis of cluster values in Cohort III, Cohort II (2014-2015) and Cohort I (2015).

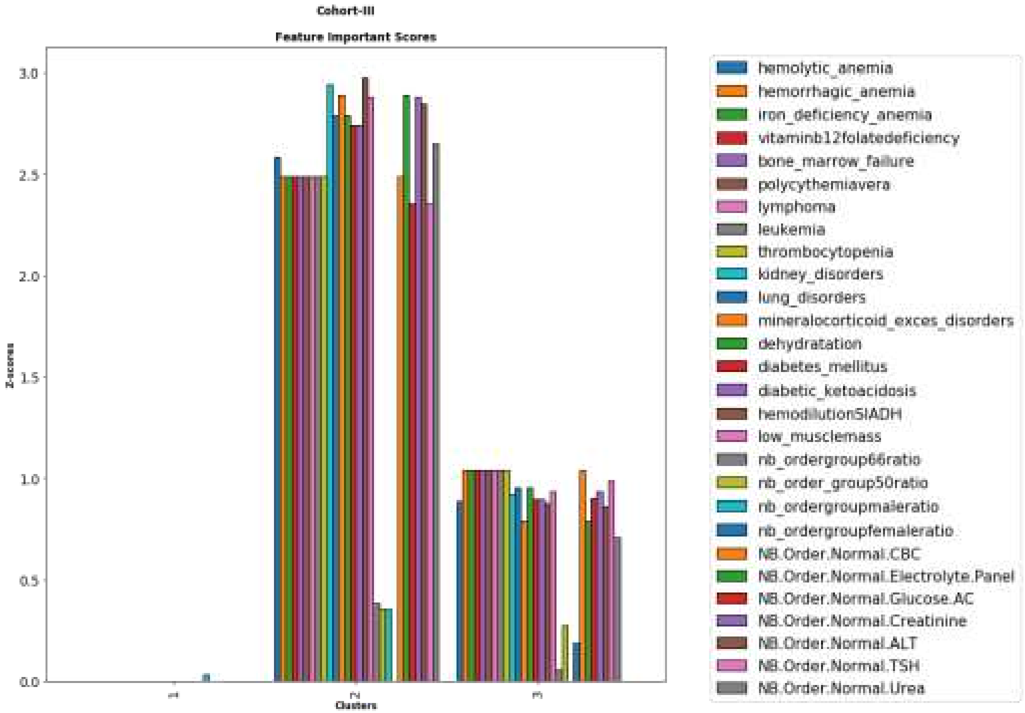

Figure 4.19 Overall Feature Importance Scores in Cohort III

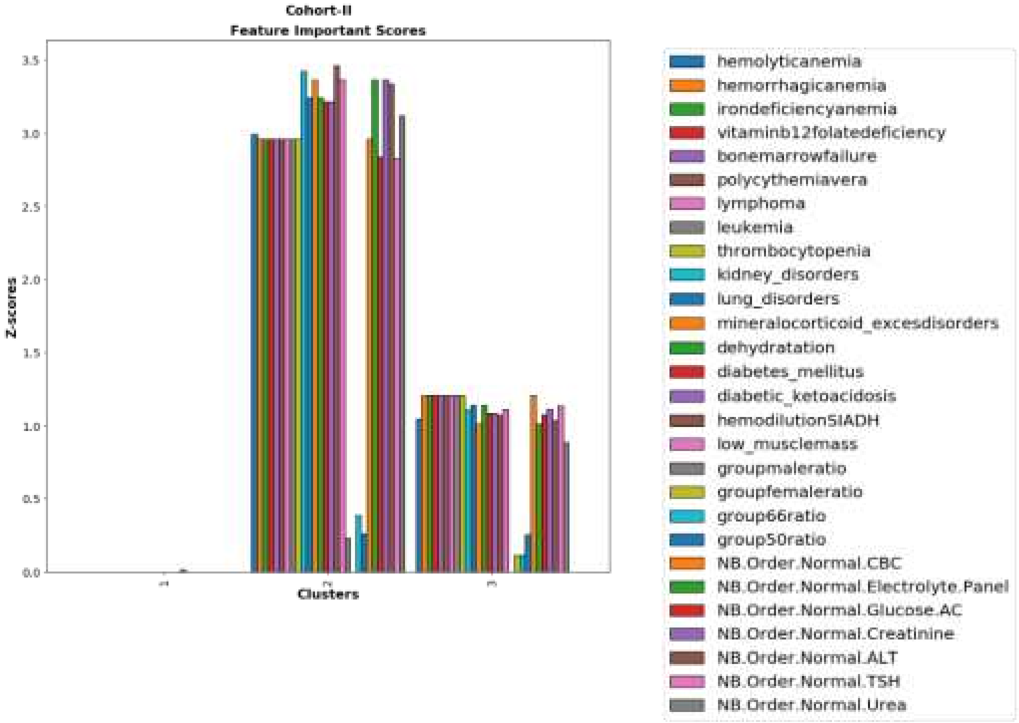

Figure 4.20 Overall Feature Importance Scores in Cohort II 2014-2015

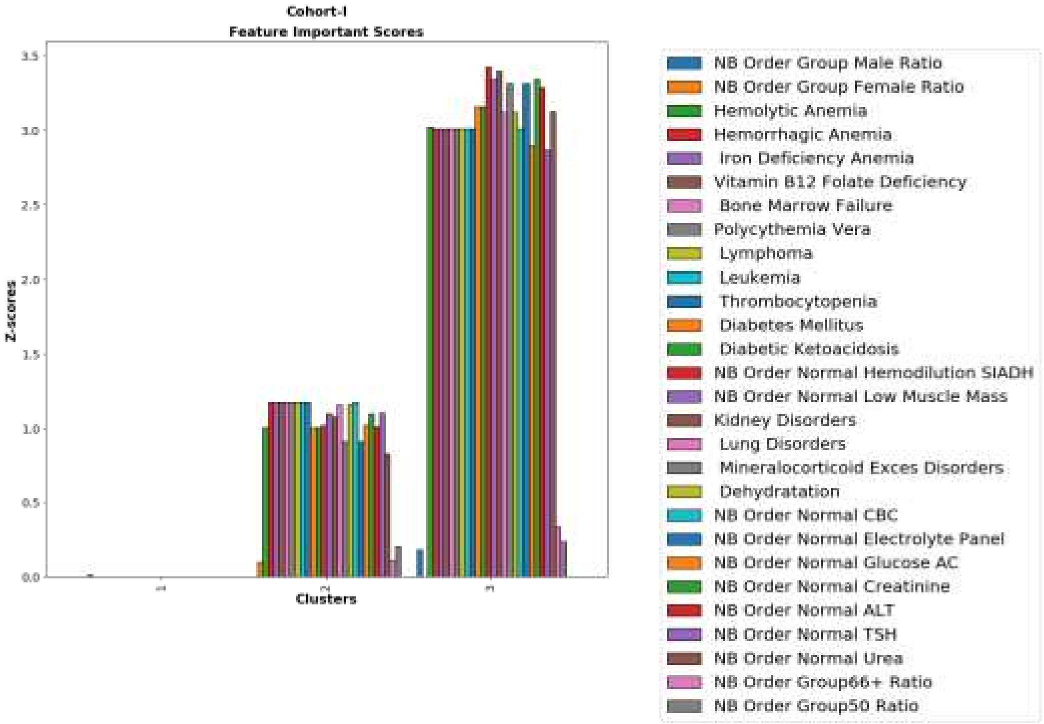

Figure 4.21 Overall Feature Importance Scores in Cohort I 2015

From the overall feature importance scores across the three time-cohorts, the phenotypes starting from the important age groups such as age group ratio 66 and age group ratio 50-65, gender information namely male and female. The patient characteristics remain the same and stable in three time-cohorts with these four presiding phenotypes in the clusters.

Following that, the hemolytic anemia, hemorrhagic anemia, iron deficiency, vitamin B12 folate deficiency, bone marrow failure and polycythemia vera from the anemia group and the hematology group consisting of important features such as lymphoma, leukemia and thrombocytopenia remain undeviating with high mean score between 3.0 and 3.3 across Cohort III, Cohort II and Cohort I.

Moreover, the highest peak disorders observed from the Figure 4.19, 4.20 and 4.21 such as kidney disorder, lung disorder, mineral excess corticoid disorder, and dehydration from biochemical1 group and hemodilution SIADH, low muscle mass, diabetic mellitus, and diabetics ketoacidosis from biochemical2 group show the cluster consistency across the physician clusters.

The tests important in the cluster values were CBC, creatinine, electrolyte panel, urea, ALT, TSH and Glucose AC in the three different time-cohorts. Thus, the important phenotypes the patient characteristics, tests and important disorders found inside each physician cluster value from the three time-cohorts namely Cohort III, Cohort II and Cohort I is persistent and maintain the cohesiveness in the test ordering pattern.

- SIMILARITY PATTERNS OF PHYSICIAN CLUSTER ACROSS TIME

To acknowledge the physician clusters are consistent or vary over time in the first cohort. We took a same subset of 20 physicians on all the three time-cohorts chosen for the analysis by taking the physician id i.e. PHY1, PHY2, PHY3 with their cluster labels who fall in the same or different clusters in 2012, 2013 and so on**.**

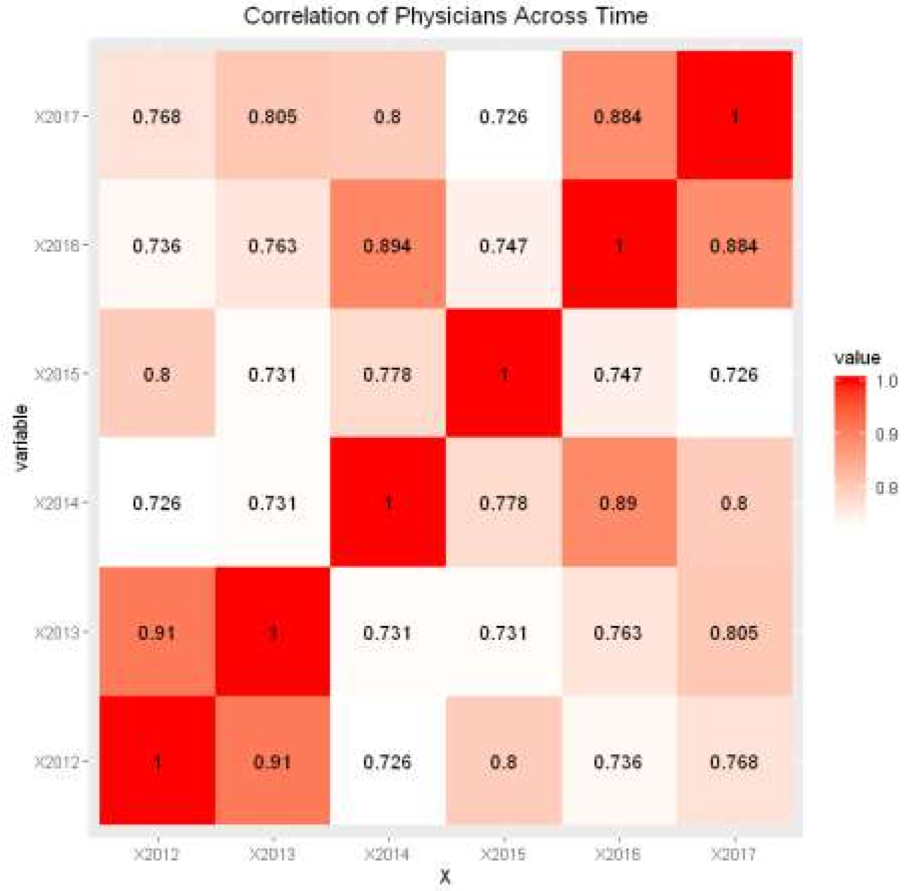

Figure 4.22 Correlation of physicians

Figure 4.22 gives correlation and similarity of physicians with the clustering results across the years. We used the external validity Rand Index statistic to identify the clustering similarities across years, by comparing the distinctive and similar pair of items found in the physician clusters. The Rand index value lies between 0 and 1, with 1 presenting the clustering are most similar whereas 0 represent clusters are not the same. The Rand index works with conditions by pairing items with four combinations of same and different subsets[97].

We compared the clustering results of the physicians across the different years from 2012, 2013, 2014, 2015, 2016 and 2017 of the first cohort. From the correlation matrix, the physicians in the clusters from 2012 and 2013 have similarity and consistent with the rand index score of 0.91. Secondly, the years across 2016 and 2017 have the most similar score of 0.884, indicating that the physicians are correlated. Likewise, the physicians correspond across the years of 2014 & 2016, 2014 & 2017, 2013 & 2017 have a rand index score of around 0.8. On an overall basis, the rand index score of the physicians clusters lies in the range between 0.7-0.9 similarity score, which presents the clusters are precisely similar for the 20 physicians across 2012, 2013 and so on in the first cohort.

4.6.2 Discussion

In this research, we presented the clustering ensemble approach to identify the physician phenotypes that determine the test ordering pattern of physicians in the three time-cohorts. We compared two clustering approaches such as individual clustering approach and clustering ensemble approach. Since the clustering ensemble approach has the ability to predict the groups based on the combination of two or three algorithms together, it has better cluster stability and since it utilizes the subsampling method for multiple cluster runs.

Before clustering, we extracted the features in the three time-cohorts using two approaches linear and non-linear dimensionality reduction. The baseline was set as PCA that reduces the data in a linear way whereas the non–linear dimensionality reduction algorithms Isomaps and Deep Auto-encoder (DAE) were experimented. We analysed different dimensions to select the best dimension for clustering ensemble approach. In order to evaluate and assess the quality of the reduced dimensions in an unlabelled data, we introduced ranking techniques termed as co-Ranking, LCMC and AUC_ln_K, R_{NX} which ranks the data based on the projection from a high dimensional data into a low dimensional space. We tested the two metrics for three dimensionality reduction algorithm for different dimensions. The PCA performed better than the non-linear dimensionality reduction algorithms and reduced into 15 dimensional space with AUC_ln_K scores of 0.86-0.93, 0.91-0.95, 0.91 in the 1st, 2nd and 3rd cohorts. The LCMC score of PCA produced a perfect embedding that is visualized in the form of line plot that peaked a score of 0.8. The main of purpose of co-Ranking is to assess the quality of the reduced dimension does not tear data apart in dimensionality reduction algorithm. Conversely, the Isomaps produced moderate embedding of 0.66, 0.67 and 0.69 in the 1st, 2nd and 3rd cohorts. Moreover, random embeddings were exhibited by using deep auto-encoders of 0.57, 0.59 and 0.53 in the 1st, 2nd and 3rd cohorts. The linear dimensionality reduction algorithm performed better than the non-dimensionality reduction algorithms in the pathology data.

We built clustering ensembles based on the centroid models k-means, k-medoids and affinity propagation and applied five consensus functions namely majority, k-modes, LCA, LCE and CSPA for generating the ensemble. The cluster ensemble majority voting performed better than the individual clustering algorithms based on the five internal evaluation indices. We utilized 15- dimensional space of PCA and Isomaps to compare the linear and non-linear dimensionality reduction algorithms as an input for clustering. Compactness and Calinski Harabasz were the two indices which differentiates PCA from Isomaps, where PCA provides better clustering structure and yields better clustering results.

In order to obtain the best k-value in a clustering ensemble approach, we used proportion of ambiguous clustering (PAC) to identify the lowest possible value from the three clustering algorithms. We visualized the optimal cluster value using Cumulative distribution function (CDF) for multiple clusters runs for the clustering algorithms. The optimal cluster k values were found at k = 3 in the 3rd cohort, and k = 3, 4 in the 2nd and 1st cohort respectively. The consensus matrix was used, to check how accurate the clustering and stable. The k-means and k-medoids executed better PAC values in the three time-cohorts than affinity propagation. We took the optimal cluster values for comparing the two approaches. The clustering ensemble majority in 3rd cohort in PCA executed better than the individual clustering algorithms based on internal evaluation indices such as Calinski Harabasz, Silhouette coefficient, Dunn index, Compactness and connectivity. Following that, majority and LCA performed better in the 2nd cohort than k-means and k-medoids, in addition, majority, LCE and k-modes achieved in the 1st cohort. By overall analysis of all the ensemble approaches, the object occurrence based method Majority performed well in the pathology dataset.

Menger and Spruit utilized the clustering ensemble approach for identifying the sub-types of psychiatric patients. In their study, k-means, GMM and affinity propagation was used as an ensemble combination and used five consensus functions such as majority, k-modes, CSPA, LCA and LCE. But the best performing algorithm was graph-based clustering ensemble approach CSPA to identify the sub-types namely depressive disorder, speech and behaviour problems in each cluster[98]. The study was carried out for 12 weeks for understanding the behaviour of the psychiatric patients.

Similarly, the most significant features were found using the variance analysis in the three time-cohorts. The age group 66 and group 50 were the most evident features in the age group criteria in the 1st, 2nd and 3rd cohorts. With the gender information, most of the cluster falls to the male population except the last year of 2017 in the 1st cohort shows the higher female ratio.

Different groups of disorders were sorted in the three time-cohorts such as anemia group, hematology group, biochemical1 group and biochemical2 group. By analysis, the anemia group and hematology group, biochemical1 group and biochemical2 group follow the same pattern proportion in the clusters throughout the time intervals of 1st, 2nd and 3rd cohorts.

To begin with, hemorrhagic anemia, iron deficiency, vitamin B12 Folate deficiency, bone marrow failure and polycythemia vera, hemolytic anemia has the same level proportion in the three time-cohorts. Adversely, the most principal features that conquer the cluster significance in biochemical1 group are kidney disorder, lung disorder, mineral excess corticoid, dehydration, liver biliary, thyroid disorder and addison disease in normal disorders, whereas the hemodilution SIADH and low muscles mass, diabetics mellitus, diabetics ketoacidosis, inflammatory conditions, parasitic infection allergy were the prominent features dominating the three time-cohorts. The abnormal disorders have the same pattern as that of the normal disorders of anemia and hematology group. But this pattern differs in the abnormal disorders of biochemical1 group with lung disorder and dehydration conquering the highest in the clusters.

The most significant features in the number of tests ordered in the cohorts was CBC, electrolyte panel, creatinine, urea, ALT TSH and Glucose AC were dominant in the clusters. On the other hand, test orders such as GGT, PT remain low important features. We interpreted the similarity pattern of the physician clusters who remain similar and consistent over time.

To conclude the above discussion was solely based on the results produced by the clustering ensemble models and the significant features were obtained from the significant clusters of these models in the three time-cohorts. We made a comparison between the linear dimensionality reduction algorithm and non-linear dimensionality reduction algorithm. We compared two clustering approaches such as individual clustering and clustering ensemble approach. We interpreted the significant features found from each cluster and concluded the overall variance analysis.

CHAPTER 5 CONCLUSION

5.1 Summary

Pathology laboratory test ordering is beneficial for understanding the outcomes of laboratory utilization process. It entails the number of test orders ordered by the physicians and is essential to give the right amount of treatment at the right time for the patients. In this research, we clustered physicians among the peers having the similar patient case-mix by utilizing the Nova Scotia Health Authority (NSHA) pathology data from 2012-2017. We conducted the study into three time cohorts based on 1-year, 2-year and 6-year interval to understand the physician ordering behaviour of test orders. Test-disease mapping was retrieved from the previous studies, was utilized in this study between the number of tests and resulted about 26 disorders in the overall study of six years. We introduced an ensemble approach for clustering, by building heterogeneous cluster ensemble based on the centroid models such as k-means, k-medoids and affinity propagation. We initiated five cluster ensembles namely majority, LCA, k-modes, LCE and CSPA. Before the clustering process, we extracted the features using three dimensionality reduction algorithms such as PCA, Isomaps and Deep auto-encoder and assessed the quality of the feature extraction algorithms using co-Ranking. Our results indicate that PCA performed the best with area under the R_{NX} curve score of 91% at the best dimension with LCMC curve at 0.9 in the overall study. We utilized a 15-dimensional space data from PCA and Isomaps on clustering. Our results for clustering ensemble were based on the five internal evaluation indices, in which majority is based on cumulative voting performed well than the individual clustering algorithms on all five indices in the time-cohorts. We used the proportion of ambiguous clustering (PAC) metric for identifying the best k-value as well as cumulative distribution functions and consensus matrix for understanding the cluster stability in a clustering ensemble approach. The best cluster k values were obtained at k = 3 in the 3rd cohort, and k = 3, 4 in the 1st and 2nd cohort respectively. We discovered the physician phenotypes in the three cohorts by using the ensemble cluster labels and the feature importance scores in clusters were calculated by using variance analysis. We found the inherent characteristics depend on the important disorders from the list of 26 disorders, tests, age groups, and sex that influence in the physician test ordering in laboratory utilization for different time-cohorts respectively.

5.2 Limitations

One of the main limitations was computational restraint while training the cluster ensemble process. The dataset was adequate, but the training process in the ensemble process took a lot of computational time. For the feature extraction process using scikit-learn was easy to train the unlabeled data. Conversely, for the cluster ensemble process we used the diceR package, we conducted experiments for a range of cluster values, it took a lot of computational time for training one cluster value. The system crashes several times while running high cluster values as well as when the clustering runs are increased above the limit of the system due to unavailable space in the RAM. In addition, for LCE ensemble, we faced several crashes while training with 'srs' and 'cts' matrix because loss of computational power. Following that, one of the functional limitation of diceR package, we were not able to train five or six clustering algorithms together at once while selecting algorithms because of high computational power and time, we had to try it only three algorithms at a time separately. We implemented co-Ranking in R but not in scikit-learn package, because the package was useful for visualizing purpose, but it did not obtain the important details of assessing the quality of the dimensions.

5.3 Future Works

The approach presented open wide opportunities for various dimensions. Since we had worked on an unlabeled data, we were able to validate the clustering with few internal validation indices. By the expert's domain and their knowledge to provide class labels on the pathology data, could help us train the ensemble approach on supervised data with different indices such as NMI, ARI, accuracy, f-measure, Cohen's kappa and various others external validity indices could be tested for validating clustering for the future work.

The main purpose for clustering ensembles was to provide views of multiple clustering algorithms to achieve one robust clustering and deploying the clustering ensemble model for laboratory utilization process aids in day-to-day to track the physician test ordering behaviour.

REFERENCES

[1] S. S. R. Abidi *et al.*, “AI-Driven Pathology Laboratory Utilization Management via Data- and Knowledge-Based Analytics. BT - Artificial Intelligence in Medicine - 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26-29, 2019, Proceedings.” pp. 241–251, 2019.

[2] M. Zhi, E. L. Ding, J. Theisen-Toupal, J. Whelan, and R. Arnaout, “The landscape of inappropriate laboratory testing: A 15-year meta-analysis,” *PLoS One*, vol. 8, no. 11, pp. 1–8, 2013.

[3] N. Delvaux, A. De Sutter, S. Van de Velde, D. Ramaekers, S. Fieuws, and B. Aertgeerts, “Electronic Laboratory Medicine ordering with evidence-based Order sets in primary care (ELMO study): protocol for a cluster randomised trial,” *Implement. Sci.*, vol. 12, no. 1, p. 147, 2017.

[4] A. P. Dorevitch, “The ‘Ulysses syndrome’ in pathology: when more is less.,” *The Medical journal of Australia*, vol. 156, no. 2. Australia, p. 140, Jan-1992.

[5] M. Rang, “The Ulysses syndrome.,” *Can. Med. Assoc. J.*, vol. 106, no. 2, pp. 122–123, Jan. 1972.

[6] D. Bates *et al.*, “Does the computerized display of charges affect inpatient ancillary test utilization?,” *Arch. Intern. Med.*, vol. 157, pp. 2501–2508, Dec. 1997.

[7] E. G. Neilson *et al.*, “The impact of peer management on test-ordering behavior.,” *Ann. Intern. Med.*, vol. 141, no. 3, pp. 196–204, Aug. 2004.

[8] M. K. Sarkar, C. M. Botz, and M. Laposata, “An assessment of overutilization and underutilization of laboratory tests by expert physicians in the evaluation of patients for bleeding and thrombotic disorders in clinical context and in real time,” *Diagnosis (Berlin, Ger.*, vol. 4, no. 1, pp. 21–26, 2017.

[9] W. Levinson and T. Huynh, “Engaging physicians and patients in conversations about unnecessary tests and procedures: Choosing Wisely Canada,” *CMAJ*, vol. 186, no. 5, pp. 325–326, Mar. 2014.

[10] S. Vega-Pons and J. Ruiz-Shulcloper, “A survey of clustering ensemble algorithms,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 3, pp. 337–372, 2011.

[11] D. Dilts, J. Khamalah, and A. Plotkin, “Using cluster analysis for medical resource decision making.,” *Med. Decis. Mak. an Int. J. Soc. Med. Decis. Mak.*, vol. 15, no. 4, pp. 333–347, 1995.

[12] M. Liao, Y. Li, F. Kianifard, E. Obi, and S. Arcona, “Cluster analysis and its

application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis," *BMC Nephrol.*, vol. 17, p. 25, Mar. 2016.

[13] R. K. Blashfield, *The classification of psychopathology: Neo-Kraepelinian and quantitative approaches*. Springer Science & Business Media, 2012.

[14] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 25, pp. 14863–14868, Dec. 1998.

[15] M. R. Weir *et al.*, "Implications of a health lifestyle and medication analysis for improving hypertension control.," *Arch. Intern. Med.*, vol. 160, no. 4, pp. 481–490, Feb. 2000.

[16] J. Clatworthy, D. Buick, M. Hankins, J. Weinman, and R. Horne, "The use and reporting of cluster analysis in health psychology: a review.," *Br. J. Health Psychol.*, vol. 10, no. Pt 3, pp. 329–358, Sep. 2005.

[17] T. Alqurashi and W. Wang, "Clustering ensemble method," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 6, pp. 1227–1246, 2019.

[18] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 2, pp. 1214–1219, 2004.

[19] A. Strehl and J. Ghosh, "Cluster ensembles - A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 583–617, 2003.

[20] A. Strehl and J. Ghosh, "Relationship-based clustering and visualization for high-dimensional data mining," *INFORMS J. Comput.*, vol. 15, no. 2 SPEC., pp. 208–230, 2003.

[21] S. Asur, D. Ucar, and S. Parthasarathy, "An ensemble framework for clustering protein-protein interaction networks," *Bioinformatics*, vol. 23, no. 13, pp. 29–40, 2007.

[22] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, Jun. 2005.

[23] A. Topchy, A. K. Jain, and W. Punch, "A Mixture Model for Clustering Ensembles," in *Proceedings of the 2004 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, 2004, pp. 379–390.

[24] X. Hu and I. Yoo, "Cluster ensemble and its applications in gene expression analysis," *Proc. Second Conf. Asia-Pacific …*, vol. 29, pp. 297–302, 2004.

[25] J. Ghosh and A. Acharya, "Cluster ensembles," *Wiley Interdiscip. Rev. Data Min.*

Knowl. Discov., vol. 1, no. 4, pp. 305–315, 2011.

[26] X. Sevillano, G. Cobo, F. Alías, and J. C. Socoró, "Feature diversity in cluster ensembles for robust document clustering," *Proc. Twenty-Ninth Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, vol. 2006, pp. 697–698, 2006.

[27] J. Ghosh, A. Strehl, and S. Merugu, "A Consensus Framework for Integrating Distributed Clusterings Under Limited Knowledge Sharing," *Proc NSF Work. Next Gener. Data Min.*, vol. 41, no. February 2003, pp. 99–108, 2002.

[28] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1866–1881, 2005.

[29] N. Iam-On and S. Garrett, "Linkclue: A matlab package for link-based cluster ensembles," *J. Stat. Softw.*, vol. 36, no. 9, pp. 1–36, 2010.

[30] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Mach. Learn.*, vol. 52, no. 1–2, pp. 91–118, 2003.

[31] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biol.*, vol. 3, no. 7, p. research0036.1, 2002.

[32] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[33] R. Tibshirani and G. Walther, "Estimating the number of clusters in a dataset via the Gap statistic," *R. Stat. Soc*, vol. 63, 2000.

[34] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data ," *Bioinformatics*, vol. 17, no. 4, pp. 309–318, 2001.

[35] G. Ogbuabor and U. F. N, "Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value," *Int. J. Comput. Sci. Inf. Technol.*, vol. 10, no. 2, pp. 27–37, 2018.

[36] G. Mirzaei, A. Adeli, and H. Adeli, "Imaging and machine learning techniques for diagnosis of Alzheimer's disease," *Rev. Neurosci.*, vol. 27, no. 8, pp. 857–870, 2016.

[37] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, "K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir," *IEEE Access*, vol. 8, pp. 151847–151855, 2020.

[38] J. Escudero, E. Ifeachor, and J. Zajicek, "Bioprofile Analysis: A New Approach

for the Analysis of Biomedical Data in Alzheimer's Disease," *J. Alzheimers. Dis.*, vol. 32, 2012.

[39] S. Vijayarani and S. Sudha, "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples," *Indian J. Sci. Technol.*, vol. 8, 2015.

[40] M. Elbattah and O. Molloy, "Clustering-aided approach for predicting patient outcomes with application to elderly healthcare in Ireland," *AAAI Work. - Tech. Rep.*, vol. WS-17-01-, pp. 533–541, 2017.

[41] K. L. and R. P., "Clustering by means of Medoids," *Statistical Data Analysis Based on the L1 Norm and Related Methods*. pp. 405–416, 1987.

[42] E. Irwansyah, E. S. Pratama, and M. Ohyver, "Clustering of Cardiovascular Disease Patients Using Data Mining Techniques with Principal Component Analysis and K-Medoids," *Preprints*, no. August, 2020.

[43] H. Li and D. Phung, "Journal of Machine Learning Research: Preface," *J. Mach. Learn. Res.*, vol. 39, no. 2014, pp. i–ii, 2014.

[44] V. Acharya and P. Kumar, "Identification and red blood cell automated counting from blood smear images using computer-aided system.," *Med. Biol. Eng. Comput.*, vol. 56, no. 3, pp. 483–489, Mar. 2018.

[45] G. Li, L. Guo, and T. Liu, "Grouping of Brain MR Images via Affinity Propagation," *Conf. Proc. (Midwest. Symp. Circuits Syst).*, vol. 2009, pp. 2425–2428, May 2009.

[46] A. Busch, T. Homeier-Bachmann, M. Y. Abdel-Glil, A. Hackbart, H. Hotzel, and H. Tomaso, "Using affinity propagation clustering for identifying bacterial clades and subclades with whole-genome sequences of Francisella tularensis," *PLoS Negl. Trop. Dis.*, vol. 14, no. 9, p. e0008018, Sep. 2020.

[47] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," *Res. Issues Data Min. Knowl. Discov.*, pp. 1–8, 1997.

[48] S. A. Sajidha, S. P. Chodnekar, and K. Desikan, "Initial seed selection for K-modes clustering – A distance and density based approach," *J. King Saud Univ. - Comput. Inf. Sci.*, 2018.

[49] D. M. A. Hussain, A. Q. K. Rajput, B. S. Chowdhry, and Q. H. Gee, *Communications in Computer and Information Science: Preface*, vol. 20 CCIS. 2008.

[50] N. Iam-On, T. Boongoen, and S. Garrett, "LCE: A link-based cluster ensemble method for improved gene expression data analysis," *Bioinformatics*, vol. 26, no. 12, pp. 1513–1519, 2010.

[51] M. Kennedy, "Virtue and virtuality: Technoethics, IT and the masters of the future," *Moral, Ethical, Soc. Dilemmas Age Technol. Theor. Pract.*, vol. 42, no. 10, pp. 1–18, 2013.

[52] C. Vidden, M. Vriens, and S. Chen, "Comparing clustering methods for market segmentation: A simulation study," *Appl. Mark. Anal.*, vol. 2, no. 3, pp. 225–238, 2016.

[53] H. G. Ayad and M. S. Kamel, "On voting-based consensus of cluster ensembles," *Pattern Recognit.*, vol. 43, no. 5, pp. 1943–1953, 2010.

[54] H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 160–173, 2008.

[55] R. Ghaemi, N. Sulaiman, H. Ibrahim, and N. Mustapha, "A survey: Clustering ensembles techniques," *World Acad. Sci. Eng. Technol.*, vol. 38, pp. 644–653, 2009.

[56] Y. Şenbabaoğlu, G. Michailidis, and J. Z. Li, "Critical limitations of consensus clustering in class discovery," *Sci. Rep.*, vol. 4, 2014.

[57] M. Hassani and T. Seidl, "Using internal evaluation measures to validate the quality of diverse stream clustering algorithms," *Vietnam J. Comput. Sci.*, vol. 4, no. 3, pp. 171–183, 2017.

[58] J. O. Palacio-Niño and F. Berzal, "Evaluation Metrics for Unsupervised Learning Algorithms," *arXiv*, 2019.

[59] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[60] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in Neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, 2014.

[61] J. Han, M. Kamber, and J. Pei, "Third Edition : Data Mining Concepts and Techniques," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2012.

[62] S. El Ferchichi, S. Zidi, K. Laabidi, M. Ksouri, and S. Maouche, "A new feature extraction method based on clustering for face recognition," *IFIP Adv. Inf. Commun. Technol.*, vol. 363 AICT, no. PART 1, pp. 247–253, 2011.

[63] A. Gracia, S. González, V. Robles, and E. Menasalvas, "A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality," *Inf. Sci. (Ny).*, vol. 270, pp. 1–27, 2014.

[64] G. Kraemer, M. Reichstein, and M. D. Mahecha, "dimRed and coRanking-

unifying dimensionality reduction in R," *R J.*, vol. 10, no. 1, pp. 342–358, 2018.

[65] K. Siwek, S. Osowski, T. Markiewicz, and J. Korytkowski, "Analysis of medical data using dimensionality reduction techniques," *Prz. Elektrotechniczny*, vol. 89, no. 2 A, pp. 279–281, 2013.

[66] E. Postma, "Dimensionality Reduction : A Comparative Review Dimensionality Reduction : A Comparative Review," no. October 2016, pp. 1–35, 2007.

[67] J. Clark and F. Provost, "Unsupervised dimensionality reduction versus supervised regularization for classification from sparse data," *Data Min. Knowl. Discov.*, vol. 33, no. 4, pp. 871–916, 2019.

[68] Z. Zhang and A. Castelló, "Principal components analysis in clinical studies.," *Annals of translational medicine*, vol. 5, no. 17. p. 351, Sep-2017.

[69] L. A. Dawson, M. Biersack, G. Lockwood, A. Eisbruch, T. S. Lawrence, and R. K. Ten Haken, "Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation.," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 62, no. 3, pp. 829–837, Jul. 2005.

[70] Y. He, R. Wang, H. Meng, L. Li, Z. Wu, and Y. Dong, "Establishment of a PCA model for skin health evaluation," *Biotechnol. Biotechnol. Equip.*, vol. 32, no. 4, pp. 1060–1064, Jul. 2018.

[71] R. Pandit and A. Shehu, "A principled comparative analysis of dimensionality reduction techniques on protein structure decoy data," *Proc. 8th Int. Conf. Bioinforma. Comput. Biol. BICOB 2016*, pp. 43–48, 2016.

[72] S. Weng, C. Zhang, Z. Lin, and X. Zhang, "Mining the structural knowledge of high-dimensional medical data using Isomap," *Med. Biol. Eng. Comput.*, vol. 43, no. 3, pp. 410–412, 2005.

[73] S. K. Prabhakar and H. Rajaguru, "Comparison of Isomap and matrix factorization with mahalanobis based sparse representation classifier for epilepsy classification from EEG signals," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 580–583.

[74] K. Raza and N. K. Singh, "A Tour of Unsupervised Deep Learning for Medical Image Analysis," pp. 1–29, 2018.

[75] A. M. Karim, M. S. Güzel, M. R. Tolun, H. Kaya, and F. V. Çelebi, "A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing," *Biocybern. Biomed. Eng.*, vol. 39, no. 1, pp. 148–159, 2019.

[76] H. S. S. Lee and D. Shen, "Latent feature representation with stacked auto-encoder for AD / MCI Latent feature representation with stacked auto-encoder for

AD / MCI diagnosis," no. December 2015, 2013.

[77] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7–9, pp. 1431–1443, 2009.

[78] A. Griparis, D. Faur, and M. Datcu, "A dimensionality reduction approach to support visual data mining: Co-ranking-based evaluation," in *2016 International Conference on Communications (COMM)*, 2016, pp. 391–394.

[79] Jan de Leeuw, "Modern Multidimensional Scaling: Theory and Applications (Second Edi- tion)," vol. 62, no. 9, 2014.

[80] J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen, "Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure," *Neurocomputing*, vol. 169, pp. 246–261, 2015.

[81] L. Chen and A. Buja, "Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis," *J. Am. Stat. Assoc.*, vol. 104, no. 485, pp. 209–219, 2009.

[82] N. Dey, S. Borra, A. S. Ashour, and F. Shi, *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*. San Diego, UNITED STATES: Elsevier Science & Technology, 2018.

[83] V. S. Sumithra and S. Surendran, "A Review of Various Linear and Non Linear Dimensionality Reduction Techniques," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 3, pp. 2354–2360, 2015.

[84] M. Pechenizkiy, A. Tsymbal, and S. Puuronen, "PCA-based feature transformation for classification: Issues in medical diagnostics," *Proc. IEEE Symp. Comput. Med. Syst.*, vol. 17, pp. 535–540, 2004.

[85] A. Sengur, "An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular heart diseases," *Comput. Biol. Med.*, vol. 38, no. 3, pp. 329–338, 2008.

[86] S. A. Zu'bi, N. Islam, and M. Abbod, "3D Multiresolution Analysis for reduced features segmentation of medical volumes using PCA," in *2010 IEEE Asia Pacific Conference on Circuits and Systems*, 2010, pp. 604–607.

[87] M. H. Yang, "Face recognition using extended isomap," *IEEE Int. Conf. Image Process.*, vol. 2, pp. 117–120, 2002.

[88] O. Samko, A. D. Marshall, and P. L. Rosin, "Selection of the optimal parameter value for the Isomap algorithm," *Pattern Recognit. Lett.*, vol. 27, no. 9, pp. 968–979, 2006.

[89] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2

mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation," *Neurocomputing*, vol. 112, pp. 92–108, 2013.

[90] D. S. Chiu and A. Talhouk, "DiceR: An R package for class discovery using an ensemble driven approach," *BMC Bioinformatics*, vol. 19, no. 1, pp. 17–20, 2018.

[91] S. K. Uppada, "Centroid Based Clustering Algorithms- A Clarion Study," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 6, pp. 7309–7313, 2014.

[92] U. Bodenhofer, A. Kothmeier, and S. Hochreiter, "Apcluster: An R package for affinity propagation clustering," *Bioinformatics*, vol. 27, no. 17, pp. 2463–2464, 2011.

[93] L. van der Maaten and G. Hinton, "Viualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[94] H. Fanaee-T and M. Thoresen, "Performance evaluation of methods for integrative dimension reduction," *Inf. Sci. (Ny).*, vol. 493, pp. 105–119, 2019.

[95] T. C. Matrix and A. G. Kraemer, "Package ' coRanking ,'" 2020.

[96] G. Brock, V. Pihur, S. Datta, and S. Datta, "ClValid: An R package for cluster validation," *J. Stat. Softw.*, vol. 25, no. 4, pp. 1–22, 2008.

[97] L. Hubert and P. Arabie, "Comparing partitions," *J. Classif.*, vol. 2, no. 1, pp. 193–218, 1985.

[98] V. Menger, M. Spruit, W. van der Klift, and F. Scheepers, "Using Cluster Ensembles to Identify Psychiatric Patient Subgroups BT - Artificial Intelligence in Medicine," 2019, pp. 252–262.