駱書羽 黃偉祥 劉天恩

景信勇 狄佳多 申泳

Introduction

Type 1 diabetes is a disease characterized by instability and unpredictability of blood glucose levels. The complex nature of this condition makes it challenging for patients to maintain optimal glucose control, often leading to frequent fluctuations between hypoglycemia and hyperglycemia.

Continuous Glucose Monitoring (CGM) systems represent a significant advancement in diabetes management, providing continuous rather than discrete point measurement of blood glucose levels. By continuously tracking blood glucose levels, CGM provides glucose patterns that offer information about the key behaviors that may be causing fluctuations. This detailed information is invaluable for making informed decisions regarding treatment plans and lifestyle modifications aimed at improving blood glucose control.

Despite the importance of CGM pattern recognition. There are some limitations. Most of the CGM datasets are too small to be effectively applied to machine learning or deep learning algorithms. This limitation necessitates the combination of multiple datasets. However, this approach is not without its challenges; there may be bias hidden within each dataset that could affect the accuracy and reliability of the results.

Another critical challenge is the shortage of healthy CGM datasets. To address this, there is a growing need for synthetic healthy CGM data based on real-world data. However, the reliability of these generated datasets remains a significant concern. Ensuring that the synthetic data accurately reflects real-world glucose patterns and behaviors is crucial for the efficacy of subsequent analyses and interventions.

Another aspect to consider is the relationship between HbA1c and Time in Range (TIR). HbA1c, a measure of average blood glucose levels over the past two to three months, has been found to have a significant correlation with TIR, which indicates the percentage of time a person's glucose levels remain within the target range. However, it has not yet been proven that TIR has a relationship with the diversity of CGM patterns.

Thus, in our study, we aim to verify the following five hypotheses:

- The diversity of daily CGM patterns has a linear relationship to HbA1c/TIR levels. We hypothesize that individuals with higher HbA1c/lower TIR levels will exhibit more diverse CGM patterns.
- When using multi-source datasets, there might be some bias patterns existed. This implies that different data sources could be separable through clustering techniques.
- We utilize dynamic time warping based techniques to synthesize healthy CGM data. This approach addresses the shortage of healthy CGM data. We aim to demonstrate that the patterns and distribution of generated CGM data are quite like those of original healthy data.
- We prepared a single source dataset and a multi source dataset. We are also evaluating domain shifting issue by applying the clustering model trained on single source dataset to multi source dataset.
- We discretize age and TIR in test samples. And we also prepared categorical treatment variables. We want to prove that samples with similar features(ex. at the same age group) have a similar CGM pattern distribution than samples with different features(ex. from different age groups).
- We are also exploring, for multi source dataset, will the clustering algorithm sacrifice smaller data sources, so that the samples coming from these sources would have a higher chance of being unclassified?

We propose using unsupervised learning to cluster daily CGM patterns and evaluate their distribution over a monitoring span. While traditional methods like k-means are susceptible to outliers, and DBSCAN struggles with border points, we aim to develop advanced techniques to mitigate these issues.

In “Identification of clinically relevant deglycation phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c.”, [9] the authors use some advanced glucose variability indexes to evaluate and do CGM phenotype clustering. But these glucose variability indexes only represent the coarse grain variability.

And both "Machine Learning–Based Time in Patterns for Blood Glucose Fluctuation Pattern Recognition in Type 1 Diabetes Management: Development and Validation Study."[10] and “The Development and Potential Applications of an Automated Method for Detecting and Classifying Continuous Glucose Monitoring Patterns” [11] use the combination of short windows (< 4 hours) CGM pattern clustering and further hierarchical clustering to aggregate the short-term patterns. But this method may lose the general view about the daily CGM pattern.

In “A Data-Driven Approach to Classifying Daily

source code: GitHub - sueyuu/2025_cgm_dataset

Continuous Glucose Monitoring (CGM) Time Series." [4], the authors use a new "motif based" clustering method which generates a set of dense, compact clusters of daily CGM pattern. The method avoids the pitfalls of k means method and DBSCAN method and provides more general view for daily CGM pattern than hierarchical clustering method. But the rigidity of RMSE distance measurements and hyperparameters setting makes the final set have over 400 motifs, which makes it too complex to be applied to the real world.

We propose our method to leverage the "motif based" method. But we replace RMSE distance measurements with dynamic time warping (DTW) distance measurements to solve the dynamic issue of daily CGM patterns. We also set a distance threshold for finding new motifs, which are the radius of already found motifs. The distance between a newly found motif and already found motifs should be larger than the distance threshold. The above rule is used to make sure the newly found motif is distant enough from the old ones, assuring they do not overlap too much. We also fine tune the hyperparameters to achieve the final motif set size under 100.

We are also trying the K medoid algorithm, we decide to set k based on the motif numbers we get from the motif algorithm. and set the outlier threshold as the radius upper bound of the motif algorithm. For evaluation, we evaluate the diversity and distribution of the daily CGM clusters of a proper monitoring span (we choose 7 days here).

Methodology

Data Sources Introduction

We found open-source data [1] which is a combination of multiple type1, type2 and healthy CGM datasets. Because the type 2 dataset it contains is quite small, we decided to ignore type 2 cases. We then selected several datasets based on subject numbers (ranging from dozens to hundreds), and it forms our multi-source CGM dataset. The multi-source dataset encompasses a total of 1236 subjects with Type 1 diabetes, with data collected over a period ranging from 3 to 6 months. Additionally, there are 418 healthy subjects with data spanning from one day to several days. (table 1)

Trial name	Sample sizes	Diabetes Type	Population Group(age)	CGM device	duration
Aleppo	225	1	25-40	Dexcom G4	6 months
Brown	168	1	14+	Dexcom G6 & t:Slim X2 with Control-IQ Technology	6 months
Lynch	90	1	6-71	Dexcom G6	13 weeks
Tambor lane	451	1	8+	FreeStyle Navigator & Dexcom SEVEN & Medtronic Paradigm	6 months
Wadwa	102	1	2-6	Dexcom G6	13 weeks
Colas	208 (17)	Healthy(type 2)	18+	iPro	<=2 day
Hall	57	Healthy & Pre-diabetes	18+	Dexcom G4	2-4 weeks
Shah	168	healthy	6+	Dexcom G6	<=10 days

Table 1. Summary of Multi-Source Dataset

Study summary of different datasets

Aleppo : The purpose of this study was to determine whether the use of continuous glucose monitoring (CGM) without blood glucose monitoring (BGM) measurements is as safe and effective as using CGM with BGM in adults (25-40) with type 1 diabetes under insulin pump treatment.

Brown : In this study, the patients with type 1 diabetes were assigned in a 2:1 ratio to receive treatment with a closed-loop system (closed-loop group) or a sensor-augmented pump (control group).

Lynch : To evaluate a transition from standard-of-care (SC) management of type 1 diabetes (any insulin delivery method including hybrid closed-loop systems plus real-time continuous glucose monitoring [CGM]) to use of the insulin-only configuration of the iLet® bionic pancreas (BP)

Tamborlane : This study was designed to test CGM as a technology to assist in diabetes care. The randomized trial was intended to determine if CGM usage had a positive effect on diabetes management.

Wadwa : The study is about young children (ages 2–6) with Type 1 diabetes using the t:slim X2 insulin pump with Control-IQ Technology and Dexcom G6 CGM

Colas : This study includes 208 subjects all of whom were healthy at study start and 17 of whom developed type 2 diabetes by study end.

Hall : This study analyzes how blood glucose fluctuates in healthy individuals by using a CGM to monitor glucose. Standardized meals (breakfast only) were given to a subset of patients in order to monitor the effect of meals on the glucose readings of healthy individuals.

Shah : to evaluate glucose control in a mixed population (ages 6 and older) and to establish reference sensor glucose ranges in healthy, non-diabetic individual

Additionally, we have another large dataset consisting of 736 patients with Type 1 Diabetes, monitored over a span of more than 4 years using the FreeStyle Libre device[2]. This dataset provides extensive insights into the long-term glucose control and variability in individuals with T1D.

We then filter out the subject with too much missing data with the following criteria [1]:

>=10% missing data for 1-day records

>=30% missing data for records spanning over 2 days

After filtering, we align and chunk each individual's data by:

If there is a missing gap for over 3 hours, chop the data.

Linearly interpolate data to every minute.

For each continuous CGM time series, make it start at 00:00 and end at 23:59 to ensure each daily data can be aligned in tensor.

For the healthy subjects, we concatenate all data together given the minimal dynamic changes in daily patterns, allowing us to treat the data as continuous due to the lack of significant overnight variations.

We choose the 3 hours gap as our largest tolerance for linear interpolation although a review paper [5] suggests only interpolate gaps that are less than about 20 minutes. But since it may generate too many small chunks in pattern evaluation tasks and we are going to apply smoothing later, we make a trade-off and finally decide 3 hours.

Figure 1 illustrates the filtering process, alignment, and chunking of the data, as well as handling of missing data gaps and interpolation

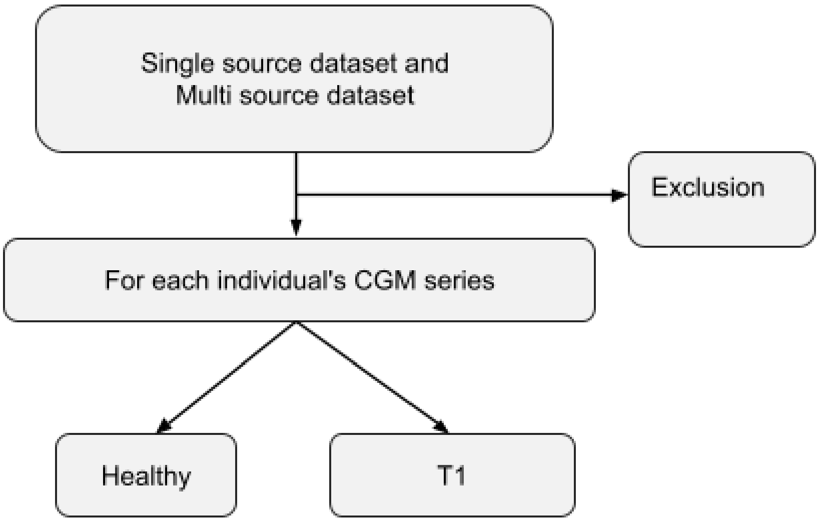


Figure 1. Flowchart of filtering and chunking CGM data

Exclusion

>=10% missing data for 1-day records.

>=30% missing data for records spanning over 2 days.

For each individual's CGM series

1. If there is a missing gap for over 3 hours, chop the data.

2. Linear interpolate each continuous data to every minute.

3. For each continuous CGM time series, make it start at 00:00 and end at 23:59 to ensure each daily data can be aligned in tensor.

Healthy

Concatenate all data into one.

T1

Don’t combine data together, as if there is a large missing gap between each continuous data, there might be some dramatic, unreasonable change when concatenate them together.

Generation of Healthy Data

We leverage two methods to generate healthy data [3]. If there is only one day data for a subject after above filtering and chunking, we first apply magnitude warping and time warping to generate over 5 days data. Then based on the data, use random guided warping to generate more data to expand total data span to 90 days. [figure 2] Random guided warping is an interpolation technique using dynamic time warping to align two time series instead of definite time points

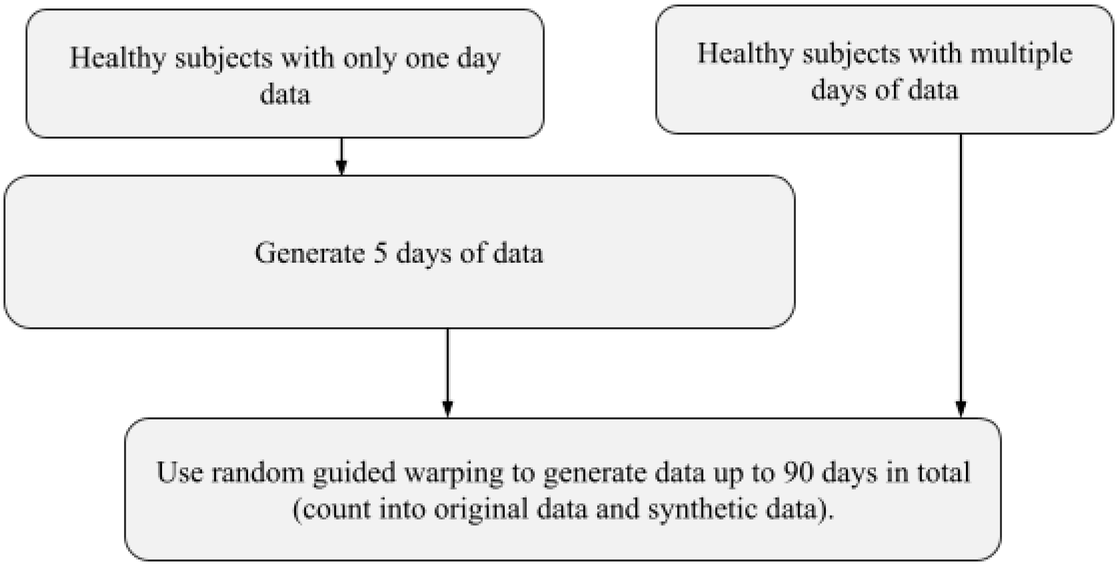

Healthy subjects with only one day data
Healthy subjects with multiple days of data
Generate 5 days of data
Use random guided warping to generate data up to 90 days in total (count into original data and synthetic data).

Figure 2. Process of generating synthetic healthy data from the existing records.

Clustering Methodology for CGM Data

Our approach to clustering Continuous Glucose Monitoring (CGM) data is inspired by the method proposed in the paper, “A Data-Driven Approach to Classifying Daily Continuous Glucose Monitoring (CGM) Time Series.” [4] While we largely adhere to the methodology outlined in the paper, we introduce several modifications.

For comparison, we also apply the k medoid algorithm, with hyperparameters like k and outlier threshold are set based on the configuration of the motif algorithm.

Algorithm introduction

1. **motif algorithm**

The same as original algorithm, we first transform the data using the following formula:

$$g\left(x_t\right) = \ln ln\left(x_t\right)^{1.084} - 5.381$$

Later, we calculate the distance between time series using dynamic time warping instead of original RMSE.

Then we follow the following steps to find out all the representative daily CGM motifs.

Given a set of daily profiles *φ* that have yet to be assigned to a cluster.

Find the pair of daily CGM profiles, say *dp_x* and *dp_y*, such that

$$f\left(dp_x, dp_y\right) \leq f\left(dp_j, dp_k\right)$$

and

$$f\left(dp_x, dp_y\right) < \gamma$$

and

$$f(dp_x,\ dp_m)\ <\ f(dp_m,\ dp_n)\ +\ 2\tau$$

and

$$f(dp_y,\ dp_m)\ <\ f(dp_m,\ dp_n)\ +\ 2\tau$$

and

$$f(dp_x,\ dp_n)\ <\ f(dp_m,\ dp_n)\ +\ 2\tau$$

and

$$f(dp_y,\ dp_m)\ <\ f(dp_m,\ dp_n)\ +\ 2\tau$$

for all $j,k\in\varphi$, and all pairs(m,n) in already found motifs. The score $f(dp_x,\ dp_y)$ is the minimum score among all pairs of daily CGM profiles in φ that is also strictly less than γ. Where *f* is the dynamic time warping distance calculation function in our definition. The parameter γ serves as a threshold criterion – the *maximum* possible score between two daily CGM profiles that will define a motif m_i. The pair $(dp_m,\ dp_n)$ is a pair of already found motif. The last four equations are used to make sure the newly found motif is distant enough from the old ones.

The dp_x and dp_y are then defined as a motif pair. Then for each daily CGM profile $dp_j\in\varphi\setminus\{dp_x, dp_y\}$, remove dp_j from *φ* if and only if

$$f\left(dp_x, dp_j\right) \leq f\left(dp_x, dp_y\right) + \tau$$

or

$$f\left(dp_y, dp_j\right) \leq f\left(dp_x, dp_y\right) + \tau$$

where *τ* is a tolerance value describing how close the match between dp_j and one of dp_x or dp_y must be for dp_j to be removed from φ.

Then do the above steps iteratively until no further motif is found.

2. k medoid algorithm

We simply apply k medoid algorithm designed for time series with initialize method forgy. We set k as the motif numbers we get from the motif algorithm. And set outlier threshold as the upper bound of the radius of the motif algorithm, which is $\gamma/2 + \gamma + \tau$.

Dividing Training, Validation, and Testing Datasets

The original paper of motif algorithm implements specific criteria for filtering out daily CGM data that exhibits abrupt increases or decreases, as well as data with excessive missing values. Our modifications to this process are as follows: we will use the filtering method described in Figure 1. And in place of the original criteria for filtering out abrupt increases or decreases, we apply Kalman smoothing method to the data.[12]

The original paper randomly assigns daily CGM data into training, validation, and testing sets in 15%, 20%, and 65% split respectively. However, there are some pitfalls hidden within this approach. Firstly, for a single subject, different daily data can be distributed into training, validation, and test sets, which may make the validation and test set results look remarkably satisfying. But because a subject's data could be in the training, validation, and test sets at the same time, the representativeness of the final motifs becomes questionable, as they could only represent the pattern distribution of the whole dataset.

To address this, if we ensure the data from a single subject could only be assigned to one of training, validation, and test sets, the result motif set will be more representative, making the conclusions more robust.[5] Secondly, the original splitting method does not employ stratification, which means some representative patterns related to certain demographics may be missing in the training set.[5] Therefore, instead of the original random assignment, we will implement a stratified splitting strategy to ensure that all demographic patterns are adequately represented.

Furthermore, because we have a multi-source dataset and a single-source dataset, if we do clustering on one of the datasets, we could use another dataset as an external validation set.[5] This approach will help verify the robustness and generalizability of our clustering results across different data contexts.

Because we aim to evaluate the diversity of CGM patterns over a monitoring span, we will first filter out subjects who lack continuous data for over 7 days, which is our primitive monitoring span. The data from these subjects will be evenly divided into training and validation sets. We will then stratify the T1 subjects included according to four demographics:

- Sex
- age group (<10, <20, 20-65, ≥65)
- mean HbA1c group (<7%, <8.5%, ≥8.5%)
- treatment group (multiple daily injection, basic pump, sensor-augmented pump, closed loop, bionic pancreas)

Healthy subjects will be stratified based on:

- sex
- age group (<10, <20, 20-65, ≥65)

With age groups roughly representing prepuberty, adolescence, adulthood, and elderly stages. The mean HbA1c group stratification is based on guidelines, with <7% indicating good control, 7-8.5% moderate control, and ≥8.5% poor control. [6][7]. And we can only stratify the t1 subjects from a single source dataset based on sex, age and hba1c because the single source dataset lacks the information about treatment group.

Based on the above ideas, we form our flow chart of dataset choosing and splitting based on our four main objectives. [figure 3]

Apply algorithm and CGM pattern distribution evaluation

We then apply different data sources to the motif clustering algorithm according to different objectives. For the training, we apply both the final training set and the CGM data that has been filtered out at the initial in figure 3. For objective 1.: to verify the relationship between diversity of daily CGM patterns and TIR/HbA1C. 4. domain shifting evaluation 5. evaluate CGM patterns distribution with different backgrounds. We are going to apply the single source dataset to avoid possible bias hidden inside the multi source dataset. And for the remaining objectives: 2. To find out if there is any possible bias hidden inside multi source dataset 3. To verify if there are any pattern differences between real healthy data and synthetic healthy data. 6. to evaluate the performance on the small data source. We are going to apply the multi source dataset, with synthetic healthy data included.

TIR for each chunked data. For further evaluation, we will randomly select out some data according to the following multi-labels.

- TIR group (<=50%, >50%, >70%, >90%) [7]
- Sex
- Age group
- Treatment group

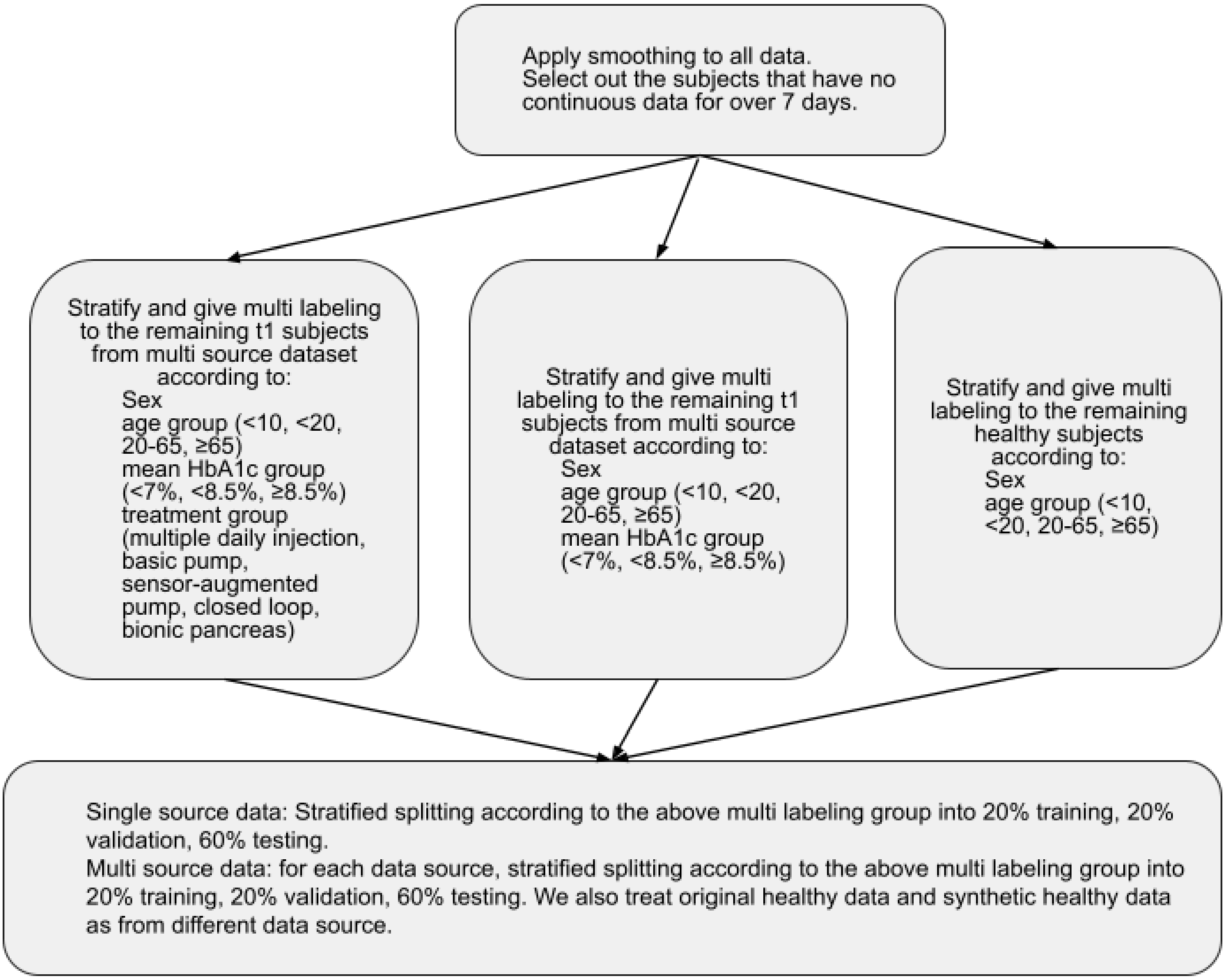
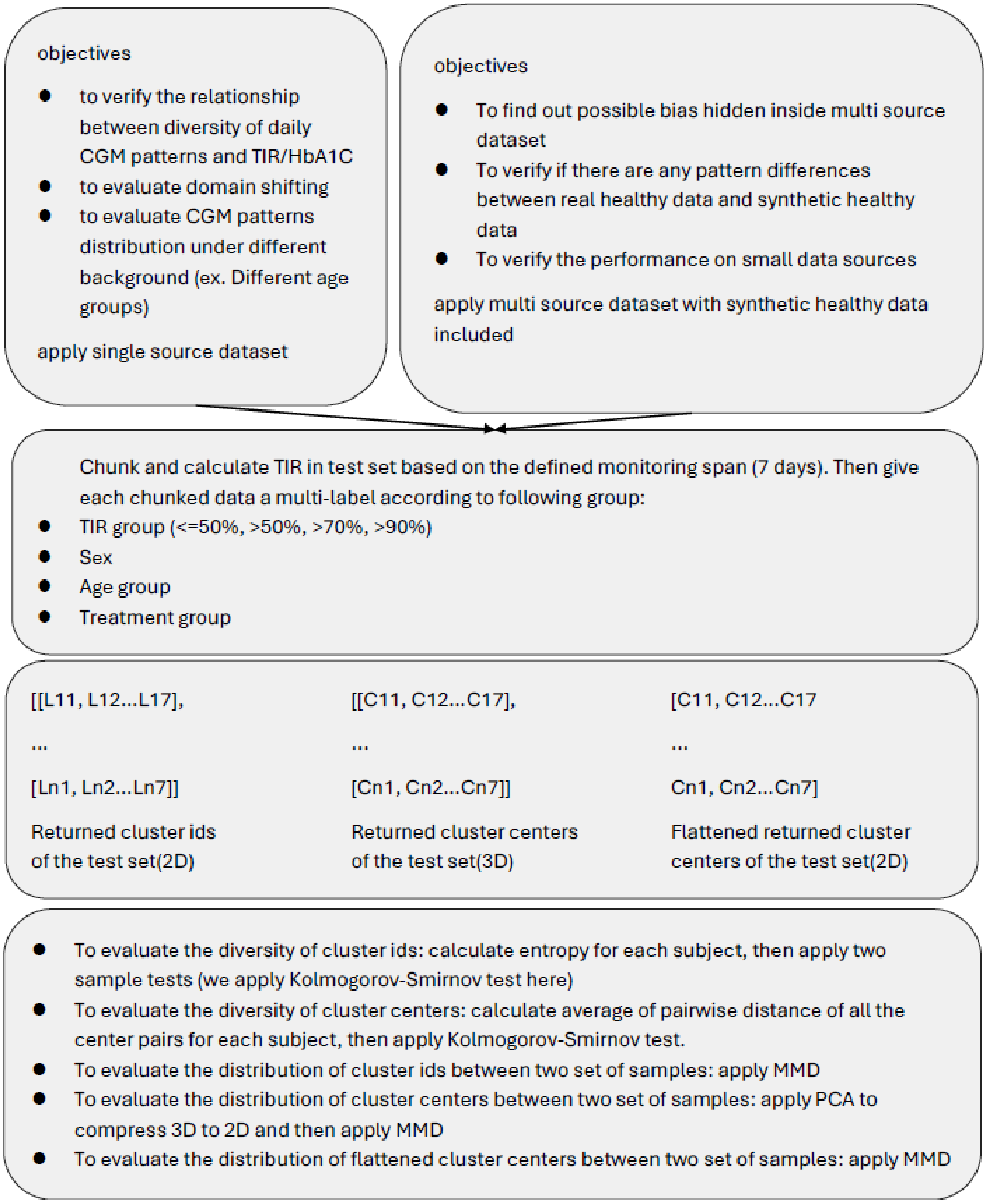

Apply smoothing to all data.
Select out the subjects that have no continuous data for over 7 days.
Stratify and give multi labeling to the remaining t1 subjects from multi source dataset according to:
Sex
age group (<10, <20, 20-65, ≥65)
mean HbA1c group (<7%, <8.5%, ≥8.5%)
treatment group (multiple daily injection, basic pump, sensor-augmented pump, closed loop, bionic pancreas)
Stratify and give multi labeling to the remaining t1 subjects from multi source dataset according to:
Sex
age group (<10, <20, 20-65, ≥65)
mean HbA1c group (<7%, <8.5%, ≥8.5%)
Stratify and give multi labeling to the remaining healthy subjects according to:
Sex
age group (<10, <20, 20-65, ≥65)
Single source data: Stratified splitting according to the above multi labeling group into 20% training, 20% validation, 60% testing.
Multi source data: for each data source, stratified splitting according to the above multi labeling group into 20% training, 20% validation, 60% testing. We also treat original healthy data and synthetic healthy data as from different data source.

Figure 3. process of splitting training, validation and test set

We then go for two directions. The first is to evaluate the diversity of CGM pattern distribution, which we can consider as a similar idea to the amount of uncertainty in an entire probability distribution of CGM patterns.[8] In this case, we consider using Shannon entropy of returned cluster ids and the average of all pairwise distance of returned cluster centers to represent diversity. The second is to evaluate the similarity between two distributions of CGM patterns. For this task, we decide to evaluate from categorical perspective and numeric perspective. Each test case has a CGM segment of length 7 days. The returned cluster ids and barycenters of a single test case are 1d vector and 2d matrix. And a set of test cases would then have a 2d matrix and a 3d tensor for returned cluster ids and barycenters. For evaluation, we apply MMD(maximum mean discrepancy) to 2d cluster ids to evaluate distribution. For returned 3d cluster centers, we use two ways. The first is to apply PCA with component=1 to 3d tensors to get a 2d matrix. The second is to flatten a 3d tensor to 2d matrix. And we apply kernel MMD as well. And based on the logic of the motif-based algorithm, there might be some daily pattern being classified as “other”. We treat each “other” case as a different, additional pattern. [figure 4]



objectives

- to verify the relationship between diversity of daily CGM patterns and TIR/HbA1C
- to evaluate domain shifting
- to evaluate CGM patterns distribution under different background (ex. Different age groups)

apply single source dataset

objectives

- To find out possible bias hidden inside multi source dataset
- To verify if there are any pattern differences between real healthy data and synthetic healthy data
- To verify the performance on small data sources

apply multi source dataset with synthetic healthy data included

Chunk and calculate TIR in test set based on the defined monitoring span (7 days). Then give each chunked data a multi-label according to following group:

- TIR group (<=50%, >50%, >70%, >90%)
- Sex
- Age group
- Treatment group

[[L11, L12...L17],

...

[Ln1, Ln2...Ln7]]

Returned cluster ids of the test set(2D)

[[C11, C12...C17],

...

[Cn1, Cn2...Cn7]]

Returned cluster centers of the test set(3D)

[C11, C12...C17

...

Cn1, Cn2...Cn7]

Flattened returned cluster centers of the test set(2D)

- To evaluate the diversity of cluster ids: calculate entropy for each subject, then apply two sample tests (we apply Kolmogorov-Smirnov test here)
- To evaluate the diversity of cluster centers: calculate average of pairwise distance of all the center pairs for each subject, then apply Kolmogorov-Smirnov test.
- To evaluate the distribution of cluster ids between two set of samples: apply MMD
- To evaluate the distribution of cluster centers between two set of samples: apply PCA to compress 3D to 2D and then apply MMD
- To evaluate the distribution of flattened cluster centers between two set of samples: apply MMD

Figure 4. process of applying algorithm and pattern evaluation

Experiments

1. motif algorithm

For motif algorithm, we set γ based on the original paper,[4] which is the 20th quantile of all pairwise time series distance. For τ, we experience 3 different values. We choose 2 motif algorithms from these 3 based on unclassified validation times series numbers and motif numbers. [table 2]

2. k medoid algorithm

We initially planned to use the original training set to train k medoid. But it crushed RAM several times so we need to downsize each time series. But then we need to recalculate γ, which is set based on the original time series. Since dynamic time warping(dtw) takes too much time to calculate and we don't have much time to redo everything. We decide to use central limit theorem to estimate γ_{new}, which is the 20th quantile of pairwise dtw distance of the downsized training time series.

We then set k and outlier threshold based on the motif algorithm and the γ_{new}.

Finally, we choose our paired motif and k medoid algorithm based on the unclassified validation numbers and validation set's silhouette score. [table 3]

We choose the final hyperparameter settings based on the results of the multi source dataset. We decide our final τ to be 0.4γ and outlier threshold for k medoid algorithm is $\gamma/2 + \gamma + 0.4\gamma$. Then we train with the same hyperparameters on the single source dataset to get the motif algorithm and k medoid algorithm for the single source dataset. The reason we don't calculate silhouette score for motifs is that we set radius for each motif so there are lots of unclassified samples in the results. The high percentage of unclassified samples may influence the reliability of silhouette scores.

gamma	tau	Motif_num	Unclassified validation	Classified validation
20th quantile of all pairwise distance	0.3* gamma	53	318	3904
ditto	0.4* gamma	28	385	3837
ditto	0.5* gamma	18	503	3719

Table 2. hyperparameters settings for multi source dataset motif algorithm experiment

K	outlier threshold	Nonclassified validation	Silhouette score
53	5.22	5	0.04
28	5.20	3	0.05

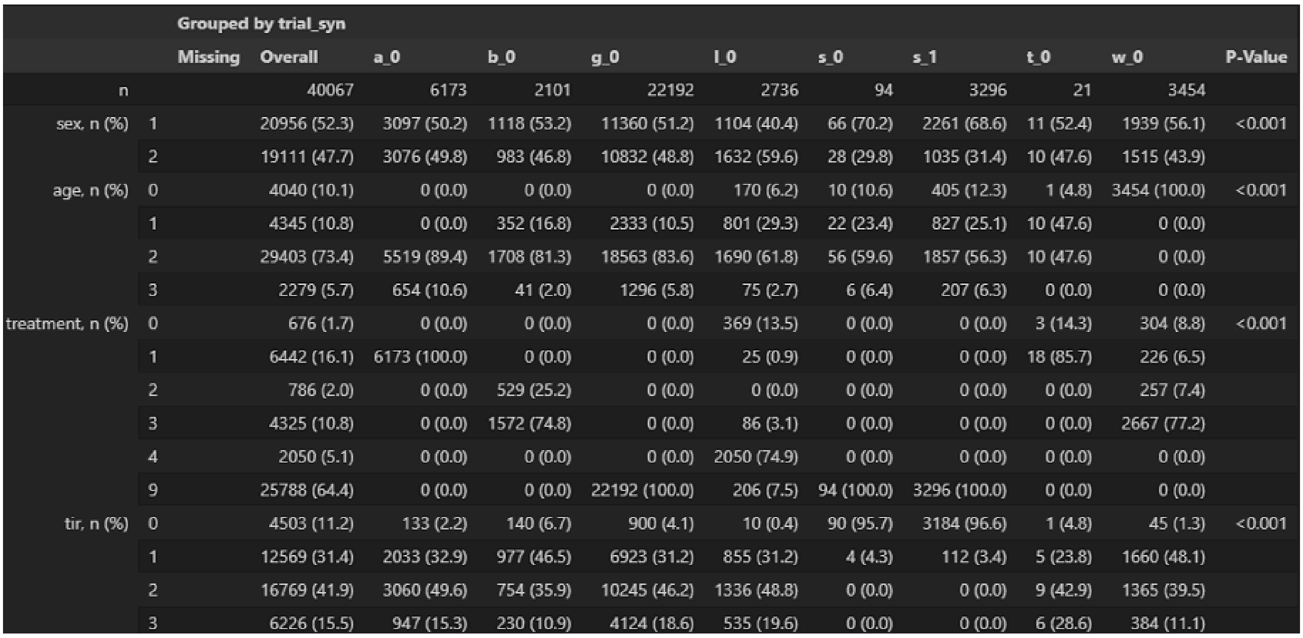
Table 3. hyperparameters settings for multi source dataset k medoid algorithm experiment

Results

Data distribution and Patient Characteristics

We get the final multi source training and validation set with 4222 daily CGM each. The single source test set has 17875 samples, each is aggregated 7-days data. Multi source training and validation sets have 4925 daily CGM each. Multi source test set has 22192 7-days CGM.

For patient characteristics. Because we use stratified splitting in this study. We only show the characteristics table of the test set.[table 4] We can see from the feature table. The distribution of number, age, treatment and tir are very different in most of the trials.



		Grouped by trial_syn										
		Missing	Overall	a_0	b_0	g_0	l_0	s_0	s_1	t_0	w_0	P-Value
n			40067	6173	2101	22192	2736	94	3296	21	3454	
sex, n (%)	1		20956 (52.3)	3097 (50.2)	1118 (53.2)	11360 (51.2)	1104 (40.4)	66 (70.2)	2261 (68.6)	11 (52.4)	1939 (56.1)	<0.001
	2		19111 (47.7)	3076 (49.8)	983 (46.8)	10832 (48.8)	1632 (59.6)	28 (29.8)	1035 (31.4)	10 (47.6)	1515 (43.9)	
age, n (%)	0		4040 (10.1)	0 (0.0)	0 (0.0)	0 (0.0)	170 (6.2)	10 (10.6)	405 (12.3)	1 (4.8)	3454 (100.0)	<0.001
	1		4345 (10.8)	0 (0.0)	352 (16.8)	2333 (10.5)	801 (29.3)	22 (23.4)	827 (25.1)	10 (47.6)	0 (0.0)	
	2		29403 (73.4)	5519 (89.4)	1708 (81.3)	18563 (83.6)	1690 (61.8)	56 (59.6)	1857 (56.3)	10 (47.6)	0 (0.0)	
	3		2279 (5.7)	654 (10.6)	41 (2.0)	1296 (5.8)	75 (2.7)	6 (6.4)	207 (6.3)	0 (0.0)	0 (0.0)	
treatment, n (%)	0		676 (1.7)	0 (0.0)	0 (0.0)	0 (0.0)	369 (13.5)	0 (0.0)	0 (0.0)	3 (14.3)	304 (8.8)	<0.001
	1		6442 (16.1)	6173 (100.0)	0 (0.0)	0 (0.0)	25 (0.9)	0 (0.0)	0 (0.0)	18 (85.7)	226 (6.5)	
	2		786 (2.0)	0 (0.0)	529 (25.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	257 (7.4)	
	3		4325 (10.8)	0 (0.0)	1572 (74.8)	0 (0.0)	86 (3.1)	0 (0.0)	0 (0.0)	0 (0.0)	2667 (77.2)	
	4		2050 (5.1)	0 (0.0)	0 (0.0)	0 (0.0)	2050 (74.9)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	
	9		25788 (64.4)	0 (0.0)	0 (0.0)	22192 (100.0)	206 (7.5)	94 (100.0)	3296 (100.0)	0 (0.0)	0 (0.0)	
tir, n (%)	0		4503 (11.2)	133 (2.2)	140 (6.7)	900 (4.1)	10 (0.4)	90 (95.7)	3184 (96.6)	1 (4.8)	45 (1.3)	<0.001
	1		12569 (31.4)	2033 (32.9)	977 (46.5)	6923 (31.2)	855 (31.2)	4 (4.3)	112 (3.4)	5 (23.8)	1660 (48.1)	
	2		16769 (41.9)	3060 (49.6)	754 (35.9)	10245 (46.2)	1336 (48.8)	0 (0.0)	0 (0.0)	9 (42.9)	1365 (39.5)	
	3		6226 (15.5)	947 (15.3)	230 (10.9)	4124 (18.6)	535 (19.6)	0 (0.0)	0 (0.0)	6 (28.6)	384 (11.1)	

Table 4. features distribution of all sources's test set. For sex, age, treatment and tir features, we use groups we defined before. a_0: Aleppo; b_0:Brown; g_0: Granada; l_0:Lynch; s_0: Shah(real data); s_1: Shah(synthetic data); t_0:Tamborlane; w_0: Wadwa

Task1 : Relationship Between CGM Pattern Diversity and Demographics (Age/TIR)

Hypothesis

This analysis was designed to test the hypothesis that the diversity of daily Continuous Glucose Monitoring (CGM) patterns is related to clinical indicators like Time in Range

(TIR). Specifically, we hypothesized that individuals with lower TIR levels (poorer glycemic control) would exhibit more diverse CGM patterns, while age would have a less significant impact.

Methodology & Conditions

From the single-source dataset, subjects were stratified into four Age groups (0: <10, 1: 10-20, 2: 20-65, 3: ≥65) and four TIR groups (0: >90%, 1: >70%, 2: >50%, 3: ≤50%). To ensure a fair comparison, a balanced number of subjects was sampled from each group (N=23 for Age analysis; N=56 for TIR analysis). Two key metrics were computed for each subject's 7-day CGM segment: **Motif Entropy** (representing pattern diversity) and **Motif Intra-sample Distance**(representing pattern spread). The distributions of these metrics were compared across groups using the Kolmogorov-Smirnov (KS) two-sample test.

Results & Analysis

The analysis revealed distinct outcomes for Age and TIR groups. When analyzed by **Age**, no statistically significant differences were found in the distributions of either Motif Entropy or Motif Intra-sample Distance. All pairwise KS tests yielded p-values well above the 0.05 significance level (e.g., $p > 0.24$ for all comparisons).

In contrast, **TIR** was found to be a primary driver of pattern diversity. A clear trend was observed where lower TIR (poorer glycemic control) correlated with higher entropy and intra-sample distance. The differences were statistically significant, particularly between the most stable group (TIR 0: >90%) and the most unstable group (TIR 3: ≤50%), which showed a KS test p-value of approximately 1.46e-10 for Motif Entropy.

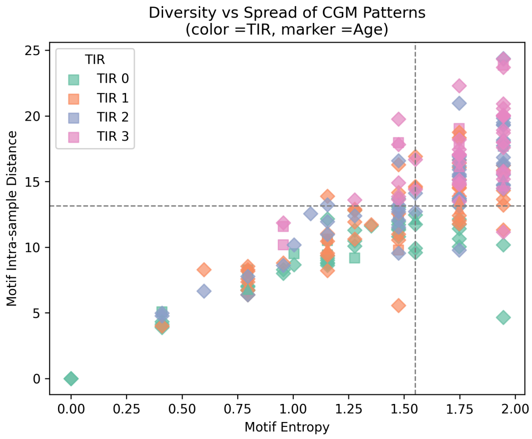


Figure 5: Motif Entropy Distribution by Age and TIR Groups

The top panel shows that the entropy distributions for Age groups 1, 2, and 3 are nearly identical, with overlapping ranges and medians. The bottom panel, however, displays a clear trend for TIR groups: as TIR decreases (from group 0 to 3), the distribution of motif entropy becomes wider and shifts towards higher values, indicating greater pattern diversity in subjects with poorer glycemic control. To further explore the relationship between the two diversity metrics, we visualized each subject as a point in a 2D space defined by entropy and intra-sample distance.

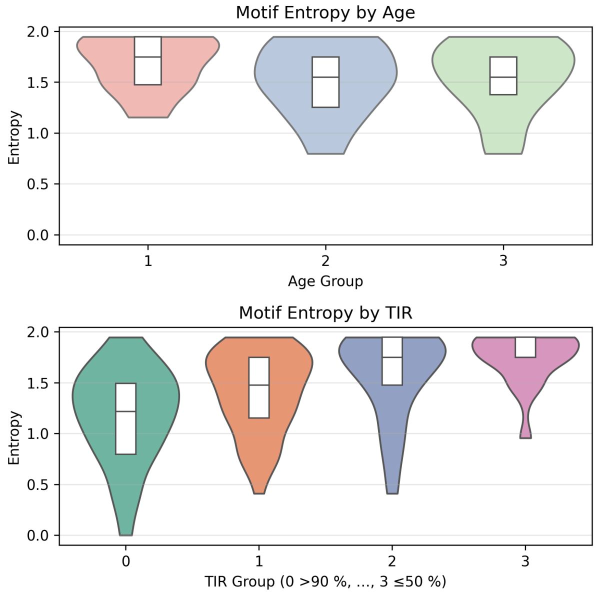


Figure 6: Scatter Plot of Motif Entropy vs. Intra-sample Distance

Each point represents a subject, colored by their TIR group and marked by their Age group. The plot is divided into quadrants by the median values for both metrics. It is evident that subjects in the TIR 3 group (pink diamonds, ≤50%) predominantly occupy the upper-right quadrant, indicating both high pattern diversity and high pattern spread. Conversely, subjects in the TIR 0 group (teal squares, >90%) are concentrated in the lower-left quadrant. The markers for Age groups are scattered across all quadrants, providing no clear clustering. This visualization reinforces that TIR, not age, is the dominant factor influencing these pattern characteristics.

The hypothesis was strongly supported for TIR but not for Age. The analysis concludes that **TIR is a primary determinant of CGM pattern diversity**, while age shows no statistically significant impact in this dataset. Subjects with poor glycemic control (lower TIR) exhibit significantly more diverse and varied CGM patterns.

It is important to note that the lack of significance in the Age analysis could be attributed to the limited statistical power from a modest sample size (N=23 per group).

However, the results for TIR are robust and statistically conclusive, suggesting that glycemic stability itself, rather than age, is the key factor influencing the complexity of daily glucose patterns.

Task 2-1 : Intra- and Inter-Group Pattern Distribution Analysis

In Task 2.1 we were doing intra-group comparison here. Our goal in this task is to see if there is any significant difference in the distribution of each subset of the same subgroup. If there is any difference then we want to see what is the cause of it and where it may lie.

Age Group 200 Samples

In this experiment we choose age group 3 (>= 65 years old) to be evaluated and since we are doing intra-group evaluation each of the classes, Class 0, Class 1 and Class 2 belong to age group 3 (>= 65 years old).

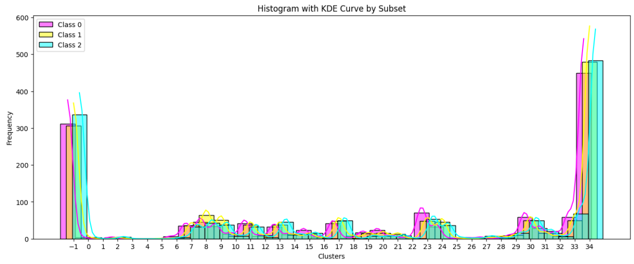


Figure 7: Histogram of Classes by Motifs

The histogram with KDE curves shows the distribution of motif-based features for 200 age group samples across three classes (Class 0, 1, and 2). Most of the features are concentrated at the edges (bins near 0, -1 and 34), with smaller peaks in the middle. Although the KDE curves for each class differ slightly, the overall distributions are very similar, meaning that the motifs are not strongly class-specific. There are only minor variations between classes, suggesting that while some motifs may be more common in one class, they are still present in others.

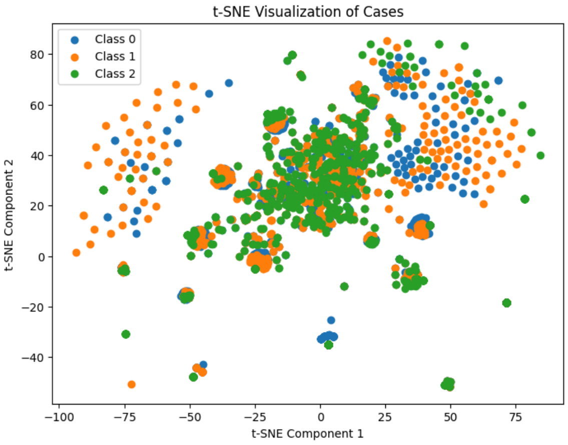


Figure 8: t-SNEof Classes by Motifs

The t-SNE plot shows how these samples group together based on their motif features. Some clusters are visible, with certain regions dominated by one class, but overall, the classes overlap quite a bit. This means the motif features contain some useful information for distinguishing classes, but not enough for clear separation. A more complex model or additional feature processing would likely be needed to improve class prediction.

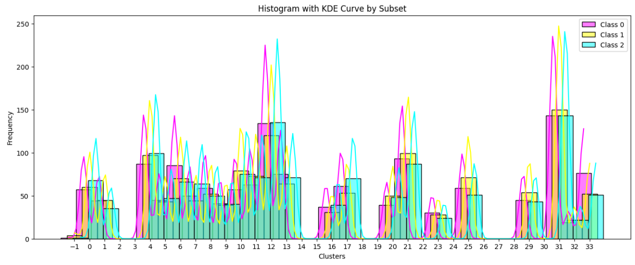


Figure 9: Histogram of Classes by Medoid

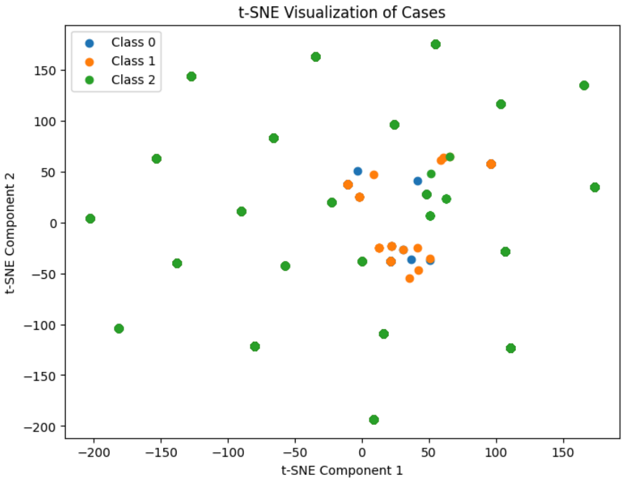
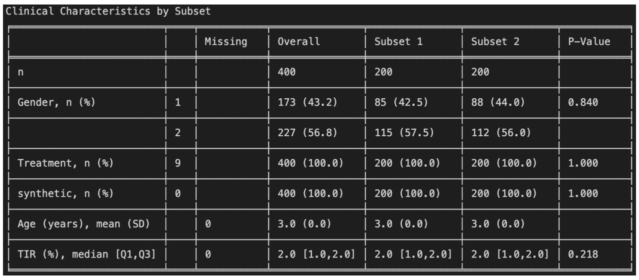


Figure 10: t-SNE of Classes by Medoid

There is no strong, distinct clustering pattern among the classes by medoid method. While some degree of grouping exists—particularly in a few clusters dominated by specific classes—the overall distribution shows significant overlap. This is evident both in the histogram with KDE curves, where different classes share many cluster regions, and in the t-SNE visualization, where class boundaries are not clearly separated. These results suggest that the feature space does not naturally partition the data into well-separated clusters based on class labels, implying that the patterns are somewhat diffuse and not strongly class-discriminative.



```
Clinical Characteristics by Subset

|                        |   | Missing | Overall        | Subset 1       | Subset 2       | P-Value |
|------------------------|---|---------|----------------|----------------|----------------|---------|
| n                      |   |         | 400            | 200            | 200            |         |
| Gender, n (%)          | 1 |         | 173 (43.2)     | 85 (42.5)      | 88 (44.0)      | 0.840   |
|                        | 2 |         | 227 (56.8)     | 115 (57.5)     | 112 (56.0)     |         |
| Treatment, n (%)       | 9 |         | 400 (100.0)    | 200 (100.0)    | 200 (100.0)    | 1.000   |
| synthetic, n (%)       | 0 |         | 400 (100.0)    | 200 (100.0)    | 200 (100.0)    | 1.000   |
| Age (years), mean (SD) |   | 0       | 3.0 (0.0)      | 3.0 (0.0)      | 3.0 (0.0)      |         |
| TIR (%), median [Q1,Q3]|   | 0       | 2.0 [1.0,2.0]  | 2.0 [1.0,2.0]  | 2.0 [1.0,2.0]  | 0.218   |
```

Figure 11: Table One of Group Age 200 Samples

Test Type	Distance / Statistic	p-value
MMD – Motif cluster IDs	0.0113	0.2000
MMD – Motif PCA barycenters	0.0121	0.1090
KS – Motif IDs	0.0243	0.8037
MMD – Motif flatten barycenters	0.0015	0.2990
MMD – K-Medoid cluster IDs	0.0130	0.0590
MMD – K-Medoid PCA barycenters	0.0110	0.2260
KS – K-Medoid IDs	0.0407	0.1963
MMD – K-Medoid flatten barycenters	0.0015	0.2910

Figure 12: Table of Statistical Result

The statistical results indicate that the pattern distributions are broadly similar across the groups, as none of the tests show a significantly low p-value. This suggests there is no strong evidence of differences in motif distributions between the Age classes.

Based on two results we have, sampling by age group using 200 samples the results show there is no significant CGM patterns distribution difference thus the pattern across these subsample are similar. There appears to be a consistent tendency in CGM pattern distributions within this group.

TIR Group 200 Samples

In this experiment, we focus on TIR group 2 (>50%) for evaluation. Since the analysis is conducted within this group, all three classes — Class 0, Class 1, and Class 2 — are subsets drawn from TIR group 2 (>50%).

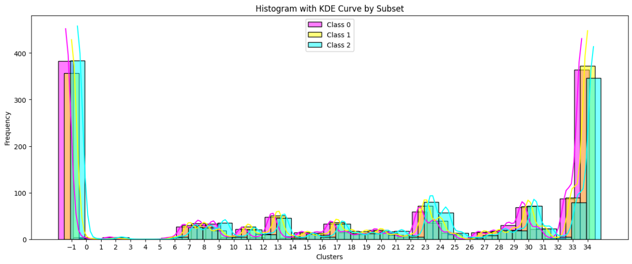


Figure 13: Histogram of Classes by Motifs

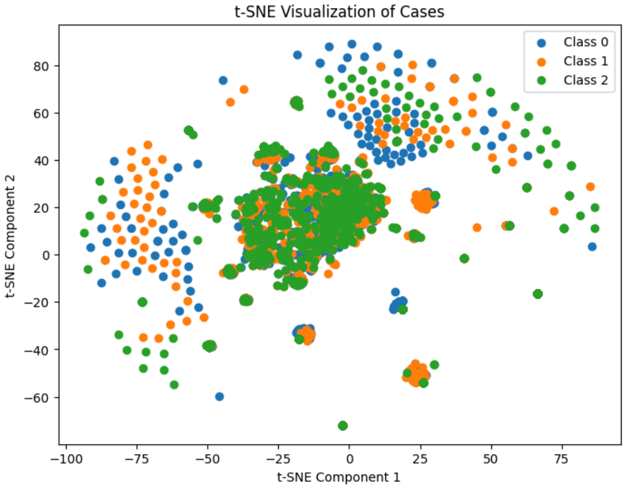


Figure 14: t-SNEof Classes by Motifs

The distribution of motif clusters across the three TIR groups (Class 0, Class 1, and Class 2) appears broadly similar, as shown in both the histogram and t-SNE visualizations. The histogram with KDE curves indicates that all classes share comparable frequency patterns, with prominent peaks around clusters -1 and 34 and smaller peaks across intermediate clusters. This suggests that the occurrence of specific motif clusters is relatively balanced among the TIR groups. Likewise, the t-SNE plot shows a high degree of overlap among the classes in the low-dimensional embedding space, with no clear separation between the clusters of each class. These observations suggest that motif-based features alone do not strongly distinguish the TIR groups, and their distribution is consistent across classes.

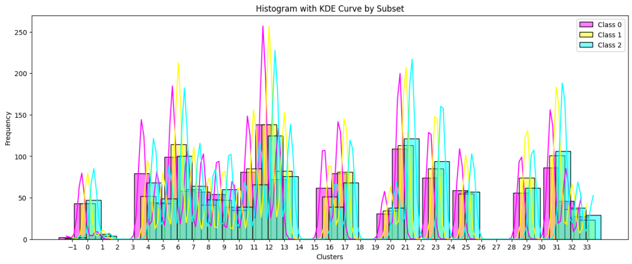


Figure 15: Histogram of Classes by Medoid

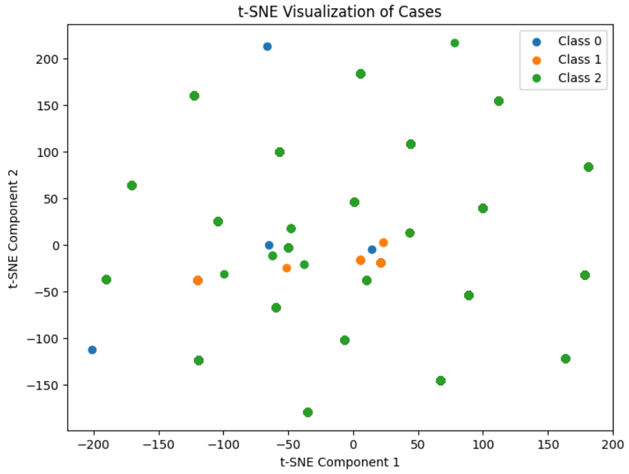
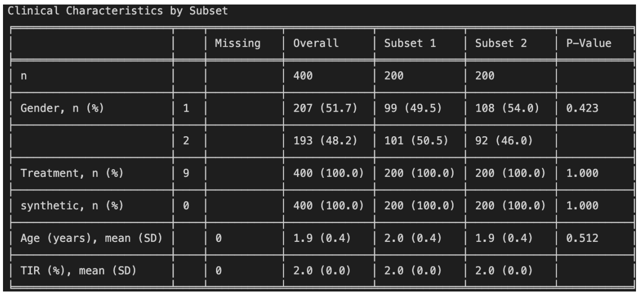


Figure 16: t-SNE of Classes by Medoid

The histogram of motif clusters derived from k-medoids clustering shows that the distribution of clusters is broadly consistent across the three TIR classes. While there are some noticeable differences in peak heights at specific cluster indices (such as clusters 6, 12, 21, and 31), the overall pattern shapes remain similar for Class 0, Class 1, and Class 2. This suggests that, although minor variations exist in how often certain motif clusters occur in each class, there is no strong, distinct clustering pattern that clearly separates the TIR groups. Thus, motif distribution based on medoid clustering appears relatively balanced among the classes. In t-SNE class 2 looks scattered and some of class 0 and class 1 looks close but there is no clear separation between the clusters of each class.



Clinical Characteristics by Subset

		Missing	Overall	Subset 1	Subset 2	P-Value
n			400	200	200	
Gender, n (%)	1		207 (51.7)	99 (49.5)	108 (54.0)	0.423
	2		193 (48.2)	101 (50.5)	92 (46.0)	
Treatment, n (%)	9		400 (100.0)	200 (100.0)	200 (100.0)	1.000
synthetic, n (%)	0		400 (100.0)	200 (100.0)	200 (100.0)	1.000
Age (years), mean (SD)		0	1.9 (0.4)	2.0 (0.4)	1.9 (0.4)	0.512
TIR (%), mean (SD)		0	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)	

Figure 17: Table One of Group TIR 200 Samples

Test Type	Distance / Statistic	p-value
MMD – Motif cluster IDs	0.0100	0.4010
MMD – Motif PCA barycenters	0.0079	0.9620
KS – Motif IDs	0.0271	0.6811
MMD – Motif flatten barycenters	0.0012	0.9270
MMD – K-Medoid cluster IDs	0.0115	0.1620
MMD – K-Medoid PCA barycenters	0.0097	0.5200
KS – K-Medoid IDs	0.0279	0.6492
MMD – K-Medoid flatten barycenters	0.0011	0.9400

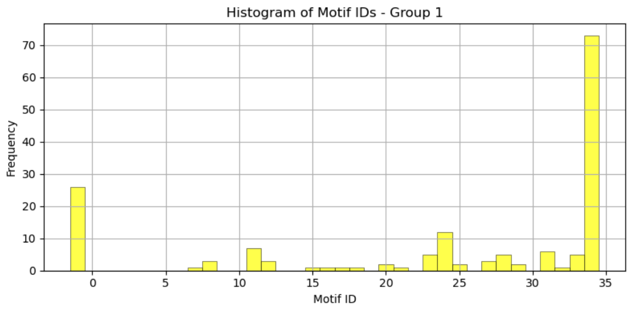
Figure 18: Table of Statistical Result

Based on the statistical test results, the pattern distribution of motif clusters across the TIR groups appears to be **broadly similar**, with **no strong evidence of significant differences** between the groups. Both Maximum Mean Discrepancy (MMD) and Kolmogorov–Smirnov (KS) tests yield **small distance/statistic values** and **high p-values**, indicating that the distributions of motif cluster IDs and their corresponding barycenters (in both PCA and flattened forms) are **not significantly different** between the groups.

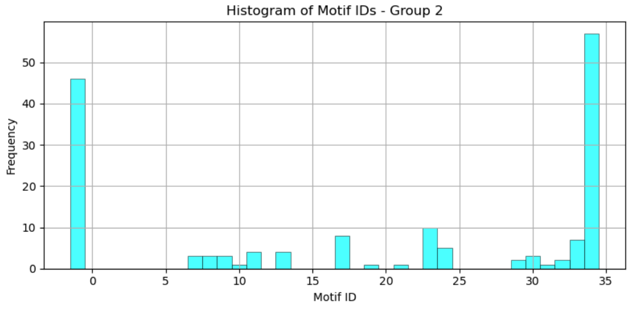
The CGM patterns in this group are stable and similar across different samples. This means the group is likely large enough and consistent, and there is no random variation caused by splitting it into subsets.

Task 2-2 : Intra- and Inter-Group Pattern Distribution Analysis

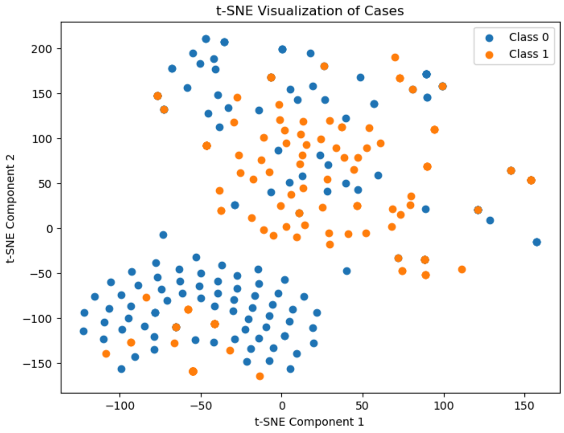
Now let us look at the inter-group comparison. This means we are comparing between different categories such as gender and age and we want to check if there are significant differences between them, creating bias. All of our comparisons will be done between the k-medoid and motif algorithms. Firstly, we will use histograms which tally which cluster ids are most common. These are produced from the motif algorithm or k-medoid algorithm, and each cluster id represents some generalization of patterns in our data. Next, we compare t-SNE results between both methods. Lastly, we use MMD and some p-values to check the statistical soundness of our results



Histogram of Motif IDs - Group 1

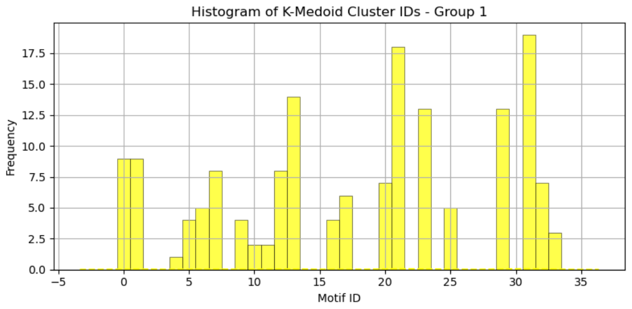


Histogram of Motif IDs - Group 2

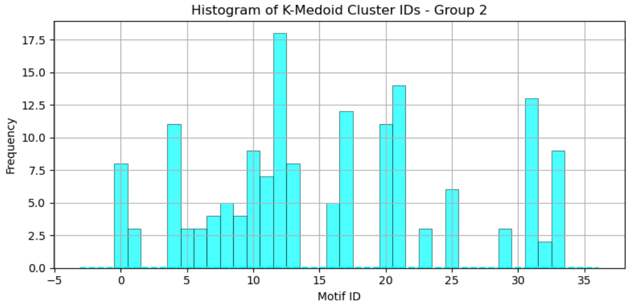


t-SNE Visualization of Cases

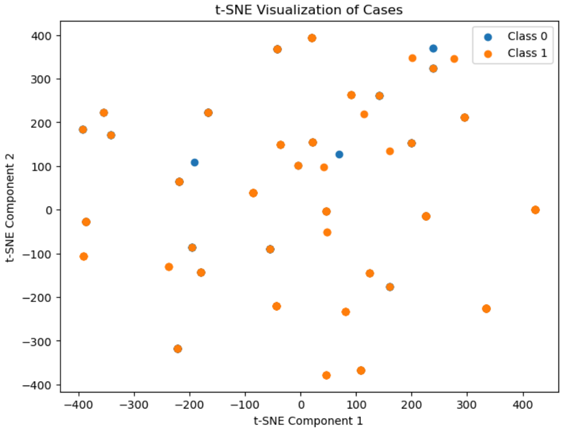
Using the motif method, we can see that both groups are quite similar, especially we can see that -1 (not grouped) and motif ID close to 35 are the most prominent. Additionally, it is quite clear that Class 0 (Age group between 20 and 65) and Class 1(Age group greater than 65) are not easily separable, and thus not dissimilar.



Histogram of K-Medoid Cluster IDs - Group 1

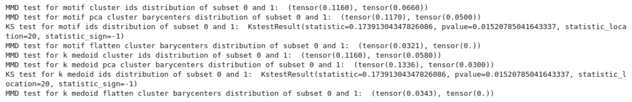


Histogram of K-Medoid Cluster IDs - Group 2



t-SNE Visualization of Cases

However, using the k-medoid method, there is not as clear a similarity, the most prominent features do not seem to match. Still, it is quite clear that Class 0 (Age group between 20 and 65) and Class 1(Age group greater than 65) are not easily separable, and thus not dissimilar.

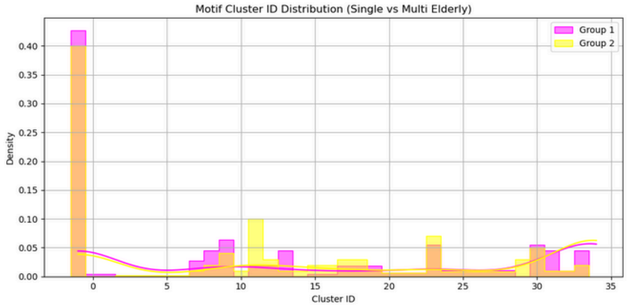


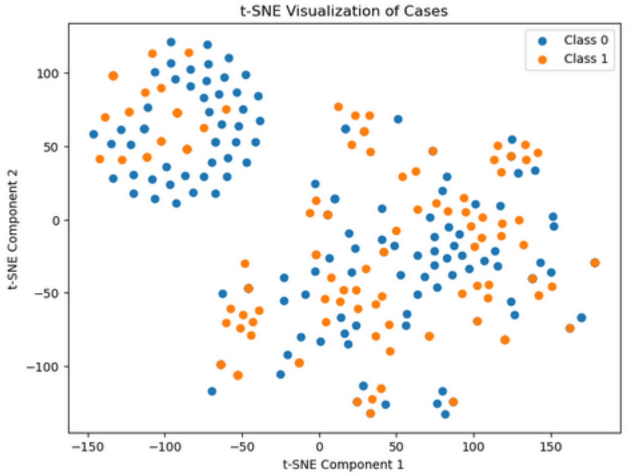
```
MMD test for motif cluster ids distribution of subset 0 and 1:  (tensor(0.1160), tensor(0.0660))
MMD test for motif pca cluster barycenters distribution of subset 0 and 1:  (tensor(0.1170), tensor(0.0500))
KS test for motif ids distribution of subset 0 and 1:  KstestResult(statistic=0.17391304347826086, pvalue=0.0152078504164337, statistic_loca
tion=20, statistic_sign=-1)
MMD test for motif flatten cluster barycenters distribution of subset 0 and 1:  (tensor(0.0321), tensor(0.))
MMD test for k medoid cluster ids distribution of subset 0 and 1:  (tensor(0.1160), tensor(0.0580))
MMD test for k medoid pca cluster barycenters distribution of subset 0 and 1:  (tensor(0.1336), tensor(0.0300))
KS test for k medoid ids distribution of subset 0 and 1:  KstestResult(statistic=0.17391304347826086, pvalue=0.0152078504164337, statistic_l
ocation=20, statistic_sign=-1)
MMD test for k medoid flatten cluster barycenters distribution of subset 0 and 1:  (tensor(0.0343), tensor(0.))
```

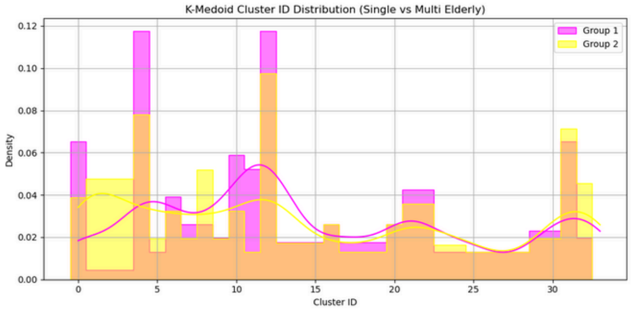
Except for the MMD for motif pca being right at significance (p=0.05) and the MMD for k-medoid ids, everything else points towards a significant difference. This signals that there might be subtle but significant differences in the raw data that are not captured by PCA.

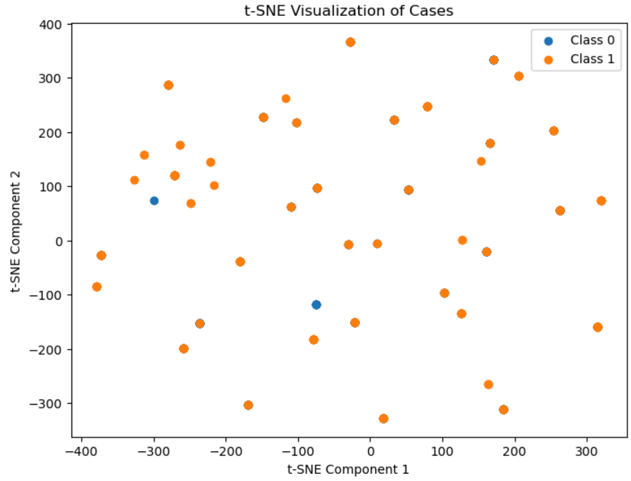
Task 2-3 : Domain Shifting: Single-Source vs. Multi-Source Dataset Comparison

Now that we have checked the characteristics of our intrinsic data versus cross-categorical inter-relationships, we can finally make the comparison we aim to address in this paper; the relationship between datasets. We will conduct the same steps as Task 2-1 and 2-2 with the only difference being the application of the algorithms between different datasets. We will be looking at age group 3 (age greater than 65)

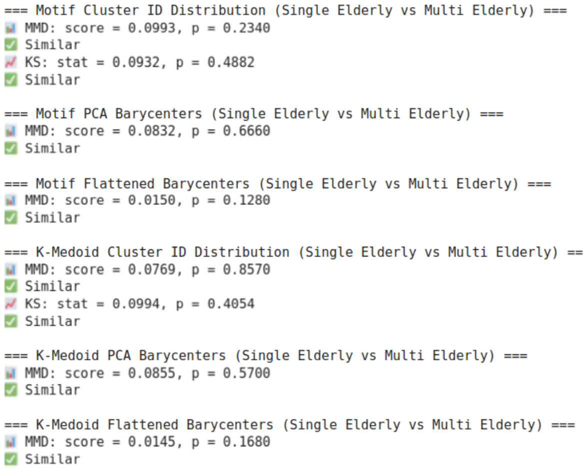






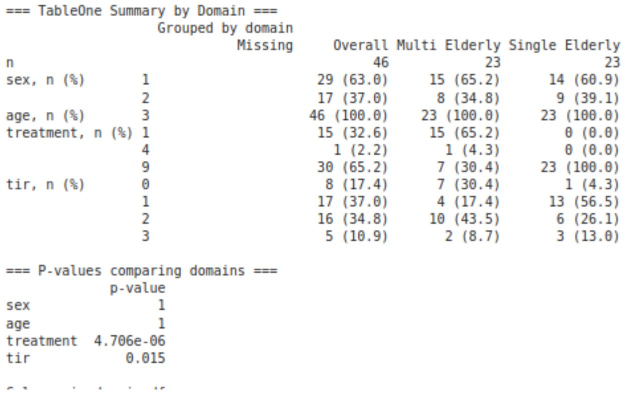


It can be observed that the histograms for both the motifs and the k-medoid have some similarities. The t-SNE like before also seem to have no clear separation, a good argument for non-dissimilarity.



```
=== Motif Cluster ID Distribution (Single Elderly vs Multi Elderly) ===
📊 MMD: score = 0.0993, p = 0.2340
✅ Similar
📈 KS: stat = 0.0932, p = 0.4882
✅ Similar

=== Motif PCA Barycenters (Single Elderly vs Multi Elderly) ===
📊 MMD: score = 0.0832, p = 0.6660
✅ Similar

=== Motif Flattened Barycenters (Single Elderly vs Multi Elderly) ===
📊 MMD: score = 0.0150, p = 0.1280
✅ Similar

=== K-Medoid Cluster ID Distribution (Single Elderly vs Multi Elderly) ===
📊 MMD: score = 0.0769, p = 0.8570
✅ Similar
📈 KS: stat = 0.0994, p = 0.4054
✅ Similar

=== K-Medoid PCA Barycenters (Single Elderly vs Multi Elderly) ===
📊 MMD: score = 0.0855, p = 0.5700
✅ Similar

=== K-Medoid Flattened Barycenters (Single Elderly vs Multi Elderly) ===
📊 MMD: score = 0.0145, p = 0.1680
✅ Similar
```

From our MMD results we can conclude that no statistical significance was found. And can make a strong case for the ability to merge datasets without creating statistical differences.



```
=== TableOne Summary by Domain ===
                   Grouped by domain
                             Missing     Overall Multi Elderly Single Elderly
n                                             46            23             23
sex, n (%)       1                       29 (63.0)     15 (65.2)      14 (60.9)
                 2                       17 (37.0)      8 (34.8)       9 (39.1)
age, n (%)       3                      46 (100.0)    23 (100.0)     23 (100.0)
treatment, n (%) 1                       15 (32.6)     15 (65.2)        0 (0.0)
                 4                         1 (2.2)       1 (4.3)        0 (0.0)
                 9                       30 (65.2)      7 (30.4)     23 (100.0)
tir, n (%)       0                        8 (17.4)      7 (30.4)        1 (4.3)
                 1                       17 (37.0)      4 (17.4)      13 (56.5)
                 2                       16 (34.8)     10 (43.5)       6 (26.1)
                 3                        5 (10.9)       2 (8.7)       3 (13.0)

=== P-values comparing domains ===
             p-value
sex                1
age                1
treatment  4.706e-06
tir            0.015
```

It can be seen that statistical differences were only found between treatment groups and also tir groups for p>0.05

Task 3 : Model Performance Analysis on Small-Scale Trials

Hypothesis

This analysis investigates the hypothesis that clustering algorithms trained on a multi-source dataset may sacrifice smaller clinical trials. We posited that samples originating from trials with fewer subjects would have a higher probability of being left "unclassified" by the models, indicating a potential performance bias related to data source size.

Methodology & Conditions

This experiment utilized the multi-source dataset, which comprises six distinct clinical trials (t, b, w, s, l, a) with subject counts ranging from 20 to 128. Both the Motif and K-medoid models were applied to samples from each trial.

The primary evaluation metric was the unclassified ratio, defined as the proportion of daily CGM patterns assigned a negative cluster ID. This ratio was calculated for each model within each trial. To assess the relationship between trial size and performance, a Spearman correlation test was performed between the number of subjects in a trial and its corresponding unclassified ratio.

Results & Analysis

The performance of the two models varied significantly across the trials. The Motif model exhibited consistently higher unclassified ratios, reaching a peak of 30.0% on the smallest trial ('t', N=20). In contrast, the K-medoid model demonstrated high stability, with a near-zero unclassified ratio across almost all trials and only a minor 5.7% ratio on the same small trial 't'.

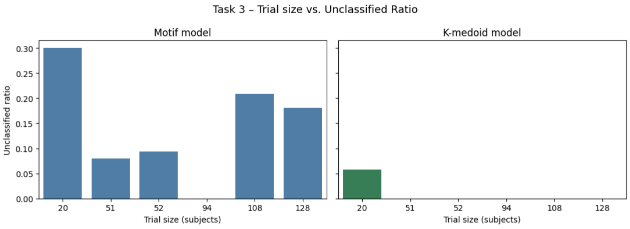


Figure 7: Unclassified Ratio by Trial Size for Motif and K-medoid Models

This chart directly compares the performance of the two models. The Motif model (blue) shows substantial unclassified ratios for most trials, especially the smallest and two of the largest ones. The K-medoid model (green) maintains a flat, near-zero ratio, highlighting its superior stability regardless of trial size.

The Spearman correlation analysis did not find a statistically significant monotonic relationship for either model. For the Motif model, the correlation was negligible ($\rho = -0.09$, $p = 0.87$). For the K-medoid model, a moderate negative trend was observed ($\rho = -0.65$), suggesting that smaller trials might have slightly higher unclassified rates, but this trend did not reach statistical significance ($p = 0.16$). A subsequent permutation test ($p \approx 0.34$) and a failed bootstrap analysis (due to the small number of trials, N=6) confirmed that this lack of significance is likely due to low statistical power.

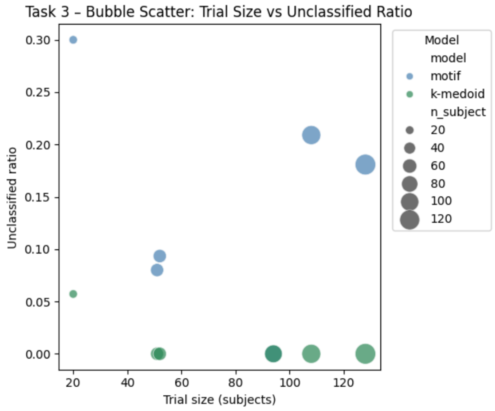


Figure 8: Bubble Scatter Plot of Trial Size vs. Unclassified Ratio.

This visualization reinforces the findings, with bubble size corresponding to the number of subjects. The K-medoid points (green) are clustered along the x-axis at a ratio near zero. The Motif points (blue) are scattered at much higher ratios, visually demonstrating that the practical performance difference between the models is stark, even if the correlation trend is not statistically significant with the available data.

Based on the observed unclassified ratios, the K-medoid model is substantially more robust and reliable than the Motif model when applied to a multi-source dataset with varying trial sizes. While the hypothesis that smaller trials are "sacrificed" was not statistically confirmed with a monotonic correlation, the practical evidence is compelling. The Motif model's high failure rate (up to 30%) on the smallest trial indicates a significant performance liability. In contrast, the K-medoid model's near-perfect classification rate across almost all trials makes it the far more suitable choice for practical applications, as it minimizes data loss and ensures stability regardless of the data source's scale.

Task 4 : Comparison of Real versus Synthetic Healthy CGM Data

Hypothesis

This analysis was designed to test the hypothesis that the patterns and distribution of generated synthetic healthy CGM data are statistically indistinguishable from those of original healthy data. The goal was to validate the practicality of using synthetic data as a substitute for real-world data in CGM research and model training.

Methodology & Conditions

From the multi-source dataset, we compared two groups of healthy subjects from trial 's': 41 subjects with real-world CGM data and 41 subjects with GAN-generated synthetic CGM data. Both the Motif and K-medoid clustering models were applied to 7-day CGM segments from each subject. The comparison was conducted by evaluating cluster ID distributions (MMD, KS tests), pattern space distributions (t-SNE, MMD on barycenters), and summary clinical features (TableOne).

Results & Analysis

The analysis revealed a high degree of similarity between the real and synthetic datasets across multiple dimensions. The statistical tests, summarized in Table 4, confirmed this finding quantitatively.

Test Item	Statistic	p-value	Interpretation
Motif ID Distribution (MMD)	0.057	0.150	No significant difference
Motif PCA Barycenter (MMD)	0.048	0.565	No significant difference
Motif Flatten ID (KS)	D=0.031	0.999	No significant difference
Motif Flatten Barycenter (MMD)	0.006	0.613	No significant difference
K-medoid ID Distribution (MMD)	0.054	0.239	No significant difference
K-medoid PCA Barycenter (MMD)	0.062	0.099	No significant difference
K-medoid Flatten ID (KS)	D=0.031	0.999	No significant difference
K-medoid Flatten Barycenter (MMD)*	NaN	0	No practical difference

Table 4: Statistical Comparison of Real vs. Synthetic Healthy Data Distributions

**Note: A result of 'NaN' with p=0 for the K-medoid flatten barycenter MMD test indicates that the input samples were nearly identical, making the statistic computationally undefined but confirming their similarity.*

As shown in the table, all MMD and KS tests yielded p-values well above the 0.05 significance threshold. This indicates that at the levels of cluster label distribution and overall pattern space, the synthetic data is statistically indistinguishable from the real data.

This statistical similarity is mirrored in the visual analysis. The histograms of cluster IDs (Figure 5) for both models show that the frequency distributions are nearly identical between the real and synthetic groups.

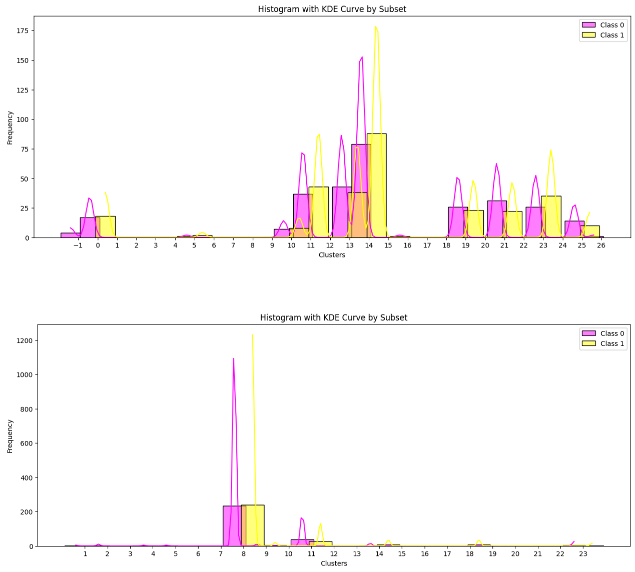


Figure 9: Cluster ID Histograms for Real (0) vs. Synthetic (1) Data. Above / Below : Motif / K-medoid ID Histogram

Furthermore, the t-SNE plots (Figure 6) show that the point clouds for the real and synthetic groups are heavily intermixed, with no clear visual separation. This confirms that they occupy the same pattern space.

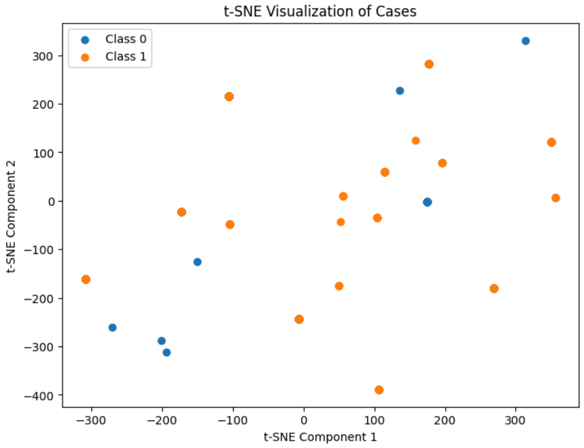
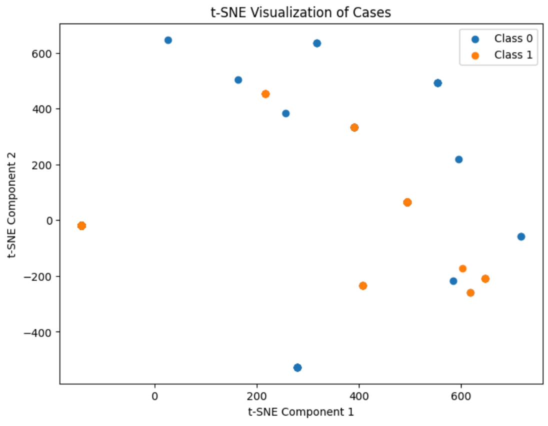


Figure 10: t-SNE Visualization of Barycenters for Real vs. Synthetic Data

Above / Below : Motif / K-medoid tsne

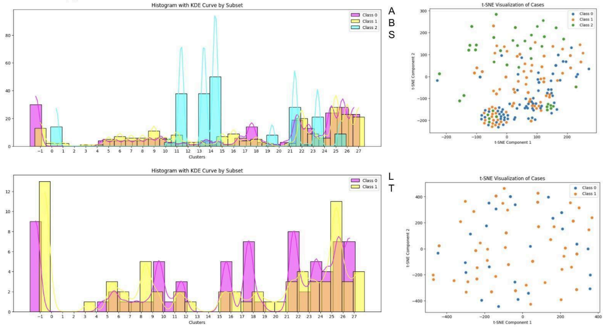


Finally, a comparison of summary clinical features (TableOne) also revealed no statistically significant differences in pattern diversity (entropy) or spread (intra-sample distance), with all p-values greater than 0.1.

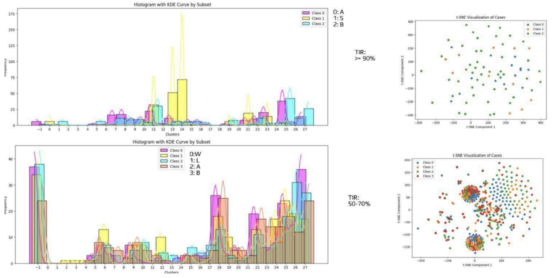
So The hypothesis is strongly supported. The synthetic healthy CGM data is statistically and visually indistinguishable from the real-world healthy data across all tested dimensions. This robust similarity confirms that the synthetic data is a highly practical and reliable substitute for real data, making it suitable for augmenting datasets in CGM research and for training machine learning models.

Task 5 : Inter-Trial Bias Analysis with Matched Demographics

Age < 65



TIR - Sample: 30



Discussions

Motif and Mediod Analysis

Throughout the motif and mediod analysis process, several key insights emerged. First, we determined that mediod analysis was ineffective for capturing meaningful distinctions within or across datasets. In contrast, motif-based clustering successfully highlighted differences in CGM patterns, making it a more reliable method for distinguishing between patient profiles and dataset characteristics.

Our analysis further revealed that demographic factors such as age, distribution were generally standardized within each T1D single-source dataset. However, significant differences in CGM pattern diversity and temporal distribution were evident between healthy and Type 1 diabetes (T1D) datasets. Notably, healthy and unhealthy datasets were statistically different in age, gender and Time in Rate (TIR).

Delving deeper into the T1D cohort, our analysis in Task 1 revealed that Time in Range (TIR) is a primary determinant of this pattern diversity. The results strongly supported that subjects with poor glycemic control (lower TIR) exhibit significantly more diverse and varied CGM patterns, while age was not a statistically significant factor. This suggests that glycemic stability itself, rather than age, is the key factor influencing the complexity of daily glucose patterns.

Cohort Selection

One critical challenge identified was the difficulty of comparing datasets across different treatment protocols. Each single-source dataset represents a distinct population under specific treatment regimens, making cross-dataset comparisons complex and potentially biased. Accurately determining and aligning treatment strategies across datasets proved particularly difficult—especially for less common or more advanced treatment modalities. This limitation underscores the importance of stratified dataset analysis and calls for cautious interpretation of multi-source comparisons in CGM research.

Validation of Synthetic Data

Furthermore, the validation of synthetic CGM data performed in Task 4 confirms its high fidelity and practical utility. The analysis, which compared real and synthetic data across multiple dimensions, found them to be statistically and visually indistinguishable. This robust similarity demonstrates that the synthetic data is a reliable substitute, suitable for augmenting datasets in CGM research and for training machine learning models without introducing significant bias.

Limitation

Our motif algorithm has some limitations. From the design of the algorithm we could know that its basic assumption is that the cluster centers would lie very close to any pair of time series that are close enough. But it may not be true. That could explain the high unclassified rate of validation data, which implies that the chosen motifs may not lie in the centers of the potential clusters.

And because there are many k to choose from in k medoid. We also hoped that the motif algorithm could give us a good guide on the most possible one among all potential k values. But it seems not true. The silhouette score of the k medoid algorithm designed based on the hyperparameters coming from the motif algorithm is not good.

For the evaluation part, although MMD is able to be used in categorical features. But kernel MMD may not. This may influence the reliability of the statistics analysis of the cluster ids distribution.

In contrast to these limitations, Task 3, which evaluated model performance on smaller trials, provided a critical insight into model robustness. Our analysis showed that the K-medoid model is substantially more reliable than the Motif model when applied to a multi-source dataset with varying trial sizes. The Motif model showed a high unclassified ratio (up to 30%) on smaller trials, indicating a significant performance liability. The K-medoid model's near-perfect classification rate ensures data integrity and stability, suggesting it is more suitable for practical applications involving diverse data sources.

Conclusions

We have some findings in our experiments.

1. The diversity of an aggregated 7 days CGM pattern has a negative relationship with TIR.
2. For elderly age group and TIR group >50% and <=70%. We evaluate the dissimilarity of different sample sets sampled from the same mother group. We find out there is no obvious intragroup dissimilarity, indicating possible cluster tendency in these two mother groups.
3. For group age adults and elderlys. There might be some pattern difference between these two groups.
4. We evaluate domain shifting for elderly age groups and we find out there is no significant CGM patterns difference between elderly groups coming from single source dataset and multi source dataset.
5. For smaller size data sources, we find out the motif algorithm yields high unclassified rate while k medoid does not.
6. From the sample set's perspective, the synthetic and real data are indistinguishable.
7. For evaluation on the multi source dataset. We find out despite obvious features differences between different data sources, the pattern distribution between different T1 trials does not show significant difference, while it is significant between the healthy trial and other T1 trials.

Future perspectives

It is needed to further evaluate if there is any cluster tendency in the CGM dataset. Because if it is not, the traditional clustering method may perform badly no matter how we adjust the hyperparameters. We can see from our tSNE results that there might be cluster tendency in CGM time series. We may further apply Hopkins statistics to explore. If there is a cluster tendency, we will further test more k with the k medoid algorithm.

We may also apply PCA with supervised classification based on pseudo labels to explore if there are any basic patterns hidden inside different groups.

Author Contribution

駱書羽

- Final evaluation

- K medoid training and validation

- Generate sex, age, treatment, a1c group columns

- Methods in time_series_augmentation to synthesize cgm datasets

- Give a pesudolabel based on sex/age/treatment/a1c grouping

黃偉祥

- Final evaluation

- Motif training and validation

- Generate sex, age, treatment, a1c group columns

景信勇

- Final evaluation

- Smoothing time series

- Generate sex, age, treatment, a1c group columns

劉天恩

- Final evaluation

- Imputate missing a1c value

- Generate sex, age, treatment, a1c group columns

狄佳多

- Final evaluation

- Training/validation/test splitting

- Generate sex, age, treatment, a1c group columns

申泳

- Final evaluation

- Generate sex, age, treatment, a1c group columns

References

[1] Xinran Xu, Neo Kok, Junyan Tan, Mary Martin, David Buchanan, Elizabeth Chun, Rucha Bhat, Shaun Cass, Eric Wang, Sangaman Senthil, & Irina Gaynanova. (2024). IrinaStatsLab/Awesome-CGM: Updated release with additional public CGM dataset and enhanced processing (v2.0.0). Zenodo. https://doi.org/10.5281/zenodo.14541646

[2] Rodriguez-Leon, C., Aviles Perez, M. D., Banos, O., Quesada-Charneco, M., Lopez-Ibarra, P. J., Villalonga, C., & Munoz-Torres, M. (2023). T1DiabetesGranada: a longitudinal multi-modal dataset of type 1 diabetes mellitus [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10050944

[3] Iwana, Brian Kenji, and Seiichi Uchida. "An empirical survey of data augmentation for time series classification with neural networks." *Plos one* 16.7 (2021): e0254841.

[4] Lobo, Benjamin, et al. "A data-driven approach to classifying daily continuous glucose monitoring (CGM) time series." *IEEE Transactions on Biomedical Engineering* 69.2 (2021): 654-665.

[5] Jacobs, Peter G., et al. "Artificial intelligence and machine learning for improving glycemic control in diabetes: Best practices, pitfalls, and opportunities." *IEEE reviews in biomedical engineering* 17 (2023): 19-41.

[6] 6. Glycemic Goals and Hypoglycemia: Standards of Care in Diabetes—2025 | Diabetes Care | American Diabetes Association

[7] Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range | Diabetes Care | American Diabetes Association

[8] https://nthu-datalab.github.io/ml/slides/03_Probability_Info-Theory.pdf

[9] Kahkoska, Anna R., et al. "Identification of clinically relevant dysglycemia phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c." *Pediatric diabetes* 20.5 (2019): 556-566.

[10] Chan, Nicholas Berin, et al. "Machine Learning–Based Time in Patterns for Blood Glucose Fluctuation Pattern Recognition in Type 1 Diabetes Management: Development and Validation Study." *JMIR AI* 2.1 (2023): e45450.

[11] Shomali, Mansur, et al. "The development and potential applications of an automated method for detecting and classifying continuous glucose monitoring patterns." *Journal of Diabetes Science and Technology* (2024): 19322968241232378.

[12] Staal, Odd Martin, et al. "Kalman smoothing for objective and automatic preprocessing of glucose data." *IEEE journal of biomedical and health informatics* 23.1 (2018): 218-226.