TILBURG UNIVERSITY

CLUSTER ANALYSIS OF CPAP PATIENT DATA USING SELF-SUPERVISED LEARNING

SHAO-WEN CHIU

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

WORD COUNT: 8684

STUDENT NUMBER

2076655 (u803296)

COMMITTEE

Dr. Marijn van Wingerden
Dr. Samaneh Khoshrou
Niek den Teuling, MSc

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2022

ACKNOWLEDGMENTS

I appreciate my university supervisor, Dr. Marijn van Wingerden, for all the support and guidance through this thesis. I greatly thank him for his patience and knowledgeable feedback during my thesis research. I also want to express my appreciation to my company supervisors, Prof. Steffen Pauws and Niek den Teuling, from Philips Research, for giving me the opportunity to work on this meaningful project. I would like to especially thank Niek for always encouraging me with insightful guidance and feedback. I want to extend my thanks to all my friends, who always makes me laugh and remain hopeful. And finally, I want to thank my family, my parents, Yung-Ling Chiu, Huang-Jung Chiu, and younger brother, Shao-Yang Chiu, for always being on my side. My deepest thank is to my beloved husband, Yen-Chia Hsu, for helping me overcome my stress, for always being my personal tutor, and for always supporting and loving me unconditionally. I could not have done this without him.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CLUSTER ANALYSIS OF CPAP PATIENT DATA USING SELF-SUPERVISED LEARNING

SHAO-WEN CHIU

Abstract

Predicting and understanding patients' subgroups is critical in helping clinicians, and device providers make early decisions about the patients' treatment plans for CPAP patients. Currently, researchers have been using clustering techniques to explore CPAP patients adherence subgroups. However, it is still unclear if we can use the clustering techniques to effectively identify patients' AHI and mask leakage subgroups. This thesis uses a self-supervised learning approach by first identifying both AHI and LargeLeakPct subgroups with unsupervised learning techniques. Then, supervised learning techniques are utilized to inspect the performance when predicting the patient subgroups. In the experiments, we implement six univariate clustering methods to extract the cluster memberships. We also implement multinomial Logistic Regression and Random Forest to predict their cluster memberships at an early stage. Results show that cluster memberships obtained from GCKM with the optimal number of clusters with Random Forest outperform other pipelines for both AHI and LargeLeakPct. The result also indicated that it is possible to identify underlying patterns of CPAP patients subgroups, furthermore, predict therapy subgroups as early as four weeks with high performance.

Keywords: longitudinal clustering, univariate clustering, obstructive sleep apnea, self-supervised learning

1 DATA SOURCE/CODE/ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data used in this thesis retains ownership of the data during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data. The code used in this thesis is not publicly available due to the confidentiality agreement with the external partner.

2 INTRODUCTION

This thesis aims to use clustering and self-supervised learning to explore patterns and identify patient subgroups in the device data obtained from Obstructive Sleep Apnea (OSA) patients on Continuous Positive Airway Pressure (CPAP) therapy. OSA has a 9% to 38% prevalence in the adult population and is a serious chronic disorder related to pauses in breathing while sleeping (Senaratna et al., 2017). Apnea-Hypopnea Index (AHI) is an indicator of the severity of the OSA disorder, which means how many apneas (temporary absence of breathing) occurred per hour over the sleep period. OSA is considered severe if the AHI is more than 30 during sleep (Quan, Gillin, Littner, & Shepard, 1999). Moreover, OSA can lead to side effects, including a higher risk of cardiovascular diseases and decreased overall quality of life (Kendzerska et al., 2014).

CPAP is an effective therapy for the OSA population (Kribbs et al., 2012), patients on this therapy wear a mask and hose that connects to a device for CPAP therapy. The device supplies compressed air to the patients while they are asleep, which prevents the occurrence of obstructive apneas, and can help to decrease the OSA symptoms. It is strongly suggested that the patients use the device daily in order to ensure effective treatment (Weaver et al., 1997), which can increase the patients' overall quality of life. Kribbs et al. (1993) indicated that the benefits of the CPAP treatment could drop significantly even with one non-attempt day. Therapy effectiveness may also be influenced by other factors. A common cause is poor mask fit, causing abnormal leakage for various reasons. It may influence the effectiveness of the CPAP therapy for the OSA population (Sopkova, Dorkova, & Tkacova, 2009).

CPAP therapy devices measure detailed daily data with many variables, including usage time, mask leakage (which we will refer to as LargeLeakPct below), AHI, and the air pressure supplied by the device. Despite such rich data, it remains a challenge to understand different OSA patient subgroups and their underlying patterns. Such information is critical for the clinicians and device providers to assist the patients in adhering to the CPAP therapy. With such information, we may also enable clinicians, device providers, and future researchers to understand how early we can predict the therapy patterns for this population.

Previously, Den Teuling, Pauws, and van den Heuvel (2020) has used K-means longitudinal (KmL) and two-step clustering (i.e. GCKM and LMKM) methods to cluster the longitudinal OSA patient data using adherence patterns. However, previous research did not inspect the AHI and LargeLeakPct variables. Also, it is unclear whether the recently developed models, such as High Dimensional Data Clustering (HDDC) (Bergé, Bouveyron, & Girard, 2012), and Deep Gaussian Mixture Modeling (Deepgmm) (Viroli & McLachlan, 2019), together with previously used models, can effectively identify patient subgroups using AHI and LargeLeakPct. Moreover, recent advance of machine learning, especially self-supervised learning techniques, provides a great opportunity to cluster OSA patient subgroups and examine if the clusters can be predicted accurately. With the self-supervised learning approach, it could be beneficial for clinicians and new patients by forecasting their therapy subgroups.

Therefore, this thesis aims to answer the following research question (RQ) that contains two sub-questions (SQ1 and SQ2):

RQ *Using a self-supervised approach, which supervised machine learning model can best predict cluster memberships derived from medical device data of OSA patients on CPAP therapy, using unsupervised longitudinal clustering?*

SQ1 *Can advanced clustering algorithms, like Deepgmm or HDDC, or statistical models, such as GCKM or LMKM outperform traditional ones, like KmL, for identifying patterns in different OSA patient groups, based on the Dunn index?*

SQ2 *Which classification algorithm, such as Random Forest or Logistic Regression, can be used to predict OSA patients' subgroups defined by the clustering algorithm with the highest accuracy?*

Results from the experiments show that the GCKM outperforms other clustering techniques for the AHI univariate clustering. Compared to other clustering techniques, the cluster memberships obtained from GCKM are the most compact (within cluster) and well-separated (among clusters). For the Large Leakage Percentage (LargeLeakPct) univariate clustering, the results indicated that LMKM outperforms others. When investigating the cluster differences, the result shows that the mean values of all the variables are significantly different across clusters that are created using the best model for both univariate clustering.

Regarding the task of classifying AHI and LargeLeakPct cluster memberships, we compared and evaluated the models based on accuracy and macro F1 score. We also analyzed the best two models for predicting AHI and LargeLeakPct cluster memberships with confusion matrix, accuracy, precision, recall, F1 score, and prevalence (i.e., percentage of data in the cluster). The results show that Random Forest with cluster memberships obtained from GCKM outperforms other combinations for both AHI and

LargeLeakPct. Moreover, we could obtain high accuracy macro F1, and good precision and recall at the early stage at the fourth week out of 13 weeks. This means that, for new patients during their CPAP treatment, it is possible to early predict their belonging subgroups with underlying patterns, which can help the clinicians and device providers intervene early if more assistance is needed. With the above information, we can conclude that using the self-supervised learning pipeline, we can predict CPAP patient subgroups (i.e., cluster memberships) with high performance. Thus, it is possible to use this pipeline to aid the patients, clinicians, and device providers.

3 RELATED WORK

This section explains each methods separately. The self-supervised learning approach remains under-explored in the CPAP patient population. In addition, there is a lack of research work related to univariate clustering for the CPAP patients population other than adherence patterns. Hence, this section focuses on the clustering methods and their application in the medical domain, including the evaluation metrics. The current research gaps and the research questions are also discussed.

3.1 *CPAP Therapy*

Self-supervised learning approach for CPAP patients is still an under-explored topic in this population. Previous research related to this population focused on how to cluster patients on adherence patterns accurately, and multivariate clustering to investigate the patterns among different subgroups.

Babbin, Velicer, Aloia, and Kushida (2015) identified four OSA adherence subgroups using dynamic clustering analysis with temporal features from longitudinal data. The four clusters including Great Users (n=22), Good Users (n=42), Low Users (n=29), and Slow Decliners (n=33). The temporal features included parameters with mean, level, and slope. The paper reported that there were significantly different patterns among different subgroups and similar patterns with-in groups, based on chi-square and MANOVAs analysis. The paper concluded that it could be meaningful for future studies to understand both individual and subgroup patterns for the patients.

Kim et al. (2021) used the regular k-means algorithm to categorize and identify the phenotype of OSA patients by integrating craniofacial risks with BMI, apnea severity, symptoms, and comorbidity. This research identified three phenotypic subgroups, including noncraniofacial phenotype

(39% patients), craniofacial skeletal phenotype (33% patients), and complex phenotype (29% patients). The three phenotypes differed mainly from BMI, OSA severity, and skeletal discrepancy, which could provide basic patterns for clinicians' decision-making process, such as alternative intervention or determining the appropriate goal for different clusters' patients. They also performed the multivariable linear regression analysis within each cluster, which found that the contributing factors for the OSA severity were observed to be different between clusters.

Den Teuling, van den Heuvel, Aloia, and Pauws (2021) utilized a "heteroskedastic hurdle growth mixture modeling" approach with generalized additive modeling to identify the different adherence patterns for OSA patients on CPAP therapy. As a result, nine group trajectories have been identified, and the findings of the different OSA patient subgroups patterns are worth exploring, including: first; the residual AHI is higher in non-adherence groups; second, the early drop-out group has the highest mask leakage variability; in contrast, the more adherent group has the lowest mask leakage variability; third, the lowest possible device pressure of 5 cmH2O was found to be of a higher rate in drop-out groups.

While these work provides insights about how to model CPAP therapy data, they are either based on simulated samples or small datasets. It remains an open challenges about whether these approaches work on real and large datasets, which is the case of this research. Moreover, this paper has a different set of target variables for patients, which have not been explored in previous works.

3.2 *Self-Supervised Learning*

Self-supervised learning is a machine learning pipeline which is commonly used on unlabelled data. In recent years, self-supervised learning has been recognized as an effective method for various fields, including computer vision and speech recognition (Zhai, Oliver, Kolesnikov, & Beyer, 2019). This pipeline employs the unlabelled data to obtain pseudo-labels from unsupervised learning (i.e., clustering).

However, medical data can be challenging to label manually because it typically needs domain experts in order to perform it accurately. Therefore, unsupervised learning can play an essential role in clustering the patients into subgroups, making pseudo-label, and enabling the following task. In this way, supervised learning (i.e., regression or classification) or unsupervised learning tasks can carry out the desired outcome (Doersch & Zisserman, 2017). The self-supervised learning is compelling due to its ability to learn insight from unlabelled data, which is common in medical fields (Zheng, Wang, Wang, & Liu, 2018).

To the best of our knowledge, this research is the first that inspects the potential of adopting the self-supervised learning pipeline to a large CPAP therapy dataset, specifically on the LargeLeakPct and AHI variables.

3.3 *Longitudinal Clustering*

Nowadays, there is a growing capability to collect large datasets, including many repeated measurements per subject (e.g., patients) over time, which is called intensive longitudinal data (ILD) (Walls & Schafer, 2006). Longitudinal clustering enables researchers to explore subgroup patterns' changing over time and the variability within and between subgroups. For instance, Den Teuling et al. (2020) explored five clustering techniques for longitudinal data on simulated datasets by comparing the actual group membership and the clustered group membership based on mean Normalized Split-Join (NSJ) scores, which is a metric to compare classification agreement. The paper indicated that the Growth Mixture Modeling (GMM) and two-step clustering (GCKM) have significantly better performance. The authors suggested GCKM over GMM when there is a limitation of computation time due to similar results compared to GMM. The researchers in another paper recommended K-means longitudinal (KmL) for its flexibility in describing the subgroups' trajectories, computational efficiency, and relatively favorable computational scaling (Teuling, Pauws, & Heuvel, 2021). Moreover, Genolini et al. (2013) introduced the KmL3D package, enabling the user to utilize multiple variables to cluster the joint trajectories, with the co-evolution taking account. With a real dataset, the researchers found that classic KmL and Kml3D could capture different clusters, with only 49.45% results matching each other (Jaccard similarity = 0.25), and tend to suggest a lower number of clusters than KmL.

Additionally, it can be challenging for longitudinal clustering when dealing with high dimensionality problems due to the nature of temporal features. To address this challenge, Bouveyron, Girard, and Schmid (2007) proposed the High Dimensional Data Clustering (HDDC) method based on the Gaussian Mixture Model (GMM) using subspace clustering and parsimonious modeling with Expectation-Maximization (EM) algorithm. In comparing different GMM model-based clustering techniques and HDDC, they found that HDDC outperforms Sphe-GMM and Vs-GMM models with a cluster recognition rate of 0.95. Bergé, Bouveyron, and Girard (2012) compared HDDC model with GMM and GMM with a variable selection approach model (clustvarsel) proposed by Scrucca and Raftery (2018). Results show that the HDDC outperformed the other methods, with a Correct Classification Rate (CCR) = 0.945. GMM obtained CCR = 0.575 and clustvarsel with CCR = 0.925.

In recent years, deep learning models received extensive attention for their capability in supervised learning (LeCun, Bengio, & Hinton, 2015). There are comparatively not many clustering techniques implemented deep learning concepts. An example is seen in the work of Viroli and McLachlan (2019), who presented the Deep Gaussian Mixture Model. The researchers implemented the Deepgmm model on various simulation datasets and compared the Deepgmm model with other clustering techniques, such as GMM, Skewed Normal Mixture Model (SNMM), Skewed-t Mixture Model (STMM), K-means, Partition around Medoids (PAM), and Ward's method (Hclust), based on Adjusted Rand Index (ARI) and misclassification rate. Results showed that Deepgmm outperformed all the simulation datasets with as high as 0.997 ARI and at the lowest as 0.002 misclassification rate. Deepgmm also outperformed other models even on the most challenging dataset related to silhouettes of vehicles.

Based on the insights from these previous works in longitudinal clustering, this thesis adopts KmL, KmL3D, two-steps clustering (i.e., GCKM and LMKM), HDDC, and Deepgmm techniques to cluster CPAP therapy data (with LargeLeakPct and AHI variables) and compare their performance.

3.4 *Optimal Number of Clusters*

Most clustering methods require specifying the number of clusters in advance. It is up to the researcher to determine the most suitable number of clusters. Unlike supervised learning tasks, it is difficult to know the optimal number of clusters due to the unknown underlying structure in real-world data. For Mixture Models, Bayesian Information Criterion (BIC) is a well-known technique to determine the optimal number of clusters (Chen & Gopalakrishnan, 1998), where a lower BIC value is preferred. The BIC value can decrease in a real dataset when there are more groups. However, more groups (i.e., clusters) might not always give the researchers more information. Therefore, a commonly used method, called the elbow method, can assist the researchers in determining the best number of clusters (Shi et al., 2021). The elbow method enables the researcher to examine the relative improvement of the BIC for each number of clusters. Usually, the BIC value will decrease when the number of groups increases. However, the visual plot of the BIC value often shows a turning point, indicating that the improvement is declining. The turning point is where the optimal cluster number lies based on this method (Hardy, 1994).

To assess the quality of clustering results, Dunn Index (Dunn, 1974) is one of the most common evaluation metrics. For example, Lynch and DeGruttola (2022) used a univariate ensemble clustering method to measure disease progression for HIV biomarkers. The researchers compared the Viral Load (VL), and CD4 cell counts clusters based on Dunn Index to see the group validity. The results showed that VL with three clusters scored 0.032 and CD4 with three clusters scored 0.024. Hence, the authors concluded that the two biomarkers performed similar result. Another work by Sobisek, Stachova, and Fojtik (2018) proposed a feature-based novel clustering method (called CluMP) and compared the algorithm with mixAK and KmL on micro-panel simulation datasets based on Dunn Index for three clusters. With experiments on the univariate clustering approach for unbalanced high noise dataset, unbalanced low noise dataset, balanced high noise data, and balanced low noise data, the Dunn index performance differed significantly. For instance, when using unbalanced high noise data, KmL achieved the highest Dunn Index 0.0863, and mixAK has the lowest Dunn Index 0.0404. On the other hand, with balanced low noise data, KmL was able to perform significantly better with Dunn Index 0.6608, and CluMP with Dunn Index 0.7064. The authors concluded that CluMP outperformed mixAK and KmL with balanced low noise data in the univariate clustering setting. The univariate clustering research observed comparatively low Dunn index in unbalanced data (for both low noise and high noise cases) and high-noise balanced data across all clustering methods.

Based on these previous works, this thesis applies BIC when choosing the optimal number of clusters within each clustering model. Then, Dunn Index is used when comparing different clustering models. Also, it is important to note that our CPAP data is unbalanced. Most patients throughout the therapy has a stable trend in AHI, and the LargeLeakPct remains stable with low value. Only a small number of patients encounter difficulties in the therapy. Moreover, our data contains high noise since there are many uncertainties in the therapy, such as the condition of the therapy machine, behaviors about how patients use the machine, individual differences among patients, etc. Thus, based on the insights from previous work, it is likely that Dunn Index would be low in our dataset.

3.5 *Connection of Research Questions to the Literature*

The literature review above shows that the self-supervised learning approach for CPAP patients' underlying therapy patterns is currently underexplored. Most of the research related to the CPAP patients was based on multivariate clustering or univariate clustering with usage time (adherence patterns).

To answer the research questions, this thesis uses and compares six clustering methods for SQ1. Each method's optimal cluster numbers are compared and evaluated based on Dunn Index. The KmL and KmL3D clustering methods are selected based on their wide usage and easy implementation with good scaling capability. GCKM and LMKM are selected based on the promising result of modeling the slowing change over time on the CPAP population. For the advanced GMM-based model, HDDC is selected due to its ability to capture clusters with high dimensional temporal features. Finally, Deepgmm is chosen because of its high performance on both simple and complex datasets.

With a self-supervised learning approach, each clustering method generates a set of labels based on their optimal cluster numbers. To answer SQ2, Logistic Regression and Random Forest are evaluated and compared with a different subset of temporal features from two weeks to 12 weeks as predictors, based on accuracy. These two classification algorithms are selected since they are widely used and easy to implement. The macro F1-score are computed to ensure the prediction quality for the unbalanced data observed. Since early prediction for the patients' subgroups can be beneficial for the patients, this thesis focuses on the possibility of early prediction with high accuracy.

4 METHOD

We address the research questions using a self-supervised learning pipeline. Longitudinal clustering methods are used to establish pseudo-labels for the follow-up prediction task. In other words, since the longitudinal dataset is unlabelled, we implement the unsupervised learning algorithms and make the generated cluster memberships the dependent variables for the supervised learning task. We use BIC with the visual elbow method to determine the optimal number of clusters for each variable. Moreover, univariate clustering results are evaluated and compared based on Dunn Index. The supervised learning step is implemented and compared based on accuracy and macro F1-score for each univariate clustering method for AHI and LargeLeakPct. First, we implement univariate clustering methods to obtain labels. Second, we use all the relevant device parameters as predictors to predict the clustering memberships. The methods in the pipeline are described below. Figure 1 demonstrated the detailed pipeline of this thesis.

4.1 *Unsupervised Learning*

4.1.1 *K-means clustering for Longitudinal Data (KmL)*

KmL is a widely used clustering method for longitudinal data based on traditional k-means (MacQueen et al., 1967). The user first identifies a

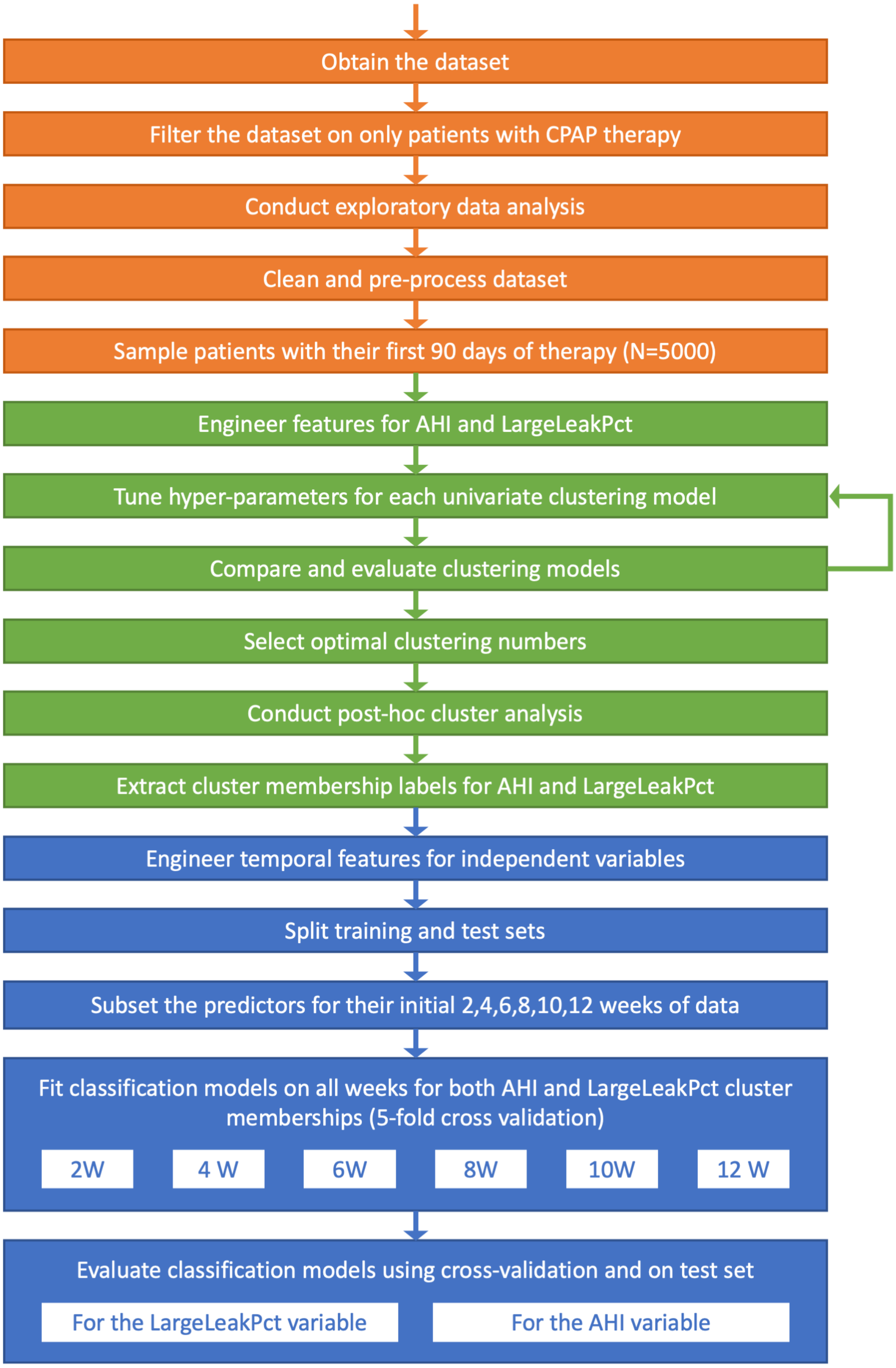

Obtain the dataset
Filter the dataset on only patients with CPAP therapy
Conduct exploratory data analysis
Clean and pre-process dataset
Sample patients with their first 90 days of therapy (N=5000)
Engineer features for AHI and LargeLeakPct
Tune hyper-parameters for each univariate clustering model
Compare and evaluate clustering models
Select optimal clustering numbers
Conduct post-hoc cluster analysis
Extract cluster membership labels for AHI and LargeLeakPct
Engineer temporal features for independent variables
Split training and test sets
Subset the predictors for their initial 2,4,6,8,10,12 weeks of data
Fit classification models on all weeks for both AHI and LargeLeakPct cluster memberships (5-fold cross validation)
2W
4 W
6W
8W
10W
12 W
Evaluate classification models using cross-validation and on test set
For the LargeLeakPct variable
For the AHI variable

Figure 1: The flow chart of the self-supervised learning pipeline

variable k, and the algorithm would find similar k clusters. K-means use the mean as a metric to determine the cluster, where the Euclidean distance is calculated in a given cluster with each data point. In this way, the data point will be assigned to a cluster based on the minimum Euclidean distance from the mean point. After one round ended without any datapoint left, a new mean is re-calculated and will be iterated several times until the "mean" stop changing. The criterion is to identify the subgroups with minimal within-cluster variance and maximum between-cluster variance. KmL utilizes the same approach; however, it uses vectors as the input and, therefore, can represent the trajectories for each cluster. In this thesis, both classic KmL and the KmL3D will be evaluated because they could provide different clustering results, which we described in the literature review. KmL will be computed based on raw features; on the other hand, we will implement KmL3D with temporal features.

4.1.2 *Two-Steps Clustering (GCKM, LMKM)*

This method models the trajectories first using a Growth Curve Model (GCM). In the second step, the k-means algorithm (MacQueen et al., 1967) is used to cluster the subject parameter estimates (i.e., the random effects). In other word, the Growth Curve Model is first trained with a mixed model and therefore, represents the longitudinal datasets as fixed effects. It also represents each subjects' deviation from the fixed effects, which referred to as random effects. Then, the k-means algorithm will be performed. The model was referred to as GCKM by Den Teuling et al. (2020) and was described previously by Twisk and Hoekstra (2012). With the GCKM approach, the longitudinal data characteristics will not be ignored. This thesis also utilizes a similar approach with linear regression instead of a growth curve model to model the trajectories. This method was referred to as LMKM in the latrend package by Den Teuling (2022).

4.1.3 *High Dimensional Data Clustering (HDDC)*

High Dimensional Data Clustering is a clustering model based on Gaussian Mixture Model and Bayesian techniques for high dimensional data. The idea of this clustering model is based on the assumption that the data can live in a lower-dimensional subspace and still be representative of the original space (Bergé et al., 2012). The method utilizes the subspace clustering algorithms based on Bouveyron et al. (2007), and the expectation-maximization (EM) algorithm is used for fitting the model parameters by maximizing the likelihood iteratively.

4.1.4 *Deep Gaussian Mixture Modeling (Deepgmm)*

Deep Gaussian Mixture Modeling (Deepgmm) is a model-based traditional Gaussian Mixture Model with multiple layers (Viroli & McLachlan, 2019). Multiple layers of learning can effectively handle the complex relationships between inputs. Deepgmm is therefore defined as a model with multiple layers with latent variables based on the deep neural network perspectives. Hence, the latent variables follow the mixture of Gaussian distributions. Due to the nature of the high dimensionality, dimensionality reduction is used at each layer to prevent an excessive number of parameters.

4.2 *Supervised Learning*

In this step, two supervised learning algorithms will be evaluated and compared to determine how early one can predict the cluster memberships created based on the univariate clustering in the previous step. This will be done using multinomial Logistic Regression and Random Forest, chosen due to their robustness and widely used in the classification research area.

4.2.1 *Multinomial Logistic Regression*

The traditional Logistic Regression was commonly used for the binary classification task. However, using the one versus all approach with multinomial Logistic Regression enables multi-class classification. The multinom function within the "nnet" package in R was utilized (Venables & Ripley, 2002). The "multinom" function in the nnet package fits multinomial Logistic Regression using the feed-forward neural networks, which enables the multinomial Logistic Regression.

4.3 *Evaluation Metrics*

4.3.1 *Dunn Index and Silhouette Score*

Dunn Index (DI) is a commonly used metric for clustering analysis. Dunn Index denotes the ratio of the minimum of inter-cluster distances and the maximum of intra-cluster distances. Therefore, the better clustering results, the higher the Dunn Index value. Below is the Dunn Index formula:

$$DI = \frac{min(\text{inter\_cluster\_distance})}{max(\text{intra\_cluster\_distance})} \tag{1}$$

In this thesis, we utilize Euclidean distance to calculate the inter and intra-cluster distance of Dunn Index.

Previous work showed that Dunn Index might be low with our dataset (see section 3.4). Thus, we also compute the Silhouette score to complement

Dunn Index and make our research more comprehensive. Silhouette score is a common evaluation metric for unsupervised learning. The score ranges from -1 to 1, where a value closer to 1 means that each cluster is more cohesive, and all clusters are better separated. In contrast, the clusters are not optimal if the Silhouette score is closer to -1 (Rousseeuw, 1987).

4.3.2 *Accuracy and Macro F1-scores*

The evaluation metric for the supervised learning task will be accuracy and F1 score. Due to the nature of multi-class classification and the imbalanced cluster memberships observed, we will use the macro F1-score to evaluate the result of the classification task. The macro F1-score computes each class' precision and recall and takes the average of their F1-score. In this way, smaller classes are considered equally important as the majority class.

5 EXPERIMENTAL SETUP

5.1 *Dataset Description*

The data of this thesis were obtained from OSA patients on CPAP therapy in the United States who are using the "Dream Mapper" application by Philips Respironics. The longitudinal dataset consisted of 54,310 CPAP patients' daily usage data over approximately two years, making it a total of 23,863,809 observations. The raw dataset has 45 variables. The most relevant variables include de-identified patient IDs, day of therapy, daily usage hours, residual AHI (which we will refer to as AHI below), average pressure, and LargeLeakPct. Table 1 shows the description of the relevant variables that are used in this thesis.

In this thesis, both AHI and LargeLeakPct will be used as the univariate clustering target, and in the supervised learning step, all the relevant variables will be transformed into temporal features as predictors. In addition, there are excessive missing values (11.6%) since not all patients used the therapy consistently every day. For example, some patients are more adhering to the therapy with few missing days, but some other patients might only attempt to take the therapy one or two days a week. Due to computational complexity, this thesis analyzes a random subset of 5,000 patients and their first 90 days of device data. Moreover, the dependent variables are selected according to domain knowledge, insights from exploratory data analysis, and literature reviews.

Table 1: CPAP machine variables descriptions

Variable Names	Description
Patient	De-identified patient ID
DayOfTherapy	The number of the day the patient has been on therapy
AHI	Number of apnea per hour
LargeLeakPct	Proportion of connected time in large leak (%)
UsageHours	The amount of time the patient was on therapy
LeakTotal	Average total leak (L/min)
PressureCpap	The average of the session-averaged CPAP pressure (cmH2O) with range [4, 20]

5.2 *Data Cleaning and Pre-processing*

During pre-processing, missing values of the usage time (i.e., the UsageHours variable) are replaced with zero, meaning there was no usage on that day. For device data other than the usage time, missing values due to no therapy attempts are replaced by the most recent non-missing value (e.g., not "NA") prior to the missing value using the Last Observation Carried Forward (LOCF) approach.

Outliers of the AHI variable can influence the quality of the analysis significantly. For example, low AHI values in short usage days are not reliable, and for the research purpose, any AHI value lower than 2.5 are not considered useful information. The patient is well-treated in both cases above. In addition, extremely high AHI values are also found unreasonable (i.e., AHI with value 40 means that the patients pause their breath 40 times in an hour). Hence, any value higher than 40 is replaced with 40, and any value lower than 2.5 is imputed as 2.5. These bounds are determined based on domain knowledge.

For clustering techniques such as KmL, LMKM, and GCKM, the AHI variable is log-transformed, which can ensure the variable approximately conform to normality. For cross-sectional clustering techniques such as KmL3D, HDDC, and Deepgmm, the input is transformed into temporal features as mean, slope, and standard deviation for each week. Also, the temporal features are transformed into Z-scores, which can make each feature equally important by scaling them.

For the classification step, and for each cluster memberships (i.e. target variable) generated per model, the temporal features of the slope, mean, and standard deviation for each week of relevant variables such as CPAP pressure, Total Leak, AHI, LargeLeakPct, and Usage Hours are selected as predictors. The temporal features are all computed as Z-scores across patients. In order to explore how early we can predict the cluster membership of the patients, we subset the features as their initial 2, 4, 6, 8, 10, and 12 weeks as predictors.

AHI and LargeLeakPct are particularly selected to perform univariate longitudinal clustering to identify the potential underlying patterns and subgroups. During exploratory data analysis, the AHI and LargeLeakPct variable are found to be positively correlated for some patients individually. But at the population level, the correlation is small (Pearson's $r = 0.12$), as discuss later in section 6.1. Moreover, the AHI and LargeLeakPct could significantly influence CPAP patients' quality of life. For example, AHI can indicate the therapy effectiveness, and LargeLeakPct is related to therapy comfort and the amount of time with effective treatment.

5.3 *Experimental Procedure*

5.3.1 *Unsupervised Learning*

AHI and LargeLeakPct variable are selected to perform the univariate clustering. The procedure of both univariate clustering is as follows. For the model that requires raw input data, such as KmL, LMKM and GCKM, the clusters are created first with a different number of clusters (i.e., $k = 2, 3, ..., 8$). Then, BIC values are compared and the most suitable number of clusters is determined using the visual elbow method.

For the kml3D model, the features are transformed to cl3d to perform with the kml3D function. The nbDrawing parameter is tuned using 5, 10, 15, and 20. The final setting is 10 since performance is not very different when increasing the nbDrawing number after 10.

Regarding the HDDC clustering method, different sets of hyperparameters are performed to explore the optimal cluster numbers. We tuned three hyperparameters in the model. The first one is the number of the cluster ($k = 1, 3, ..., 8$). Secondly, we examine the applied algorithm, which has three possibilities: Expectation-Maximisation ("EM"), Classification E-M ("CEM"), and Stochastic E-M ("SEM"). Thirdly, we investigate the initialization of model parameters, which has three choices: "k-means", "random", and "param". The final setup is to use "EM" as algorithm, "k-means" as initialization, and "AkjBkQkDk" as model.

For Deepgmm, a grid search on different hyperparameters is performed. Deepgmm has the option to adjust the hidden layers from 1 to 3. In this research, the number of hidden layers is set to one. With all the combinations of the other hyperparameters, BIC values are calculated, and the optimal number of clusters is selected. The final setting of the

Deepgmm model is: init="kmeans", r=5, iteration=50, eps=0.001 and init-est="factanal".

5.3.2 *Supervised Learning*

The cluster memberships generated in the unsupervised learning step for each combination of clustering models and variables are used as the dependent variable for supervised learning. The temporal features of relevant variables of their mean, standard deviation, and slope described previously are selected as independent variables. The dataset is split (using R language's createDataPartition function) into 80% of the training set and 20% testing set by patients. Then, 5-fold cross-validation is computed for Logistic Regression and Random Forest classifier, providing an unbiased estimation of the true error, which can be used when tuning the model to prevent overfitting. Logistic Regression and Random Forest are applied to explore how early we can predict the cluster memberships with high accuracy and macro F1-score. To achieve this goal, the independent variables are subset to their first 2, 4, 6, 8, 10, and 12 weeks as predictors for both classification models.

5.4 *Programming Language and Packages*

This thesis uses the R language (version 4.0.5) for experiments. For univariate clustering, the implementation and evaluation of KmL (Genolini, Alacoque, Sentenac, & Arnaud, 2015a), LMKM, and GCKM applies the latrend package with version 1.2.3 (Den Teuling, 2022). HDDC clustering uses the HDclassif package with version 2.2.0 (Bergé et al., 2012). Deepgmm clustering uses the Deepgmm package with version 0.1.62 (Viroli & McLachlan, 2020). The kml3d package with version 2.4.2 is used for kml3D clustering (Genolini, Alacoque, Sentenac, & Arnaud, 2015b). For evaluating HDDC, Deepgmm, and kml3D clustering, the clValid package with version 0.7 is used (Brock, Pihur, Datta, & Datta, 2008). For supervised learning, the caret package with version 6.0.88 is used for Random Forest and evaluation (Kuhn, 2021). The nnet package with version 7.3.16 is utilized for the multiclass Logistic Regression (Venables & Ripley, 2002).

6 RESULTS

This section firstly presents the key results of the exploratory data analysis. Secondly, we present the optimal number of clusters and the results for each univariate clustering for AHI and LargeLeakPct across different clustering methods. Finally, a comparison of the different supervised learning methods results is presented.

6.1 *Exploratory Data Analysis*

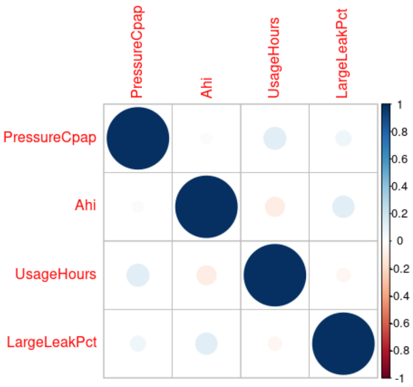


Figure 2: Population level correlation plot

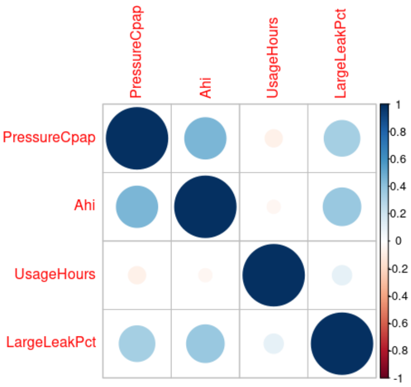


Figure 3: One example of patient level correlation plot

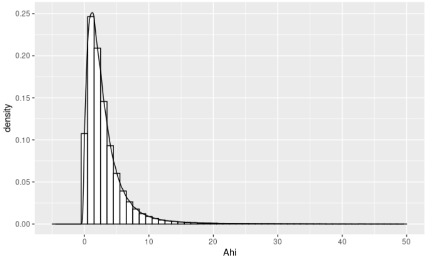


Figure 4: AHI density plot for the first 90 days

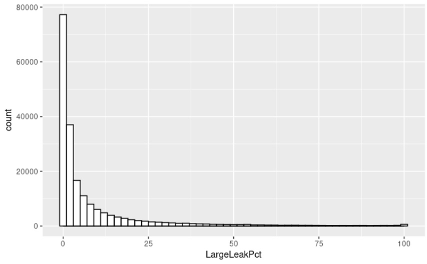


Figure 5: LargeLeakPct distribution excluding zero value (90 days)

Figure 2 shows the population level correlation plot on relevant variables. It is observed that the usage hours and AHI are slightly and negatively correlated (Pearson's $r = -0.1$). On the other hand, AHI and LargeLeakPct are slightly and positively correlated with each other (Pearson's $r = 0.12$) on population level. Figure 3 shows one example of the patient level correlation plot. We found that LargeLeakPct is significantly correlated with AHI positively (Pearson's $r = 0.38$). Also, the Pressure is positively correlated with LargeLeakPct (Pearson's $r = 0.34$) and AHI (Pearson's $r = 0.45$). Figure 4 shows the AHI variable density plot, and Figure 5 shows the distribution of LargeLeakPct distribution exclude zero value. It is noted that 57% of the LargeLeakPct values are zero (486,345 out of 841,197 observation).

Table 2: Optimal number of clusters (k^*), Dunn Index, and Silhouette score

Model	AHI			LargeLeakPct		
	k^*	Dunn Index	Silhouette	k^*	Dunn Index	Silhouette
KmL	4	0.045	0.22	3	0.063	0.65
KmL3D	3	0.047	0.46	3	0.068	0.63
LMKM	5	0.062	0.16	3	0.071	0.73
GCKM	3	0.066	0.25	3	0.054	0.64
HDDC	3	0.002	-0.04	3	0.00002	-0.25
Deepgmm	3	0.023	0.34	3	0.013	0.55

6.2 *Results from Unsupervised Learning*

Table 2 shows the optimal cluster numbers based on the lowest BIC values. The Silhouette score and Dunn Index for each clustering method are also presented. Take KmL for AHI univariate clustering as an example. Figure 6 shows that the 8-cluster approach has the lowest BIC value, which indicates the best model fit. However, in our case, we apply the visual elbow technique to select 4 as the optimal number of clusters.

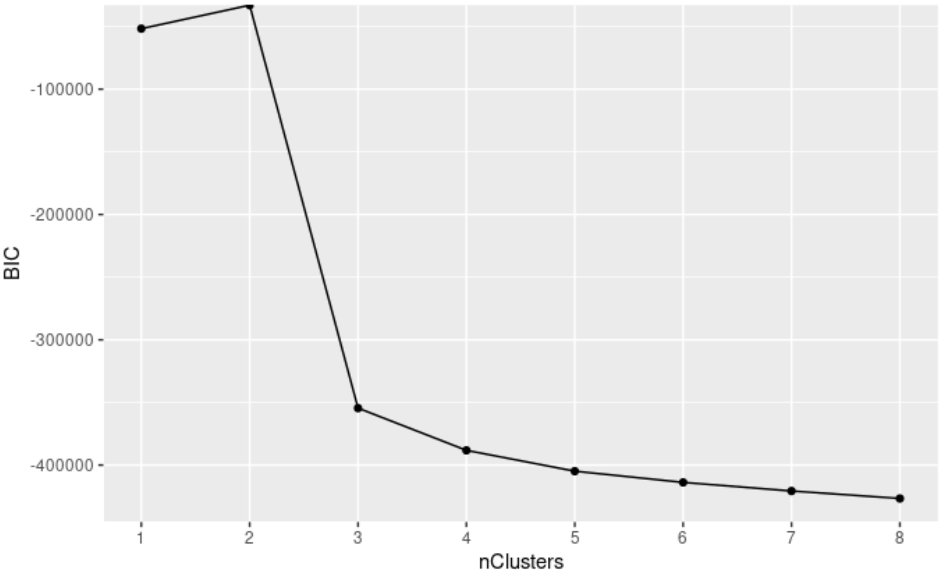


Figure 6: KmL Bayesian Information Criterion for AHI

Table 3: Cluster mean per variable obtained from AHI GCKM clustering

Variable Names	Cluster 1 (N=1405, 28.1%)	Cluster 2 (N=3077, 61.54%)	Cluster 3 (N=518, 10.36%)	F-statistic
AHI	5.05	2.96	10.39	3500***
UsageHours	5.58	5.84	5.33	20.06***
PressureCpap	8.29	8.36	8.71	6.111**
LargeLeakPct	2.93	1.73	5.17	71.93***
LeakTotal	34.88	33.76	39.51	52.91***

*p<.05 **p<.01 ***p<.001

6.2.1 *AHI Clustering*

Results of AHI clustering show that GCKM with 3 clusters outperforms other methods with Dunn Index 0.066. Table 3 contains the number of patients and their percentage per cluster with the GCKM univariate clustering for AHI. Figure 7 shows the raw trajectories and Figure 8 relative trajectories for each cluster memberships for GCKM AHI clustering.

Table 3 shows the mean device parameters per cluster memberships obtained from GCKM. Patients from cluster 2 have the lowest on AHI, LargeLeakPct, and LeakTotal, with the highest UsageHours. In contrast, cluster 3 exhibited the highest AHI, PressureCpap, LargeLeakPct, and LeakTotal, but with the lowest UsageHours. Results in Table 3 also included the ANOVA test, which described the variance of means for each variable between the different clusters with statistically significant differences. Results show that for the subgroups retrieved from GCKM for AHI univariate clustering, the supplied pressure significantly differs from each cluster (p-value < 0.01). At the same time, all other variables, including AHI, UsageHours, LargeLeakPct, and LeakTotal, are significantly different across the clusters (p-values < 0.001).

6.2.2 *LargeLeakPct Clustering*

Results from Table 2 show that LMKM with 3 clusters outperforms other methods with Dunn Index 0.071 and Silhouette score 0.73 for LargeLeakPct univariate clustering. Figure 9 shows the raw trajectories for each cluster memberships for LMKM LargeLeakPct clustering. Also, Figure 10 presents the relative trajectories.

Table 4 presents the distribution and the device parameters mean per cluster memberships retrieved from LMKM for LargeLeakPct. ANOVA test is also included in Table 4, indicating that each variable among different cluster memberships differs from each other significantly (p-value < 0.01).

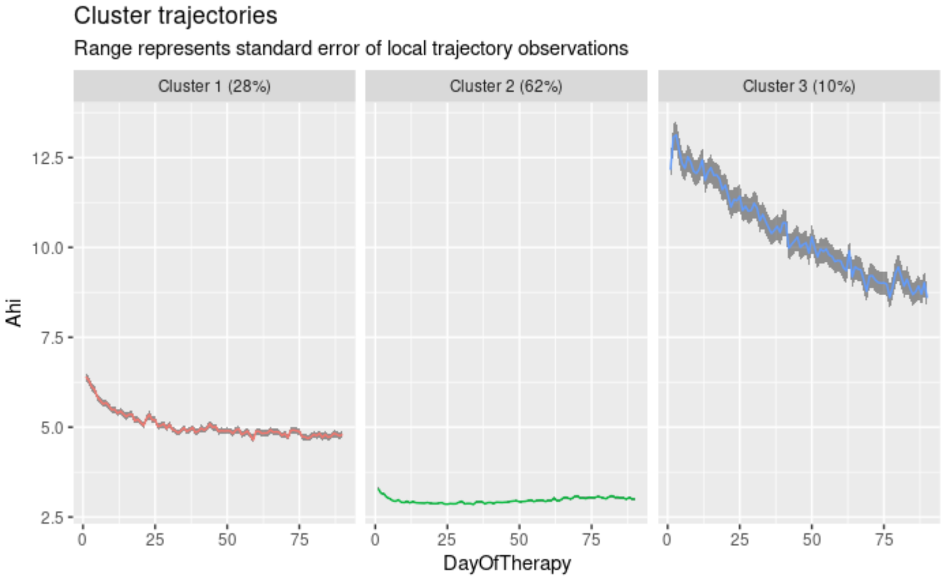


Figure 7: GCKM raw trajectories for AHI (higher value means worse)

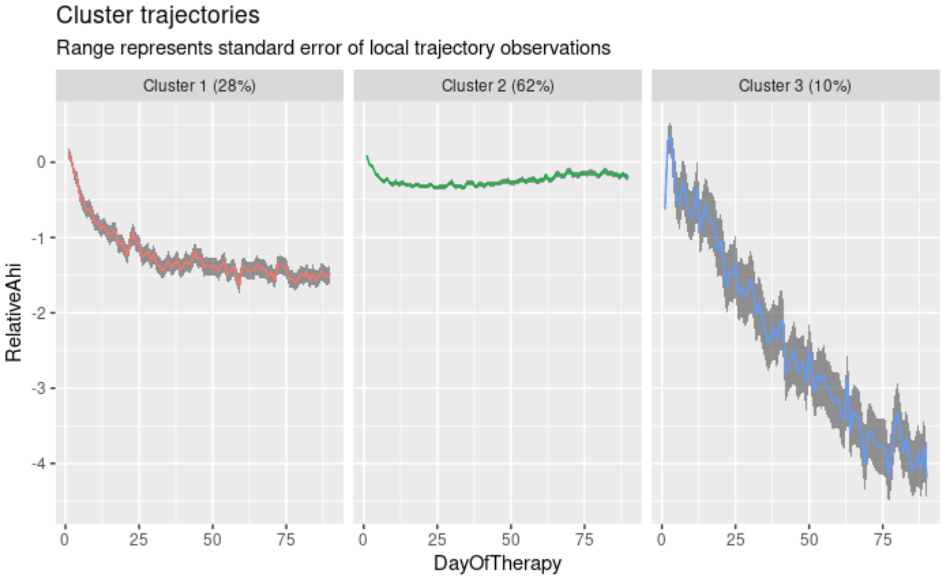


Figure 8: GCKM relative trajectories for AHI

Patients from cluster 2 have the lowest AHI, PressureCpap, LargeLeakPct, and LeakTotal among the clusters, with the highest UsageHours. On the other hand, patients from cluster 1 and 3 have higher LargeLeakPct, LeakTotal, and AHI. However, their UsageHours and PressureCpap are different from each other. Patients from cluster 1 have significantly lower

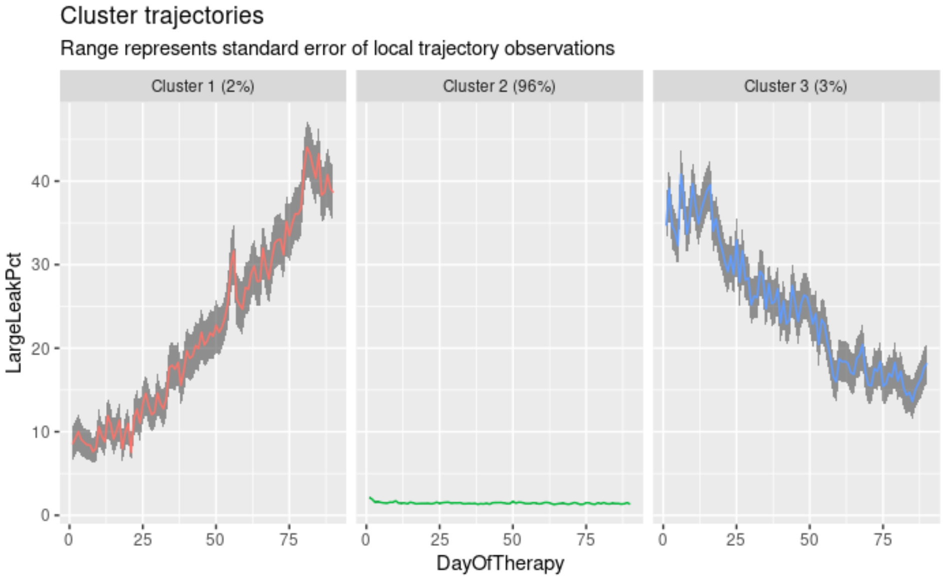


Figure 9: LMKM raw trajectories for LargeLeakPct (higher value means worse)

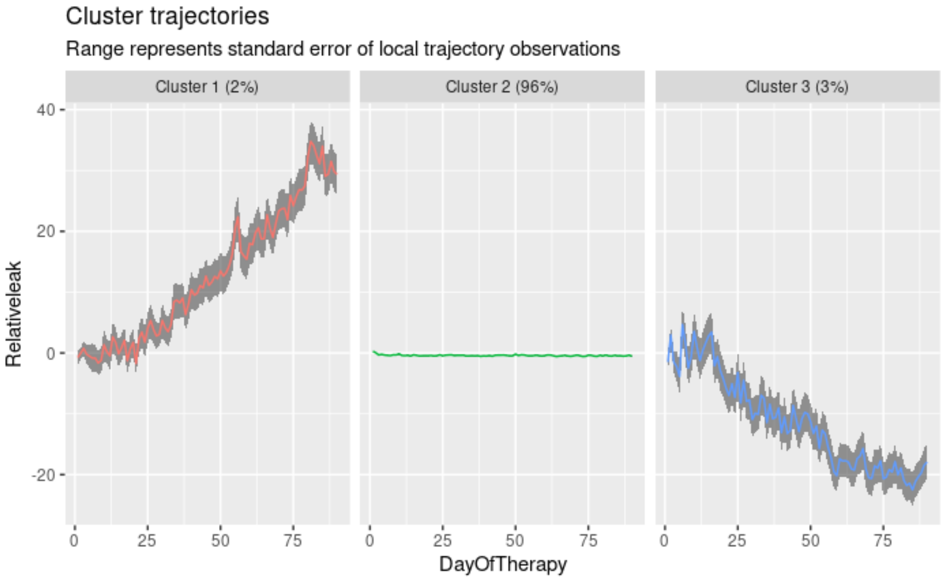


Figure 10: LMKM relative trajectories for LargeLeakPct

UsageHours and lower PressureCpap compare to cluster 3. On the other hand, the patients from cluster 3 have higher UsageHours comparing to cluster 1 and higher PressureCpap comparing to cluster 1 and cluster 2.

Table 4: Cluster mean per variable obtained from LargeLeakPct LMKM clustering

Variable Names	Cluster 1 (N=87, 1.74%)	Cluster 2 (N=4785, 95.7%)	Cluster 3 (N=128, 2.56%)	F-statistic
AHI	6.83	4.20	6.85	84.37***
UsageHours	4.55	5.76	5.23	20.59***
PressureCpap	8.36	8.36	9.07	5.63**
LargeLeakPct	22.18	1.46	24.90	2498***
LeakTotal	54.66	33.64	59.77	510.5***

*p<.05 **p<.01 ***p<.001

Table 5: Accuracy (mean ± standard error) for Logistic Regression (LR) and Random Forest (RF) using 5-fold cross-validation for AHI with 4 weeks of data as predictors

	KmL	KmL3D	LMKM	GCKM	HDDC	Deepgmm
LR	0.83 ±0.01	0.91 ±0.008	0.75 ±0.007	0.935 ±0.011	0.79 ±0.003	0.87 ±0.006
RF	0.84 ±0.018	0.92 ±0.01	0.76 ±0.015	0.938 ±0.007	0.80 ±0.012	0.87 ±0.009

6.3 *Results from Supervised learning*

6.3.1 *Classification for AHI*

Appendix A (page 38) visualizes the classification results for AHI clusters memberships from all the clustering methods with Logistic Regression and Random Forest classifier across different weeks. The results show that the Random Forest classifier with AHI cluster memberships obtained from GCKM outperforms other combinations based on accuracy and macro F1-score. Moreover, the result shows that Random Forest with GCKM AHI with high performance (accuracy = 93.8% and macro F1-score = 90.9%) as early as four weeks. The 5-fold cross-validation results are shown in Table 5 and Table 6. Table 11 shows the accuracy and macro F1-scores on test set of the two best models (GCKM and KmL3D with Random Forest, selected by cross-validation). Appendix E (page 46) shows the confusion matrix from Random Forest for GCKM and KmL3D labels on test set. The above information shows that the results can be generalized to test set with Random Forest. Table 8 displays the statistics per cluster for the best two models, including the prevalence, accuracy, precision, recall and F1.

Table 6: Macro F1-scores for Logistic Regression (LR) and Random Forest (RF) using 5-fold cross-validation for AHI with 4 weeks of data as predictors

	KmL	KmL3D	LMKM	GCKM	HDDC	Deepgmm
LR	0.73	0.75	0.52	0.906	0.78	0.77
RF	0.77	0.79	0.11	0.909	0.79	0.77

Table 7: Accuracy and macro F1-score for AHI GCKM and KmL3D on test set with Random Forest with 4 weeks of data

	Accuracy	Macro F1
GCKM	0.93	0.90
KmL3D	0.92	0.80

Table 8: Statistics per cluster (denoted as C_i, where i is the cluster number) for the best two models of AHI using 4 weeks of data

	Random Forest with GCKM			Random Forest with KmL3D		
Statistics	C_1	C_2	C_3	C_1	C_2	C_3
Prevalence	0.29	0.62	0.09	0.78	0.20	0.02
Accuracy	0.89	0.97	0.81	0.97	0.79	0.54
Precision	0.86	0.96	0.91	0.95	0.80	0.81
Recall	0.89	0.97	0.82	0.97	0.79	0.54
F-1	0.88	0.97	0.86	0.96	0.80	0.65

Table 9: Accuracy (mean ± standard error) for Logistic Regression (LR) and Random Forest (RF) using 5-fold cross-validation for LargeLeakPct with 4 weeks of data as predictors

	KmL	KmL3D	LMKM	GCKM	HDDC	Deepgmm
LR	0.95 ±0.006	0.92 ±0.005	0.96 ±0.004	0.97 ±0.005	0.76 ±0.008	0.883 ±0.015
RF	0.96 ±0.002	0.93 ±0.008	0.96 ±0.007	0.98 ±0.007	0.78 ±0.005	0.884 ±0.013

Table 10: Macro F1-scores for Logistic Regression (LR) and Random Forest (RF) using 5-fold cross-validation with 4 weeks of data as predictors for LargeLeakPct

	KmL	KmL3D	LMKM	GCKM	HDDC	Deepgmm
LR	0.39	0.74	0.56	0.88	0.56	0.75
RF	0.41	0.77	0.57	0.90	0.75	0.76

6.3.2 *Classification for LargeLeakPct*

Appendix B (page 40) presents the classification results for LargeLeakPct clusters memberships. The results show that the Random Forest classifier with LargeLeakPct cluster memberships obtained from GCKM outperforms other combinations based on accuracy and macro F1-scores. According to the figures in Appendix B, the result indicated that we can perform classification with merely four weeks of initial device data since the accuracy and macro F1-score did not increase significantly after four weeks. The 5 folds cross-validation scores are shown in Table 9 and Table 10. Table 11 shows the accuracy and macro F1-scores of the best two models (with KmL3D and GCKM cluster memberships) on test set. Appendix E (page 46) shows the confusion matrix for the results obtained from KmL3D and GCKM with Random Forest on test set. The results conclude that the Random Forest with cluster memberships generated from GCKM can be generalized well on test set. Furthermore, Table 12 displays the statistics per clusters for the best two models for LargeLeakPct subgroups prediction.

Table 11: Accuracy and macro F1-score for LargeLeakPct GCKM and KmL3D on test set with Random Forest with 4 weeks of data

	Accuracy	Macro F1
GCKM	0.98	0.89
KmL3D	0.91	0.68

Table 12: Statistics per cluster (denoted as C_i, where i is the cluster number) for the best two Models of LargeLeakPct using 4 weeks of data

	Random Forest with GCKM			Random Forest with KmL3D		
Statistics	C_1	C_2	C_3	C_1	C_2	C_3
Prevalence	0.07	0.92	0.01	0.86	0.12	0.02
Accuracy	0.81	0.99	0.70	0.97	0.61	0.35
Precision	0.88	0.98	1.00	0.95	0.67	0.60
Recall	0.81	0.99	0.7	0.97	0.61	0.35
F-1	0.84	0.99	0.82	0.96	0.64	0.44

7 DISCUSSION

This thesis explores the common patterns in the first 90 days of therapy on AHI and large leakage percentage (LargeLeakPct) for CPAP patients using a self-supervised learning approach. First, we used longitudinal clustering to create the labels based on AHI or LargeLeakPct. We then evaluated the clustering results based on the Dunn Index and the mean value of device parameters (i.e., variables in Table 1) per cluster. Second, we used different sets of weeks of device parameters to assess the performance of predicting cluster memberships by accuracy and macro F1-scores.

Considering the complexity of the self-supervised learning approach for multi-class classification, the result of this thesis is promising and is able to be interpreted clinically. With this research, we found that it is possible for clinicians and device provider to understand and predict the underlying patterns of each patients subgroups in an early manner. In the following subsections, we explain the insights obtained from the experiment results to inform the CPAP therapy practice and data analysis.

7.1 *Sub Research Question 1*

7.1.1 *Insights for CPAP Therapy Practice*

The result shows that GCKM can summarize CPAP patient therapy patterns into three AHI clusters, each with different characteristics that can be explained by clinicians. Specifically, the minority clusters reveal problematic therapy patterns, which indicates that patients belonging to the minority clusters may need early assistance.

As mentioned in the result section (section 6.2.1), AHI univariate clustering shows that the GCKM with 3 clusters outperforms other clustering algorithms. Clusters generated by the GCKM model are unbalanced, with the largest cluster having 61.54% patients. Using the elbow method with BIC, GCKM created meaningful clusters since the mean of device parameters are different among clusters significantly. Additionally, we observe that patients from cluster 2 are the most stable, adherent, and treated subgroups among the 3 clusters, which represented the majority of the patients (61.54%). On the other hand, although the AHI trajectory of patients from cluster 3 shows a decreasing trend over time, the AHI value is still comparatively high and unstable (i.e., with high variance) compared with other clusters, with also the highest large leakage percentage and lowest usage hours (10.36% of patients). Although the trajectories from cluster 1 (28.1%) patients seem stable, they have higher AHI and large leakage percentage than the other more stable clusters. It is also noted that cluster 3 has the highest average Pressure setting among the three clusters.

Similar to the insight obtained from AHI clustering using GCKM, the LMKM model can create three meaningful LargeLeakPct clusters, which can help identify large leakage patterns in CPAP therapy devices. The experiment shows that for the LargeLeakPct univariate clustering, LMKM with 3 clusters outperforms other methods based on Dunn Index. It is also noted that LMKM also obtained the highest on Silhouette score. The clusters obtained using LMKM are extremely unbalanced, with the majority cluster having 95.7% of patients. Based on the ANOVA test, the average of device variables significantly differed from each other among clusters. The LargeLeakPct trajectory for patients in cluster 2 is the most stable, has the highest usage hours, and has the lowest AHI. Moreover, the trajectories of cluster 1 and cluster 3 show problematic patterns, where the LargeLeakPct of cluster 3 (2.56% of patients) has a higher AHI and LargeLeakPct with a decreasing LargeLeakPct trend over time. Also, patients from cluster 3 have higher usage hours than cluster 1. On the other hand, the LargeLeakPct of patients from cluster 1 (1.74%) shows an increasing trend over time, with lowest usage hours and similar LargeLeakPct and AHI values with cluster 3. These problematic patterns indicate that patients from cluster 1 and 3 may have abnormal leakage over time during their therapy with higher AHI, meaning they may receive less effective treatment and thus need further assistance in receiving CPAP therapy. It would be worth further research on the relationships among LargeLeakPct, AHI, UsageHours, and PressureCpap over time in the future. With all the above findings, we conclude that LMKM can identify patients' mask fit patterns.

7.1.2 *Insights for CPAP Data Analysis*

For CPAP data analysis, the first insight is that Dunn Index may not be a fair evaluation metric when assessing the performance of univariate clustering with our AHI and LargeLeakPct data. It is observed that the Dunn Index, in general, obtained from this study is comparatively low when compared with other research (see section 3.4). We suspect it is due to our univariate clustering approach with unbalanced data. A previous study (Sobisek et al., 2018) showed that with unbalanced and high noise data, the Dunn index is significantly lower than balanced and low noise data. Furthermore, another work also demonstrated that univariate clustering with unbalanced data obtained Dunn Index lower than 0.05 (Lynch & DeGruttola, 2022). Additionally, due to the low performance of the Dunn Index, we compute the Silhouette score to make our research more comprehensive. Our result shows that Silhouette score performs adequately with some discrepancy with Dunn Index. Future research might be worth exploring other suitable evaluation metrics for high noise and imbalanced data.

The second insight is that although HDDC has been proven useful in the literature for other tasks, it does not perform well on our data to cluster CPAP therapy patterns meaningfully. Surprisingly, HDDC performed with the lowest Dunn Index and Silhouette score among all of the methods with both univariate clustering tasks, with only 0.002 and 0.00002 Dunn Index for AHI and LargeLeakPct univariate clustering, respectively. Our assumption is that various temporal features make the input features high dimensional, which matches the HDDC assumption we discussed in the method section. However, compared with other methods using temporal features, such as Deepgmm, and KmL3D, the performance of other models is much better than HDDC. Appendix C and D (see page 42, and page 44) show the mean values of device parameters per cluster of each clustering models. It is found that HDDC identifies comparatively not very different clusters among all the models, and hence, it might explain the low Dunn Index and Silhouette score.

7.2 *Sub Research Question 2*

Experiments show that using Random Forest to predict GCKM-generated clusters outperform other combinations and can perform well with as early as four weeks of data based on accuracy. Our self-supervised learning approach for early prediction of patients' cluster memberships has provided good results with GCKM univariate clustering. Based on the findings of our clustering analysis, we can identify patient cluster membership at an early stage. With our classification task, we found that with only four weeks of device data, it is possible to predict CPAP patients' AHI patterns during therapy with 93.8% accuracy and 90.9% macro F1-scores, meaning the model could capture most of the cluster memberships of patients correctly. Also, for the LargeLeakPct patterns, we obtained excellent performance with 98% accuracy and 90% macro F1-scores with cluster memberships generated from GCKM, which means it is also possible to examine as early as the fourth week for the abnormal mask leakage patterns among CPAP patients. When we explore the performance in each class, we conclude that Random Forest can predict GCKM cluster memberships with adequate recall and excellent precision for both AHI and LargeLeakPct classification. In sum, for new patients, it is possible to predict their subgroups with high accuracy and macro f1-score. Clinicians and device providers can assist the patients early if the patients are in the problematic subgroups.

7.3 *Limitation and Future Direction*

We identified several major limitations in this research. First, we only sample 5,000 patients with their first 90 days of CPAP data due to the complexity of the task. In the future, it is recommended to re-sample the patients several times or use full data to investigate how well the proposed approaches in this research can be generalized to the population. Secondly, we only adopt BIC to decide the optimal number of clusters. In order to make the research more comprehensive, it might be beneficial to take different methods into account, such as Davies–Bouldin index and Calinski-Harbasz score. Thirdly, we did not extensively tune the model hyper-parameters since our research goal for SQ2 focuses on understanding which supervised learning model can predict the SQ1 output with the highest accuracy, and how early we can predict it. Future research can tune hyper-parameters in our supervised learning models using grid-search. Finally, our Logistic Regression did not include regularization. Future research can also explore ways of adding regularization terms in the Logistic Regression. For the future study on clustering longitudinal CPAP patient data, it may be worth exploring the possibility of adopting unsupervised deep learning techniques. For example, De Jong et al. (2019) proposed a deep learning-based model, Variational Deep Embedding with Recurrence (VaDER), to perform cluster analysis. They discussed that VaDER is advantageous for longitudinal data with missing values because it can treat missing values directly by integrating model training with imputation. The paper displayed that they could accurately classify the clusters for the simulated dataset (cluster purity > 0.9). The researchers specified that they could effectively classify patients with Parkinson's and Alzheimer's diseases into subgroups using VaDER.

Moreover, for the supervised learning step in this research, the Random Forest classifier performs well for both classification tasks. However, for the future work, it would be interesting to implement the XGBOOST or Deep Learning approach due to their proven ability to predict multi-class problems with high performance.

It is also suggested that future work could apply a semi-supervised learning approach to ensure the quality of both unsupervised learning and supervised learning tasks. It could be difficult to label all data in medical datasets since domain experts are needed to label each observation manually, which could be laborious, time-consuming, and expensive. However, with semi-supervised learning, it may be more practical to have the domain experts label a small size of the datasets and train the models on mixed labeled and unlabeled datasets. In this way, the clusters will contain the labels provided by domain experts, and therefore the labels can be used to further assess the quality of clusters.

8 CONCLUSION

With rich data obtained from CPAP patients, researchers can further analyze different OSA patient subgroups and their underlying patterns by clustering analysis, allowing the clinicians and device providers to better assist the patients in adhering to the CPAP therapy. The result can also be further analyzed by self-supervised learning techniques to examine if the clustering memberships can be predicted with high performance, which can enable clinicians, device providers, and future researchers to understand how early we can predict the therapy patterns for CPAP patients. With this study, it can be concluded that it is possible to predict and understand the subgroups of CPAP patients with high performance in early-stage with a self-supervised learning pipeline, which can be beneficial for clinical practice. In future research, the results from the present study can be improved by implementing a semi-supervised learning approach, finding suitable evaluation metrics, and experimenting with other supervised and unsupervised learning models.

REFERENCES

Babbin, S. F., Velicer, W. F., Aloia, M. S., & Kushida, C. A. (2015). Identifying longitudinal patterns for individuals and subgroups: an example with adherence to treatment for obstructive sleep apnea. *Multivariate behavioral research*, *50*(1), 91–108.

Bergé, L., Bouveyron, C., & Girard, S. (2012). Hdclassif: An r package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, *46*(6), 1–29.

Bergé, L., Bouveyron, C., & Girard, S. (2012). HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, *46*(6), 1–29. Retrieved from http://www.jstatsoft.org/v46/i06/

Bouveyron, C., Girard, S., & Schmid, C. (2007). High-dimensional data clustering. *Computational statistics & data analysis*, *52*(1), 502–519.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. *Journal of Statistical Software*, *25*(4), 1–22. Retrieved from https://www.jstatsoft.org/v25/i04/

Chen, S. S., & Gopalakrishnan, P. S. (1998). Clustering via the bayesian information criterion with applications in speech recognition. In *Proceedings of the 1998 ieee international conference on acoustics, speech and signal processing, icassp'98 (cat. no. 98ch36181)* (Vol. 2, pp. 645–648).

De Jong, J., Emon, M. A., Wu, P., Karki, R., Sood, M., Godard, P., ... Fröhlich, H. (2019). Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience*, *8*(11), giz134.

Den Teuling, N. (2022). latrend: A framework for clustering longitudinal data [Computer software manual]. Retrieved from https://github.com/philips-software/latrend (R package version 1.2.3)

Den Teuling, N., Pauws, S., & van den Heuvel, E. (2020). A comparison of methods for clustering longitudinal data with slowly changing trends. *Communications in Statistics-Simulation and Computation*, 1–28.

Den Teuling, N., van den Heuvel, E. R., Aloia, M. S., & Pauws, S. C. (2021). A latent-class heteroskedastic hurdle trajectory model: patterns of adherence in obstructive sleep apnea patients on cpap therapy. *BMC medical research methodology*, *21*(1), 1–15.

Doersch, C., & Zisserman, A. (2017). Multi-task self-supervised visual learning. In *Proceedings of the ieee international conference on computer vision* (pp. 2051–2060).

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, *4*(1), 95–104.

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015a). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, *65*(4), 1–34. Retrieved from http://www.jstatsoft.org/v65/i04/

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015b). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, *65*(4), 1–34. Retrieved from http://www.jstatsoft.org/v65/i04/

Genolini, C., Pingault, J.-B., Driss, T., Côté, S., Tremblay, R. E., Vitaro, F., ... Falissard, B. (2013). Kml3d: a non-parametric algorithm for clustering joint trajectories. *Computer methods and programs in biomedicine*, *109*(1), 104–111.

Hardy, A. (1994). An examination of procedures for determining the number of clusters in a data set. In *New approaches in classification and data analysis* (pp. 178–185). Springer.

Kendzerska, T., Mollayeva, T., Gershon, A. S., Leung, R. S., Hawker, G., & Tomlinson, G. (2014). Untreated obstructive sleep apnea and the risk for serious long-term adverse outcomes: a systematic review. *Sleep medicine reviews*, *18*(1), 49–59.

Kim, S.-J., Alnakhli, W. M., Alfaraj, A. S., Kim, K.-A., Kim, S.-W., & Liu, S. Y.-C. (2021). Multi-perspective clustering of obstructive sleep apnea towards precision therapeutic decision including craniofacial intervention. *Sleep and Breathing*, *25*(1), 85–94.

Kribbs, N. B., Pack, A. I., Kline, L. R., Getsy, J. E., Schuett, J. S., Henry, J. N., ... Dinges, D. F. (1993). Effects of one night without nasal cpap treatment on sleep and sleepiness in patients with obstructive sleep apnea. *Am Rev Respir Dis*, *147*(5), 1162–1168.

Kribbs, N. B., Pack, A. I., Kline, L. R., Smith, P. L., Schwartz, A. R., Schubert, N. M., ... Dinges, D. F. (2012). Objective measurement of patterns of nasal cpap use by patients with obstructive sleep apnea. *American Review of Respiratory Disease*.

Kuhn, M. (2021). caret: Classification and regression training [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=caret (R package version 6.0-88)

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lynch, M. L., & DeGruttola, V. (2022). Ensemble clustering of longitudinal bivariate hiv biomarker profiles to group patients by patterns of disease progression. *International journal of data science and analytics*, 1–14.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

Quan, S., Gillin, J., Littner, M., & Shepard, J. (1999, 01). Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. editorials. *Sleep*, *22*, 662-689.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Scrucca, L., & Raftery, A. E. (2018). clustvarsel: A package implementing variable selection for gaussian model-based clustering in r. *Journal of Statistical Software*, *84*.

Senaratna, C. V., Perret, J. L., Lodge, C. J., Lowe, A. J., Campbell, B. E., Matheson, M. C., ... Dharmage, S. C. (2017). Prevalence of obstructive sleep apnea in the general population: a systematic review. *Sleep medicine reviews*, *34*, 70–81.

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 1–16.

Sobisek, L., Stachova, M., & Fojtik, J. (2018). Novel feature-based clustering of micro-panel data (clump). *arXiv preprint arXiv:1807.05926*.

Sopkova, Z., Dorkova, Z., & Tkacova, R. (2009). Predictors of compliance with continuous positive airway pressure treatment in patients with obstructive sleep apnea and metabolic syndrome. *Wiener Klinische Wochenschrift*, *121*(11), 398–404.

Teuling, N. D., Pauws, S., & Heuvel, E. v. d. (2021). Clustering of longitudinal data: A tutorial on a variety of approaches. *arXiv preprint arXiv:2111.05469*.

Twisk, J., & Hoekstra, T. (2012). Classifying developmental trajectories over time should be done with great caution: a comparison between methods. *Journal of clinical epidemiology*, *65*(10), 1078–1087.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/ (ISBN 0-387-95457-0)

Viroli, C., & McLachlan, G. J. (2019). Deep gaussian mixture models. *Statistics and Computing*, *29*(1), 43–51.

Viroli, C., & McLachlan, G. J. (2020). deepgmm: Deep gaussian mixture models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=deepgmm (R package version 0.1.62)

Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data*. Oxford University Press.

Weaver, T. E., Kribbs, N. B., Pack, A. I., Kline, L. R., Chugh, D. K., Maislin, G., ... others (1997). Night-to-night variability in cpap use over the first three months of treatment. *Sleep*, *20*(4), 278–283.

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1476–1485).

Zheng, X., Wang, Y., Wang, G., & Liu, J. (2018). Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, *107*, 55–71.

APPENDIX A

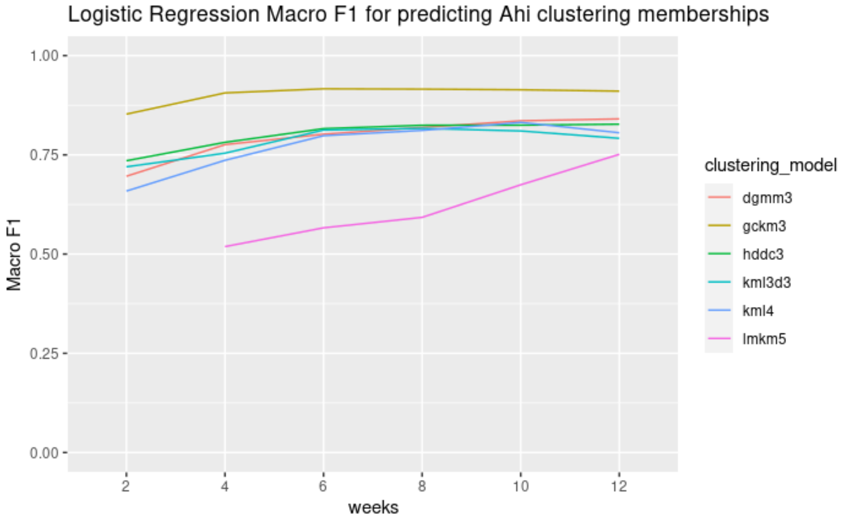


Figure 11: Logistic Regression macro F1-scores for AHI classification

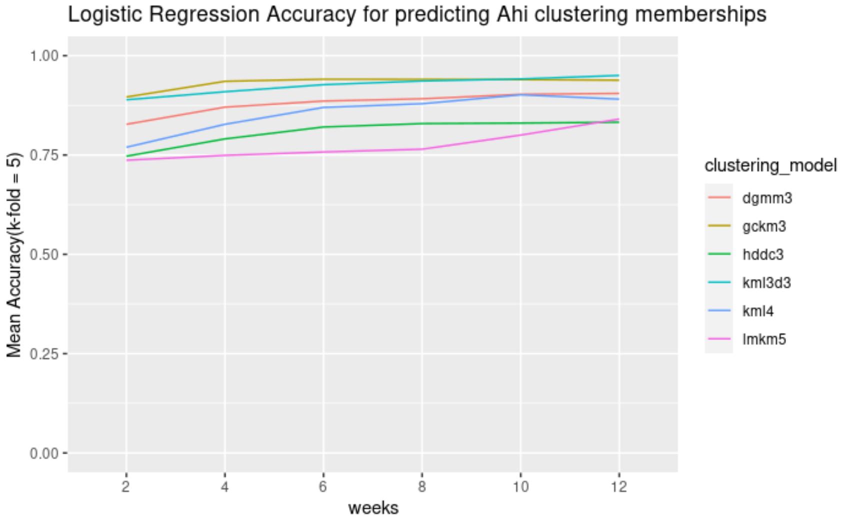


Figure 12: Logistic Regression mean Accuracy (5-fold cross validation) for AHI classification

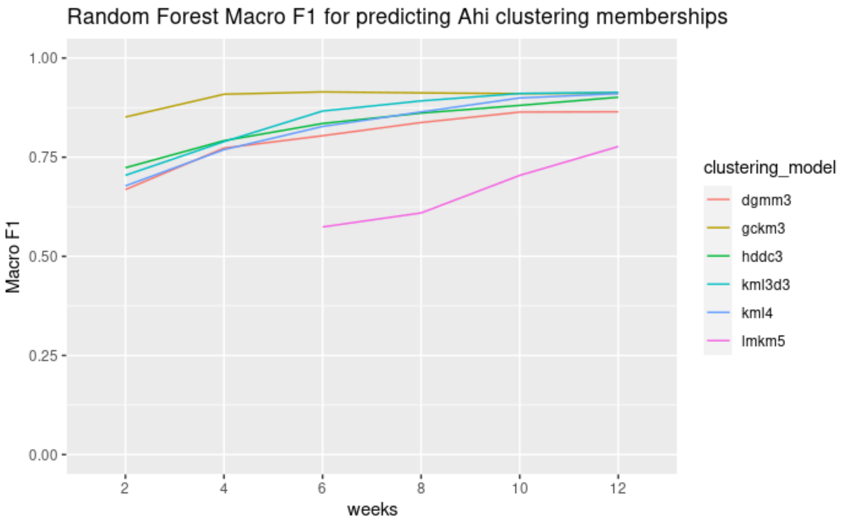


Figure 13: Random Forest macro F1-scores for AHI classification

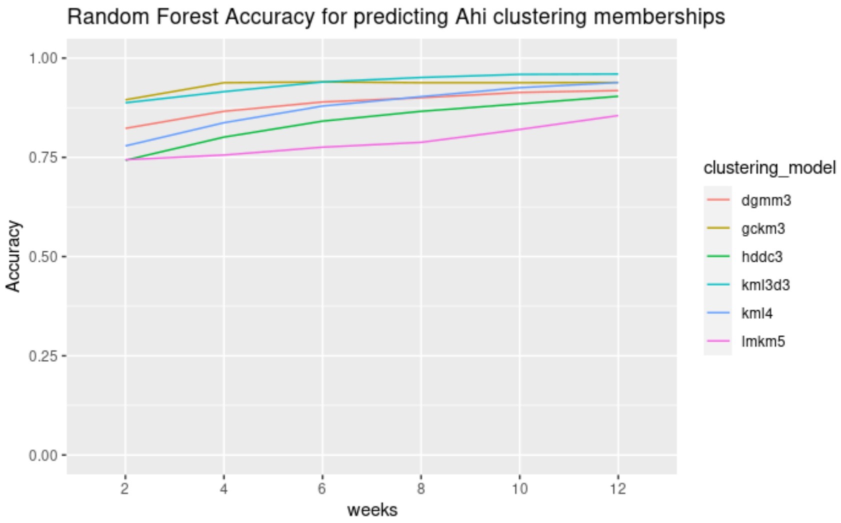


Figure 14: Random Forest mean Accuracy (5-fold cross validation) for AHI classification

APPENDIX B

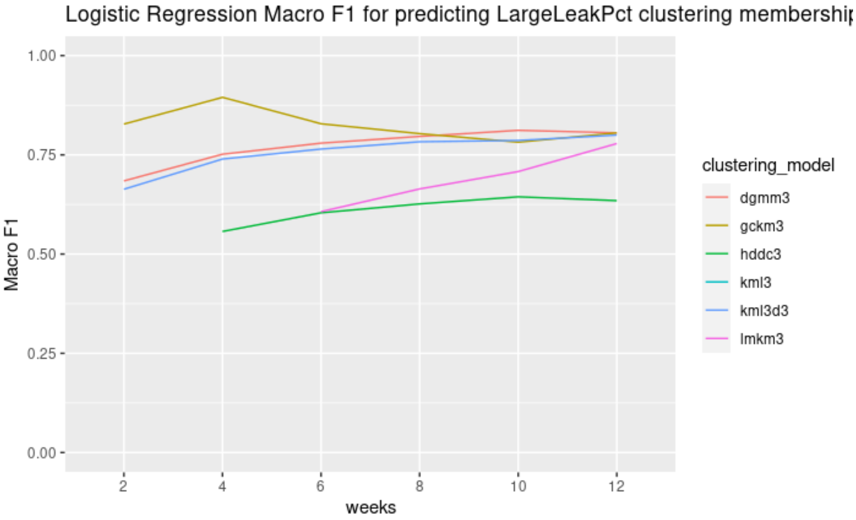


Figure 15: Logistic Regression macro F1-scores for LargeLeakPct classification

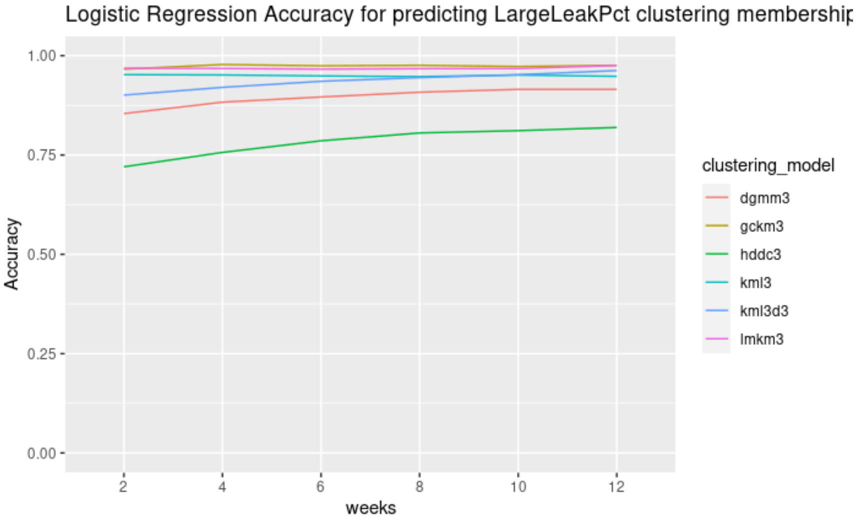


Figure 16: Logistic Regression mean Accuracy (5-fold cross validation) for LargeLeakPct classification

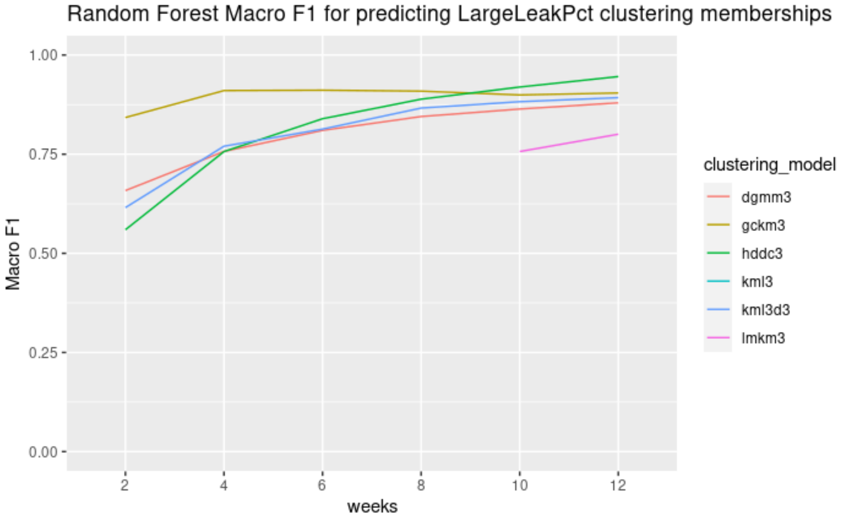


Figure 17: Random Forest macro F1-scores for LargeLeakPct classification

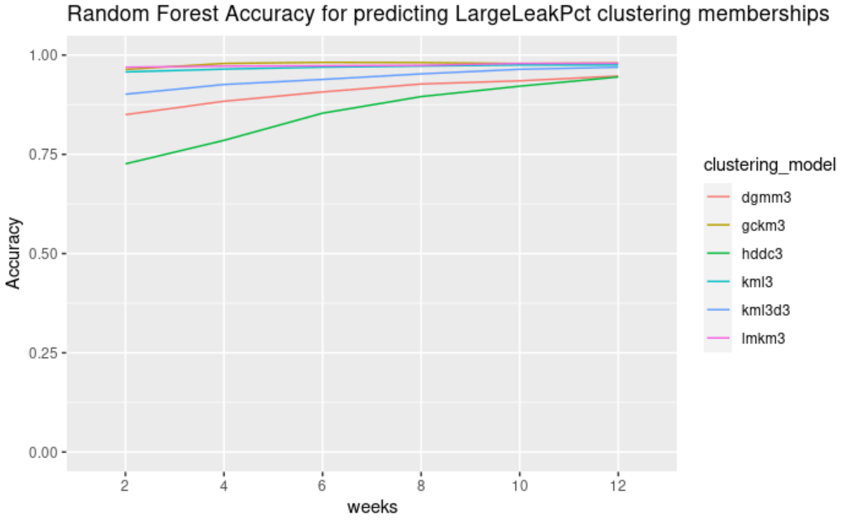
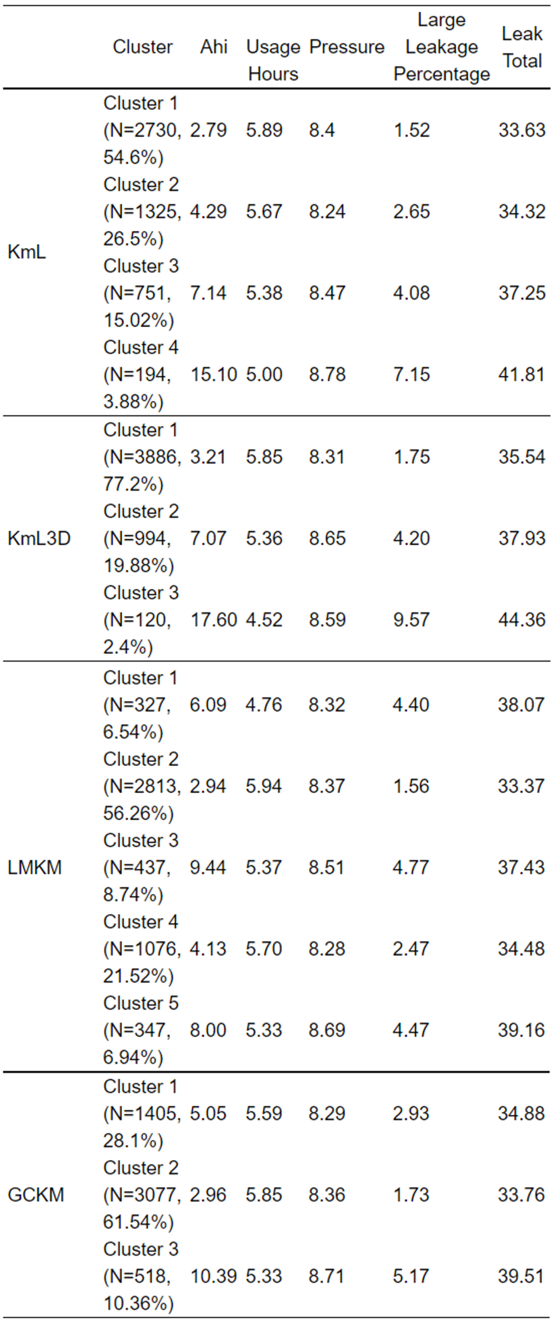


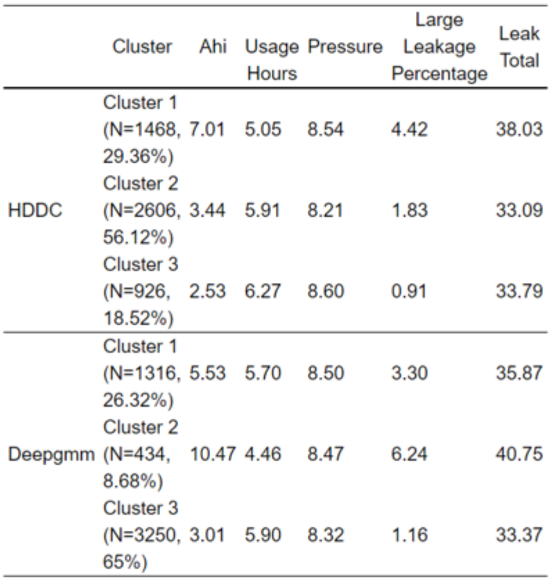
Figure 18: Random Forest mean Accuracy (5-fold cross validation) for LargeLeakPct classification

Appendix C: Cluster variables average for AHI univariate clustering



	Cluster	Ahi	Usage Hours	Pressure	Large Leakage Percentage	Leak Total
KmL	Cluster 1 (N=2730, 54.6%)	2.79	5.89	8.4	1.52	33.63
	Cluster 2 (N=1325, 26.5%)	4.29	5.67	8.24	2.65	34.32
	Cluster 3 (N=751, 15.02%)	7.14	5.38	8.47	4.08	37.25
	Cluster 4 (N=194, 3.88%)	15.10	5.00	8.78	7.15	41.81
KmL3D	Cluster 1 (N=3886, 77.2%)	3.21	5.85	8.31	1.75	35.54
	Cluster 2 (N=994, 19.88%)	7.07	5.36	8.65	4.20	37.93
	Cluster 3 (N=120, 2.4%)	17.60	4.52	8.59	9.57	44.36
LMKM	Cluster 1 (N=327, 6.54%)	6.09	4.76	8.32	4.40	38.07
	Cluster 2 (N=2813, 56.26%)	2.94	5.94	8.37	1.56	33.37
	Cluster 3 (N=437, 8.74%)	9.44	5.37	8.51	4.77	37.43
	Cluster 4 (N=1076, 21.52%)	4.13	5.70	8.28	2.47	34.48
	Cluster 5 (N=347, 6.94%)	8.00	5.33	8.69	4.47	39.16
GCKM	Cluster 1 (N=1405, 28.1%)	5.05	5.59	8.29	2.93	34.88
	Cluster 2 (N=3077, 61.54%)	2.96	5.85	8.36	1.73	33.76
	Cluster 3 (N=518, 10.36%)	10.39	5.33	8.71	5.17	39.51

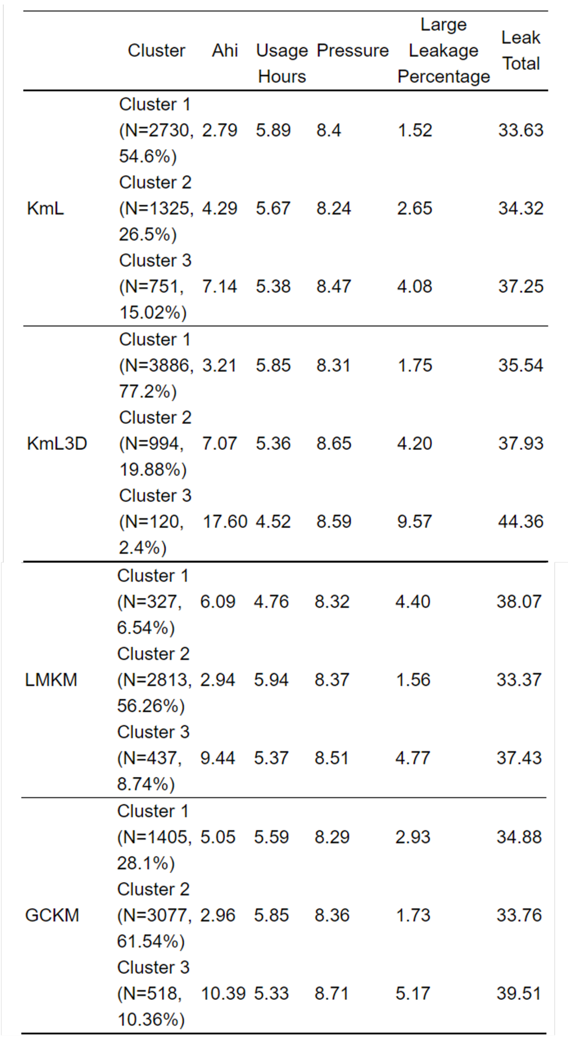
Figure 19: Cluster variables average for AHI univariate clustering-1



	Cluster	Ahi	Usage Hours	Pressure	Large Leakage Percentage	Leak Total
HDDC	Cluster 1 (N=1468, 29.36%)	7.01	5.05	8.54	4.42	38.03
	Cluster 2 (N=2606, 56.12%)	3.44	5.91	8.21	1.83	33.09
	Cluster 3 (N=926, 18.52%)	2.53	6.27	8.60	0.91	33.79
Deepgmm	Cluster 1 (N=1316, 26.32%)	5.53	5.70	8.50	3.30	35.87
	Cluster 2 (N=434, 8.68%)	10.47	4.46	8.47	6.24	40.75
	Cluster 3 (N=3250, 65%)	3.01	5.90	8.32	1.16	33.37

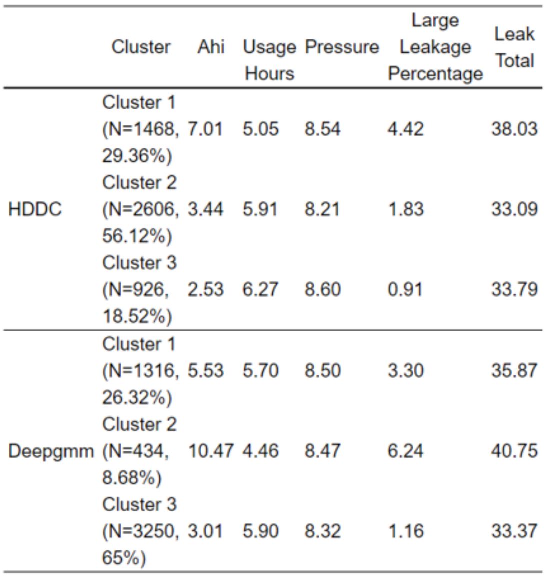
Figure 20: Cluster variables average for AHI univariate clustering-2

appendix d: cluster variables average for largeleakpct univariate clustering



	Cluster	Ahi	Usage Hours	Pressure	Large Leakage Percentage	Leak Total
KmL	Cluster 1 (N=2730, 54.6%)	2.79	5.89	8.4	1.52	33.63
	Cluster 2 (N=1325, 26.5%)	4.29	5.67	8.24	2.65	34.32
	Cluster 3 (N=751, 15.02%)	7.14	5.38	8.47	4.08	37.25
KmL3D	Cluster 1 (N=3886, 77.2%)	3.21	5.85	8.31	1.75	35.54
	Cluster 2 (N=994, 19.88%)	7.07	5.36	8.65	4.20	37.93
	Cluster 3 (N=120, 2.4%)	17.60	4.52	8.59	9.57	44.36
LMKM	Cluster 1 (N=327, 6.54%)	6.09	4.76	8.32	4.40	38.07
	Cluster 2 (N=2813, 56.26%)	2.94	5.94	8.37	1.56	33.37
	Cluster 3 (N=437, 8.74%)	9.44	5.37	8.51	4.77	37.43
GCKM	Cluster 1 (N=1405, 28.1%)	5.05	5.59	8.29	2.93	34.88
	Cluster 2 (N=3077, 61.54%)	2.96	5.85	8.36	1.73	33.76
	Cluster 3 (N=518, 10.36%)	10.39	5.33	8.71	5.17	39.51

Figure 21: Cluster variables average for LargeLeakPct univariate clustering-1



	Cluster	Ahi	Usage Hours	Pressure	Large Leakage Percentage	Leak Total
HDDC	Cluster 1 (N=1468, 29.36%)	7.01	5.05	8.54	4.42	38.03
	Cluster 2 (N=2606, 56.12%)	3.44	5.91	8.21	1.83	33.09
	Cluster 3 (N=926, 18.52%)	2.53	6.27	8.60	0.91	33.79
Deepgmm	Cluster 1 (N=1316, 26.32%)	5.53	5.70	8.50	3.30	35.87
	Cluster 2 (N=434, 8.68%)	10.47	4.46	8.47	6.24	40.75
	Cluster 3 (N=3250, 65%)	3.01	5.90	8.32	1.16	33.37

Figure 22: Cluster variables average for LargeLeakPct univariate clustering-2

APPENDIX E: CONFUSION MATRIX FOR AHI AND LARGELEAKPCT

Table 13: Confusion Matrix (Random Forest for GCKM AHI with four weeks of data)

		Predicted		
		Cluster 1	Cluster 2	Cluster 3
Ground Truth	Cluster 1	251	21	19
	Cluster 2	22	594	0
	Cluster 3	8	0	84

Table 14: Confusion Matrix (Random Forest for Kml3D AHI with four weeks of data)

		Predicted		
		Cluster 1	Cluster 2	Cluster 3
Ground Truth	Cluster 1	750	39	0
	Cluster 2	27	156	11
	Cluster 3	0	3	13

Table 15: Confusion Matrix (Random Forest for GCKM LargeLeakPct with four weeks of data)

		Predicted		
		Cluster 1	Cluster 2	Cluster 3
Ground Truth	Cluster 1	56	6	2
	Cluster 2	13	908	0
	Cluster 3	4	2	8

Table 16: Confusion Matrix (Random Forest for KmL3D LargeLeakPct with four weeks of data)

		Predicted		
		Cluster 1	Cluster 2	Cluster 3
Ground Truth	Cluster 1	834	43	0
	Cluster 2	26	75	11
	Cluster 3	0	4	6