HeartInsightify: Interpreting Longitudinal Heart Rate Data for Health Insights through Conformal Clustering

Prathamesh Dharangutte* Zongxing Xie† Jie Gao* Elinor Schoenfeld† Yindong Hua† Fan Ye†

***Abstract*—Heart rate, a commonly accessible health data from most wearables, carries rich information of a person's well-being, yet remains of limited deep health applications, due to the lack of groundtruth of health events and their impact on heart rate patterns. Specifically, standard health analytics usually are designed based on well-modeled health conditions thus known data patterns and rich training data. To bridge the gap, we propose *HeartInsightify*, an exploratory framework that facilitates the process of deriving health-relevant measurable indicators from longitudinal heart rate data, without any of the above knowledge. *HeartInsightify* focuses on comparative and qualitative study, using model-free statistical methods such as conformal prediction, to study similarities, perform clustering and detect outliers, and build multi-resolutional data summaries, allowing human experts to efficiently examine and verify their health relevance. We conduct extensive experiments to evaluate *HeartInsightify* using individuals' free-living heart rate data collected through Fitbit over 6 years. We illustrate the process of analyzing heart rate data for its health relevance and demonstrate the effectiveness of *HeartInsightify*. We envision that *HeartInsightify* lays the groundwork for personalized health analytics with continuous monitoring data from wearables.**

I. Introduction

Continuous health monitoring [24] with integrated diagnostic devices worn on the body and used in the home holds great potential to identify and prevent early manifestations of diseases [1]. Data before and after health events at home are critical for causality studies, yet they are already gone at the time of a later clinical visit. Continuous monitoring would be instrumental from this perspective, enabling the construction of an individual health profile with which one can make predictions and even take early intervetions.

The prevalence of wearable devices [32], [10] provides great opportunities for continuous monitoring because they are accessible to the general population. In this paper, we specifically focus on heart rate data, which is available in most wearable technologies and has long been used to assess the overall well-being. A number of diseases are associated with changes in resting heart rate, and rapid or gradual changes in heart rate over time. A high resting heart rate has been associated with an increased risk for coronary heart disease, and cardiovascular associated mortality [7], [8], [13]. A study of stroke risk found that for each 10bpm increase in resting heart rate there was a 10% increase in risk for stroke [21]. These studies though longitudinal in nature only obtained measurements at given points in time. We hypothesize that if heart rate data are collected continuously, data evaluation could sooner predict onset of these heart related conditions at a point where they are either preventable or more treatable.

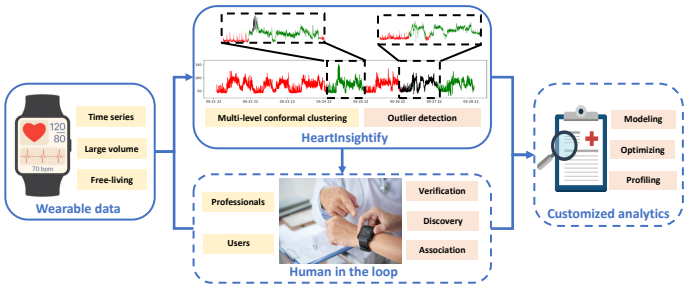


Fig. 1: The workflow of analyzing longitudinal heart rate data for its health relevance. *HeartInsightify* bridges the gap between raw wearable data, human intuition, and customized health analytics. The resulting multi-level clusters and detected outliers preserve human-understandable information, allowing human in the loop to verify and associate the data patterns to potential health conditions.

However, despite some literature showing that various health conditions (physical and/or cognitive, e.g., clinical deterioration [14], stress and depression [31]) exhibit certain heart rate patterns, practical applications of free-living heart rate data for daily health management are still limited. This is because the existence of potential health conditions and their corresponding patterns in the heart rate data are ill-defined or unknown a priori. Such unknowns pose new challenges to the traditional health analytic paradigm where the specific incidents, events and corresponding patterns are known, and sufficient training data exist for development of analytics. What is needed in this situation is reverse engineering longitudinal heart rate to identify the patterns, detect the health incidents, events, thus onset, progression of conditions from daily lives. We summarize intrinsic challenges as follows:

- *Unspecified health relevant patterns to be detected.* The ground truth of when, what health incidents, events exist and how they evolve, is seldom available because even the user cannot fully remember or record them. Due to the general lack of ground truth labels of corresponding conditions in longitudinal heart rate data, existing supervised learning approaches are inapplicable. Moreover, there is a high variation across subjects due to fundamental differences in basic demographics, genetics, health conditions and variations in living environments. Even for the same individual, multiple factors can contribute to the change in trends and patterns in heart rate, such as diseases, drug response, activities, social

*Computer Science, Rutgers University, {ptd39,jg1555}@rutgers.edu
† Stony Brook University, {zongxing.xie, elinor.schoenfeld, yindong.hua, fan.ye}@stonybrook.edu

interactions, sleep stages and even motion artifacts. How all these factors impact and produce exactly what data patterns, is largely unknown.

• *Variation in time scales.* Different health conditions may manifest in noticeable patterns at different time scales. For example, an overall rising trend of resting heart rate may last for weeks during an infection, yet it may fluctuate within hours during different sleep stages. Those happening at fine time granularity may not even be noticeable to the user (e.g., persons with hypertyroid disease have a faster heart rate, and those with hypothyroid disease have a slower heart rate, however, none of them presents noticeable changes at a fine granularity), and how data patterns are impacted are not yet well studied or documented in the literature. The reliability of analysis results cannot be guaranteed in the absence of knowledge regarding the time scale of health-relevant events and the corresponding patterns in the heart rate data, and such uncertainties is further compounded by the inherent variability in heart rate measures.

In this context, we introduce *HeartInsightify* (see Figure 1), an exploratory framework, that facilitates the process of deriving from free-living wearable data, health relevant measurable indicators correlated to unknown, non-predetermined health incidents and events, for practical usefulness regarding health analytics. The development of *HeartInsightify* is driven by three major design choices:

1) *Using model-free data analysis methods.* To potentially accommodate the unknowns in input data, we opt for algorithms that make no assumptions on the input data. Instead, each cluster is represented by a set of "core points" which can be understood as representatives of this cluster. New/incoming data points are compared with the core points of each cluster to decide which cluster they belong. This step is carried out by using a model-free statistical method called conformal prediction, which allows one to test if a new data point belongs to an empirical distribution or not. The core points are updated with new data streams to reflect the current trend, with historical data gradually discounted.

2) *Handling data patterns at varying resolutions.* We develop a multi-level processing framework. At the local level we partition the heart rate data into chunks of a given time window size. We calculate similarities of these pieces by using Dynamic Time Warping metric [19] on time series data (which allows for the shifting/misalignment in the temporal domain and allows for missing data in the input). The cluster name is used as "label" of the chunk, with which we produce a compact representation of the heart beat rate data for a day as a discrete sequence where the alphabet comes from the cluster labels. The compact data representation is crucial for system scalability but also shows to be insightful as data patterns at the 'right' resolution can be distinctive with the unnecessary details hidden/summarized away.

3) *Comparative and qualitative study.* Since ground truth labels or even hypotheses are not available, our focus is on comparative data analysis – finding similarities, dissimilarities, correlations and outliers in the data streams, for the same subject over time or for different subjects living in the same area or the same household. Even without ground truth of 'normal' patterns, segments of repetitive trends are grouped due to circadian rhythm which allows the detection of outliers that may be indicative of potential health relevant conditions for further examination by health professionals (such as disease progression, treatment response, or health outcomes).

We conduct extensive experiments to evaluate *HeartInsightify* using free-living heart rate data from individuals' Fitbit over 6 years. We illustrate the process of analyzing heart rate data for its health relevance and demonstrate that COVID-affected days can be detected at a ratio of 12/16 according to users' annotation. We also demonstrate positive correlation between the scale of data representation and the reliability of analysis results, which necessitates *HeartInsightify* for integrating continuous monitoring data over longer time scales. Although we focus solely on the heart rate data in this study, our method is applicable to time series data of any modality without loss of generality. Our key contributions are:

- We define a new paradigm of analyzing longitudinal heart rate data for health insights, and identify intrinsic challenges primarily from uncertainties of the existence, evolution and extent of potential conditions and their patterns in free-living wearable data.
- We present *HeartInsightify* as an exploratory framework that effectively transforms a large volume of longitudinal data into a manageable set of human-understandable data representations, for incorporating human feedback and integrating with downstream processing models.
- We conduct extensive experiments using individuals' heart rate data collected through Fitbit over 5 years to illustrate the process of analyzing longitudinal heart rate data and demonstrate the effectiveness of *HeartInsightify*.

II. Related Work

Health analytics. There has been a growing interest in health analytics [11] due to the aging population. With large amounts of health data (e.g., MIMIC datasets [12]) becoming publicly available, a substantial body of work in health analytics [30], [29] has emerged, primarily falling into three categories: data mining [46], representation learning [47], [41], and predictive models [45]. Although such studies have shown promising results, they rely on traditional health data (e.g., EHR) collected in sophisticated clinical environments [26], [27], not easily accessible in our day-to-day life.

Health monitoring. To enhance the accessibility of health measurements, researchers (particularly those in the sensing community) proposed a range of sensing technologies for at-home health monitoring. There are mainly two categories: device-based and device-free. In device-based settings, wearable devices have been explored as body area sensors to measure various physiological and/or physical parameters, such as cardiac signals [43], blood pressure [2], respiration [37], [4], jaw motion [33], and gait [15]. In device-free settings, sensors are deployed in the built environments to capture and analyze health relevant information, including vital signs [20], [38],

indoor trajectories [39], body postures [48] and daily activities [23]. Although a rich set of health measurements have been demonstrated feasible and promising, they are usually limited to short-term settings in controlled environments.

Healthcare applications of wearable data. With the prevalence of wearable devices, longitudinal wearable data has become widely available for analyzing health-related factors [32], [10]. Researchers have instantiated the concept of digital phenotyping [22] through wearable data analysis which provides personalized health insights (e.g., sleep quality [28], stress level [31]) and recommendations regarding lifestyle choices [25] (e.g., dietary behaviors [34] and physical activities [3]). Moreover, wearable data has been demonstrated to be indicative of disease progression (e.g., Parkinson [17]) and outcomes [5], [14], thereby facilitating timely diagnosis and medical intervention. While these studies have identified correlations between the wearable data and predetermined health-related factors, discovering unspecified patterns from extensive wearable data still remains an ongoing challenge.

Time series data analysis. Longitudinal wearable data represents a time series. Many methods using deep learning approaches have shown promises for processing time series data [16] and extracting features representative of a certain condition or event [42], [47], [44], however, they lack interpretability, especially when dealing with an open world problem where the patterns of events/conditions to be discovered is unspecified. Change point detection has been extensively studied to identify moments of changes in statistical properties of a time series [35]. However, the detected change points may primarily arise from normal fluctuations in physiological parameters during repetitive daily routines, but provide limited information regarding the temporal trend.

III. An Exploratory Framework: HeartInsightify

A. Problem Formulation

We propose to build compact multi-resolution data summaries for a data stream of health monitoring data. Our design consideration is to prioritize scalability and flexibility with both data modalities and applications. The output (i.e., data summaries) from one scale is the input of the processing in the next (coarser) resolution scale. At each resolution scale, we also allow users/health professionals to examine and provide feedback such as the normal patterns or patterns that would warrant an alert. The parameters in our processing module includes: the resolution scales for which the data summaries need to be computed, for each resolution the choice of algorithms for computing summaries and (when available) patterns of interest to be identified. We discuss the choice of parameters in Section IV. Last the data summaries will be continuously updated with a steam of incoming data.

Below we describe at a high level our approach for processing large-scale streaming health data. The following discussion focuses on the particular data set we work with: heart rate data from fitbit over multiple years. We believe the considerations are applicable to other continuous health monitoring data.

a) Representation and clustering for heart rate data: Raw Fitbit data comprises of sequence of time-stamped heart rate data. For heart rate data, there are two important resolutions that we consider: the heart rate pattern in a day (at smaller time scale) and the daily patterns (days with similar activities). Within a day, the heart rate is heavily influenced by the activity and intensity level. Heart rate during sleep is typically lower on average compared to that when the subject stays awake and heart rate increases dramatically during strenuous activities. Therefore, in our design we first divide the continuous heart rate data into smaller *time segments* and perform clustering on these segments. Next, we build a compact representation for heart rate data at day level using these cluster labels. Henceforth, we refer to this as the *daily representation*, which more concretely represents a day as a vector with entries corresponding to the cluster label of the corresponding time segment. We can now cluster daily patterns using this representation and also use the same idea to cluster longer sequences. Here we remark that this choice of resolution is tailored for heart rate data. Other health signals would reveal different activity patterns and might involve different considerations. For example, blood sugar level is mainly influenced by meal time and intake of food or medication.

b) Identifying change and outliers: With clustering at different scales, we now consider trends or changes at the resolution of days. One pattern of interest at this level is *anomaly detection* – days when the subject deviates from a typical routine (e.g., due to health events), as well as trends that can be identified due to seasonal changes (e.g., holiday break) or changes in living styles (e.g., daily gym time). When data is available from multiple subjects, we would also be able to correlate the identified patterns to detect population level or sub-group level trends. This may reveal differences in both demographics and lifestyles due to culture, geographical location and socio-economical status.

B. Algorithm Design

a) Clustering segments: For clustering segments, we consider each segment to correspond to data for 30 mins. Clustering at this level using smaller length of time segment serves two purposes: 1) Since we are now dealing with smaller time segments, it becomes computationally less extensive for distance computation and comparison, and 2) Similarities of heart rate segments reveal meaningful patterns that correlate with daily activity patterns. Changes or trends in activity level are more informative than changes at the raw time scales.

We use Dynamic Time Warping as the distance measure between time segments, which is widely used in time series applications, and use a model-free method called Conformal DBSCAN [36] for clustering. We refer to [36] for a detailed description of the algorithm and comparisons. At a high level, Conformal DBSCAN first identifies a cluster and compares other segments in order to grow this cluster, one cluster at a time. For each cluster, the algorithm maintains a list of core data points – the data points believed to be in a cluster with high confidence. Throughout the process of clustering,

the algorithm compares new segments to the core data points to determine the cluster membership. For clustering segments of 30 min, core points across different clusters form a natural interpretation - activity patterns of different shape and heart rate levels. Conformal DBSCAN requires a discrepancy measure for which we use sum of distances to k-nearest neighbors.

Implementing Conformal DBSCAN using sum of distances to k-nearest neighbor requires $O(n^2\ell^2)$ time for clustering n points of length ℓ. This computation can be expensive for clustering smaller segments. To overcome this, we modify Conformal DBSCAN to process data in a streaming fashion, which we describe next.

b) Addressing scalability for clustering segments: We first start by clustering a small subset of the continuous stream of data. Then use the core points from these clusters to later assign cluster label to incoming stream of data, while updating the set of core points. We still risk increasing the size of core points, requiring us to compute distance of a new segment with a large set. To overcome this, we limit the the maximum number of core points and adapt a randomized scheme to update the set of core points. Any time the addition of a new segment increases the size of core points beyond this threshold, we randomly select a segment within this set with probability depending on the recency of point. Concretely, we assign a number to data points in order of their arrival. Let i be the number assigned to x which belongs to core points. If the number of core points exceed the threshold, x will be removed from core points with probability proportional to $1/i$. Recall our setup is to deal with continuous time series data, and removing older points from the set of core points has the added benefit of comparing with newer points which will account for small changes in distribution for a cluster.

c) Runtime analysis: To quantify the computational savings of this approach, consider that we run initial clustering on subset of size m from n data points ($m << n$) with segments of length ℓ. Naive implementation scales as $O(n^2\ell^2)$ for distance matrix computation and $O(n^2 \log k)$ for Conformal DBSCAN (updating core points by maintaining distances to k-nearest neighbors can be done for n using min-heap and hash sort in $O(n \log k)$ time). Following this, we need $O(m^2\ell^2 + m^2 \log k)$ for initial clustering of m segments. For clustering remaining segments, setting the limit on number of core points to be a constant adds additional $O(n\ell^2 \log k)$ to runtime. So the total time complexity improvement is from $O(n^2\ell^2 + n^2 \log k)$ to $O(m^2\ell^2 + m^2 \log k + n\ell^2 \log k)$.For small m, run time has linear dependency on the data size.

d) Clustering daily representations: Now for clustering daily patterns, we cluster heart rate data with 24 hr segments as a data point and use cluster labels for representation. Using 30 min segments we represent 24hrs as 48 continuous cluster labels, each corresponding to 30 min segment of the day. This lets us form representation for any arbitrarily long sequence for clustering. We refer to this approach henceforth as *multi-level* clustering. For multi-level clustering, we use a combination of edit distance and DTW as our distance measure. To measure distance between two different days, we compare cluster label for respective 30 min time segment and for those that have different labels we sum up the DTW distance. Comparing at a coarse level, this measure ignores the small changes in similar patterns and focuses on potential change in activities. Note that we can also cluster days by considering 24hr heart rate data, but now the length of segments (ℓ) is large and recall computing DTW scales quadratically with ℓ. We consider this approach for one of our experiments (discussed in detail in Section IV). Note that for clustering longer sequences, one may also use the naive implementation with multi-level representation as with longer segments and fewer data points, computing the distance matrix becomes feasible.

e) Identifying outlier days: In order to identify outlier days, we consider the multi-level representation. Using this representation, we then consider edit distance coupled with a standard outlier detection algorithm (Local Outlier Factor). We discuss our findings in Section IV.

IV. Experiments

We conduct extensive experiments to demonstrate the effectiveness of *HeartInsightify* and illustrate the process of analyzing longitudinal heart rate data for the discovery of unspecified health relevant information. In experiments, we aim to answer following questions:

- **Q1:** How does *HeartInsightify* facilitates making sense of heart rate data for its **health relevance**?
- **Q2:** When integrating continuous monitoring data, how does the scale of data representations impact the **reliability** of analysis results?

A. Experimental Setup

We use heart rate data of two subjects, referenced hereafter as subject 1 and subject 2, over a period of 6 years (2016 to 2021). For majority of our results we focus on the year 2021 for subject 1 and 2019 for subject 2 as they correspond to the time period with most health related events (e.g, Covid) as confirmed by the subjects. Below we report year-wise statistics about heart rate data from Fitbit for subject 1 in Table I.

TABLE I: Subject 1 data statistics (in bpm) across years.

	2016	2017	2018	2019	2020	2021
Mean	74.49	72.76	68.36	71.33	69.66	71.77
Std. Dev	2.06	2.04	1.88	2.32	2.33	1.87
75th percentile	86	83	76	78	76	83

a) Preprocessing: The heart rate data obtained from Fitbit consists of data sampled at uneven frequency and time periods with missing data. In order to deal with these issues, we use standard pre-processing techniques for time series data: linear interpolation and resampling. We resample at 10s interval, and use median for interpolation. This results in $\sim$315k data points for a year. We then group these data points in time segments of 30 min, ending up with $\sim$17.5k data points, each as a vector of dimension 180.

b) Choosing the length of time segment: The choice of length of time segment allows a trade-off between the resolution at which we hope to capture the patterns and the number of data points. For our goal of identifying days that are different from the regular ones, we focus on activities that form a part of individuals routine activities like workout, commuting to work, etc. With this aim, we choose 30 mins as the length of time segment, as this is small enough to capture meaningful variation but long enough to have relatively fewer data points. Note that different analysis warrant different length of time segments. E.g, for trends in sleep patterns and identifying days where changes in sleep patterns occur, one could consider 5 mins segments, focusing only on night time data.

TABLE II: Clustering algorithm parameters.

Parameter	2016	2017	2018	2019	2020	2021
k	15	15	17	14	15	8
ε	0.6	0.7	0.64	0.6	0.45	0.8

c) Choosing the algorithm parameters: For initial clustering, we consider the data for month of January for each year and simulate the rest of data in streaming fashion. Since Conformal DBSCAN does not require specifying number of clusters a priori, we discuss our process for choosing the parameters. Conformal DBSCAN starts by identifying the densest set of data points.We observed that for heart rate data, time segments in this initial set of densest cluster usually corresponded to sleep. As a result, we use the following heuristic: select the parameters resulting in appropriate number of points being clustered for sleep segments. More concretely, expecting an average 7hr of sleep for the month of January, we would expect $\sim$ 400 data points within this cluster. We believe this approach provides us clustering specific to the individual while only requiring a very high level knowledge of sleep pattern for a month. We report the parameters (k : number of neighbors for the discrepancy score measure, and ε : parameter to control growing phase of clusters) used for our experiments in Table II. All our experiments were run on a virtual server with 8 core CPU and 64GB RAM.

B. Sensemaking for Health Relevance (Q1)

For recognizing health related events in a year, our goal is to identify the days that deviate from normal days which are different for different individuals. The multi-scale representation captures activity level information specific to individual. As a result the days detected as outliers have a natural interpretation – days where activities are significantly different from the usual ones for the individual. For detecting outliers we focus on data from 2021 for subject 1 and use $k = 8$ and $\varepsilon = 0.75$ for clustering daily patterns. Figure 2 shows representative samples of daily-basis data representation corresponding to each day-wise cluster. The color code is used to visualize the distribution of local temporal patterns encoded by the corresponding cluster assignments. The resulting data representation accommodates both short and long time scales, and is indicative of nuanced changes in health status. The outlier

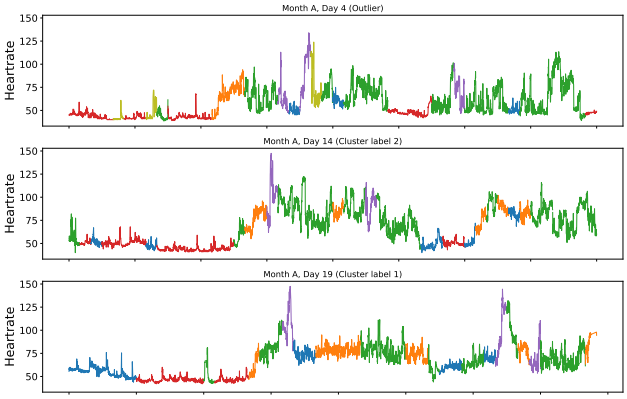


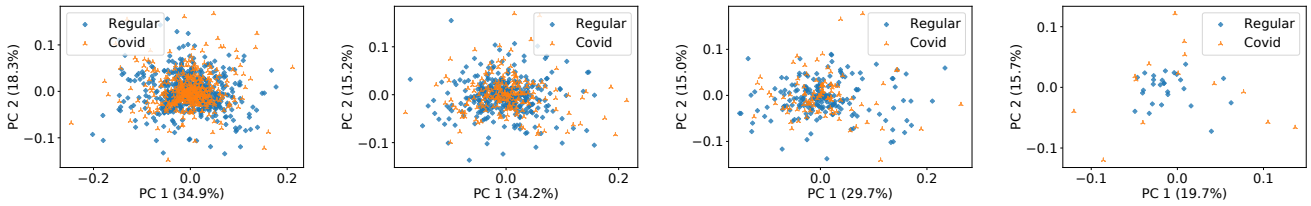
Fig. 2: Representative samples of daily-basis data representation corresponding to each day-wise cluster.

day (first row) has longer sleep like pattern (cluster colored red) during mid-day which is different from other clusters for daily patterns (row 2, 3).

As we do not have access to labeled data, the outliers detected were manually labeled by subject 1 as part of the human evaluation. From among the days that were detected as outliers the noteworthy days confirmed by human evaluation corresponded to a significant health event and the week individual contracted Covid. We also observe sudden increase in the number of outlier days detected for the month of November 2021 and December 2021 which was confirmed as the time period with health events. Apart from health related events, other labels provided by human annotations for days detected as outliers consisted of "*very sedentary*", "*day with change in workout time*" and "*recovery*" which corresponded to restful days after a health event.

In order to understand how good multi-level representation is for detecting outliers, we examine days considered as outliers by running Conformal DBSCAN of 1-day segment of raw data (data points not clustered are treated as outliers). With smaller time segments, 12 of 16 Covid days were identified as outliers whereas multi-level representation identified 2. On the other hand, smaller time segments only identified 6 of 17 recovery days as outliers whereas multi-level representation identified 12. Out of 45 outlier days for smaller time segments, 14 were false positives and of 53 outlier days for multi-level representation, 21 were false positives. As the choice of representing same data is different for these approaches, the reason for a day being considered outlier may be different.

1) Remarks: The outliers suggested by our method are reasonably reliable for distinguishing Covid-affected days from regular days, and can be further combined with heuristics about continuity in the Covid period to identify a complete period. In contrast, we observe a relatively low detection rate for the days annotated as "recovery days", during which the subject would stay more "sedentary" than usual. The low detection rate of "recovery days" is due to the fact that the difference between regular and recovery days is nuanced. This, however, is also fixable when combined with simple heuristics. The false positive rate achieved in this implementation is



(a) 15-min (b) 30-min (c) 45-min (d) Daily-basis

Fig. 3: Visualizations of different data representations through 2D PCA projections.

below $1/3$, and it is tunable based on the significance level configured in conformal prediction framework to be updated in the loop according to trade offs in the downstream analysis.

C. Integrating Data for Reliability (Q2)

In the longitudinal heart rate data, there exist various sources of variability that may impact of reliability of analysis results. For example, short-scale measures may vary a lot during the day depending on daily routine activities at different times. Motion artifacts may also contribute to the variability/noise to heart rate data. We conduct experiments to demonstrate that long-scale data presentation through multi-level data integration is **more resilient to variability thus reliable** than short-scale data representation.

1) Discriminating between regular and Covid periods: We first evaluate the discriminative power using different data representations in distinguishing between regular days and Covid-affected days. Empirically, we choose $15, 30, 45$ minutes (which fall within the typical range of the clinical visit duration) as short-scale data representations, to preserve local time-series patterns.

TABLE III: Wilcoxon rank sum test (Regular vs. Covid).

Data representation	Statistical significance	Effect size
15-min	****	0.1173
30-min	****	0.1236
45-min	****	0.1302
daily	****	1.1277

Note: the symbols *, **, ***, **** denote the statistical significance at 10%, 5%, 1%, and 0.1% levels, respectively.

Figure 3 shows the 2D PCA projections [9] of different representations of longitudinal heart rate data during regular days and Covid-affected days. It visually shows that daily-basis data representation is more discriminative than short-period ones. We further quantify the discriminative power of the representation based on the observation that the data from the same category usually has a similar distribution of intra-group pairwise distances, and vice versa. We contrasting distributions of pairwise distances among regular and Covid-affected days using respective data representations, and DTW is used to measure distance between time series data. We apply Wilcoxon rank sum test (also known as the Mann-Whitney U test, which does not assume a specific distribution for the data) to verify if there is significant difference between the distributions of regular days and Covid-affected days (see Table III). Surprisingly, the distribution differences between regular and Covid-affected days using different data representations are all comparably significant according to the p-value, be it short-scale (15, 30, 45 minutes) or daily-basis data representation. While p-value does not show much difference between different data representations, the effect size quantified via Cohen's d [40] shows that using daily-basis data representation is more discriminative than short-scale (15, 30, 45 minutes) data representations.

2) Clustering and outlier detection: Next, we assess the reliability of the framework for clustering and outlier detection. Variability in heart data data stems mainly from two sources that may impact the analysis results: motion artifacts that add Gaussian noise to heart rate measures [18], and uncertainty in the personal schedule of physical activities during a day that may change the order of certain heart rate trends. To evaluate the reliability of our framework, we add artificial perturbations according to the two sources of variability in real data.

TABLE IV: Normalized Mutual Information between original and perturbed clustering results.

σ	NMI (30-min)	NMI (1-day, multi-level)
0.1	0.7666	0.8027
0.5	0.7019	0.7443
1	0.6972	0.6816

In Table IV, we compare the reliability of clustering results between using 30-min representation and 1-day representation. We quantify the reliability of clustering results according to Normalized Mutual Information score (NMI), which ranges from 0 to 1; "0" means no agreement with original results, indicating fully skewed, while "1" indicates full agreement [6]. For clustering time segments, we add Gaussian noise of different magnitudes (standard deviation $\sigma = 0.1, 0.5, 1$). For multi-level clustering, Gaussian noise is added in obtaining the initial cluster labels of 30 min segments, which are then used to represent 1-day time segment. Then we performing clustering on this representation of 1-day segments. The results show that integrating data into 1-day time segments in a multi-level setting is more resilient to the perturbations than using the 30-min one, as the impact of perturbations is confined in the local scale when the Gaussian noise is limited ($\sigma \leq 0.5$).

Furthermore, we aim to understand how much perturbation in the local patterns (represented and encoded by clustering

TABLE V: The relation between the fraction of flip in the data representation and detection as outlier.

Fraction of flip	Ratio of becoming outlier
3/48	15%
6/48	54%
12/48	93%

results) will cause 1-day segments be skewed and detected as outlier. Towards this, we start by selecting days randomly from the year that were not recognized as outliers. For these days, we randomly select and change their local cluster assignments to cluster labels other than the current ones. We then run outlier detection algorithm for data corresponding to the entire year (using same parameters). We wish to observe how much perturbation (e.g., the ratio of cluster assignments that are flipped) is needed before a particular day is considered to be an outlier. To this extent, we select 3, 6 and 12 labels out of 48 in a 1-day data representation to flip (corresponding to 6.25%, 12.5% and 25%, respectively). Averaging over repeated experiments, with same parameters for the outlier detection algorithm across all runs, we report the ratio of days randomly selected ones that were detected to be outliers (Table V). We observe that flipping 25% of local labels in a daily-basis data representation (corresponding to 6hrs) would cause the majority (93%) of normal samples to become outliers.

D. Correlation analysis

Access to longitudinal heart rate data for multiple subjects provides further opportunities for investigating interesting patterns at group and population level. As an elementary study in this direction, we try to investigate the cross correlation of heart rate date across two cohabiting subjects, comparing the heart rate data month wise for two cohabiting subjects. For comparison, we also perform cross correlation on the data represented using cluster label for 30 min segments (we use $k = 12$ and $\varepsilon = 0.9$ for clustering subject 2's data).

For evaluation, we obtain feedback from the subjects on pattern we expect to observe, according to which we expect to observe months of February and March to be relatively similar (in terms of activities and daily patterns) while May and September to be different. We focus on correlation values for these months due to space constraints.

TABLE VI: Cross correlation between subjects

	February	March	May	September
Heart rate data	0.22	0.37	0.44	0.42
30 min segments	0.33	0.30	0.29	0.16

As observed, the cross correlation between sequences better aligns with human feedback with cluster labels as representation rather than raw data. We attribute this to the fact that data itself is at finer level and noisy whereas representation using 30 min segments enables us to capture details at a coarser level, enabling us to focus on trends and changes in a more meaningful way.

1) Remarks: In this case study, the long-scale data representation (e.g., daily basis) shows better consistency (compactness) among the same period of a certain condition (e.g., Covid), and is discriminative from the non-Covid period. In contrast, short-period data representation varies largely over a day, subject to condition-irrelevant factors, not discriminative between Covid and non-Covid periods. We conclude that long-scale representation is more resilient to perturbations than short ones as the impact of perturbations is confined to local.

V. Conclusion and Future Work

This work is a starting point towards an ecosystem that exploits continuous health data collected through wearable devices for health monitoring and management. The recent breakthrough in machine learning techniques and large-scale health data collection through wearable devices have demonstrated feasibility of using such data for well defined events/activities/health conditions (e.g., [15], [28]). But extending such efforts for unknown, diverse health conditions without high quality ground truth label faces a major gap. Solving this issue requires the inclusion of both users and health professional in the loop to identify, define and label patterns that are of interest to the subject. This work provides basic data processing and prepares data representations that facilitate the discovery and confirmation of interesting patterns, and is on the trajectory of eventually fully exploiting wearable measurements for health benefits. Therefore the work in this paper is considered as complementary and instrumental to existing literature.

There are a number of opportunities for future work including and not limited to the following directions: 1) our design choice is empirically motivated by the particular Fitbit data set and guided by health professionals' recommendations. In general, we would like to develop a set of metrics to evaluate different data analysis methods, including user studies with health professionals in the loop to evaluate the identification of interesting short-term and long term patterns; 2) study a larger pool of subjects and correlate data analysis of interesting groups (individuals within the same demographic group, geographical location, profession, or with similar health concerns) to reveal new discoveries on commonalities and individualities; 3) apply the framework to a diverse set of health related signals for multi-modality analysis.

VI. Acknowledgements

Jie Gao and Prathamesh Dharangutte would like to acknowledge support from NSF CCF-2118953. Fan Ye, Elinor Schoenfeld, Zongxing Xie and Yindong Hua would like to acknowledge support from NSF-2119299.

References

[1] G. Appelboom, E. Camacho, M. E. Abraham, S. S. Bruce, E. L. Dumont, B. E. Zacharia, R. D'Amico, J. Slomian, J. Y. Reginster, O. Bruyère, et al. Smart wearable body sensors for patient self-assessment and monitoring. *Archives of public health*, 72(1):1–9, 2014.

[2] N. Bui, N. Pham, J. J. Barnitz, Z. Zou, P. Nguyen, H. Truong, T. Kim, N. Farrow, A. Nguyen, J. Xiao, et al. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *Mobicom*, pages 1–17, 2019.

[3] M. A. Case, H. A. Burwick, K. G. Volpp, and M. S. Patel. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Jama*, 313(6):625–626, 2015.

[4] A. Chara, T. Zhao, X. Wang, and S. Mao. Respiratory biofeedback using acoustic sensing with smartphones. *Smart Health*, 28:100387, 2023.

[5] H. Cos, D. Li, G. Williams, J. Chininis, R. Dai, J. Zhang, R. Srivastava, L. Raper, D. Sanford, W. Hawkins, et al. Predicting outcomes in patients undergoing pancreatectomy using wearable technology and machine learning: prospective cohort study. *JMIR*, 23(3):e23595, 2021.

[6] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2):189–201, 2009.

[7] R. F. Gillum, D. M. Makuc, and J. J. Feldman. Pulse rate, coronary heart disease, and death: the NHANES I epidemiologic follow-up study. *Am. Heart J.*, 121(1 Pt 1):172–177, Jan. 1991.

[8] P. Greenland, M. L. Daviglus, A. R. Dyer, K. Liu, C. F. Huang, J. J. Goldberger, and J. Stamler. Resting heart rate is a risk factor for cardiovascular and noncardiovascular mortality: the chicago heart association detection project in industry. *Am. J. Epidemiol.*, 149(9):853–862, May 1999.

[9] G. Grinstein, M. Trutschl, and U. Cvek. High-dimensional visualizations. In *KDD - Visual Data Mining Workshop*, page 120, 2001.

[10] J. L. Hicks, T. Althoff, R. Sosic, P. Kuhar, B. Bostjancic, A. C. King, J. Leskovec, and S. L. Delp. Best practices for analyzing large-scale health data from wearables and smartphone apps. *NPJ digital medicine*, 2(1):45, 2019.

[11] K. Jin, J. W. Simpkins, X. Ji, M. Leis, and I. Stambler. The critical need to promote research of aging and aging-related diseases to improve health and longevity of the elderly population. *Aging and disease*, 6(1):1, 2015.

[12] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[13] W. B. Kannel, C. Kannel, R. S. Paffenbarger, Jr, and L. A. Cupples. Heart rate and cardiovascular mortality: the framingham study. *Am. Heart J.*, 113(6):1489–1494, June 1987.

[14] D. Li, J. Vaidya, M. Wang, B. Bush, C. Lu, M. Kollef, and T. Bailey. Feasibility study of monitoring deterioration of outpatients using multimodal data collected by wearables. *ACM Transactions on Computing for Healthcare*, 1(1):1–22, 2020.

[15] H. Li, H. Chen, C. Xu, Z. Li, H. Zhang, X. Qian, D. Li, M.-c. Huang, and W. Xu. Neuralgait: Assessing brain health using your smartphone. *Proceedings of IMWUT*, 6(4):1–28, 2023.

[16] X. Li, V. Metsis, H. Wang, and A. H. H. Ngu. Tts-gan: A transformer-based time-series generative adversarial network. In *AIME 2022*, pages 133–143. Springer, 2022.

[17] L. Lonini, A. Dai, N. Shawen, T. Simuni, C. Poon, L. Shimanovich, M. Daeschler, R. Ghaffari, J. A. Rogers, and A. Jayaraman. Wearable sensors for parkinson's disease: which data are worth collecting for training symptom detection models. *NPJ digital medicine*, 1(1):64, 2018.

[18] Y. Maeda, M. Sekine, and T. Tamura. Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography. *Journal of medical systems*, 35:969–976, 2011.

[19] A. Mueen and E. Keogh. Extracting optimal performance from dynamic time warping. In *Proc. 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2129–2130, 2016.

[20] P. Nguyen, X. Zhang, A. Halbower, and T. Vu. Continuous and fine-grained breathing volume monitoring from afar using wireless signals. In *IEEE INFOCOM 2016*, pages 1–9. IEEE, 2016.

[21] W. T. O'Neal, W. T. Qureshi, S. E. Judd, J. F. Meschia, V. J. Howard, G. Howard, and E. Z. Soliman. Heart rate and ischemic stroke: the REasons for geographic and racial differences in stroke (REGARDS) study. *Int. J. Stroke*, 10(8):1229–1235, Dec 2015.

[22] J.-P. Onnela. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*, 46(1):45–54, 2021.

[23] S. Pan, M. Berges, J. Rodakowski, P. Zhang, and H. Y. Noh. Fine-grained recognition of activities of daily living through structural vibration and electrical sensing. In *the 6th Buildsys*, pages 149–158, 2019.

[24] A. Pantelopoulos and N. G. Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(1):1–12, 2009.

[25] M. S. Patel, D. A. Asch, and K. G. Volpp. Wearable devices as facilitators, not drivers, of health behavior change. *Jama*, 313(5):459–460, 2015.

[26] V. Sachdev, Y. Gu, J. Nichols, W. Li, S. Sidenko, D. Allen, C. Wu, and S. L. Thein. A machine learning algorithm to improve risk assessment for patients with sickle cell disease. *Blood*, 134:893, 2019.

[27] V. Sachdev, X. Tian, Y. Gu, J. Nichols, S. Sidenko, W. Li, A. Beri, W. A. Layne, D. Allen, C. O. Wu, et al. A phenotypic risk score for predicting mortality in sickle cell disease. *British journal of haematology*, 192(5):932–941, 2021.

[28] A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, S. Taheri, et al. Sleep quality prediction from wearable data using deep learning. *JMU*, 4(4):e6562, 2016.

[29] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE JBHI*, 22(5):1589–1604, 2017.

[30] D. Shillan, J. A. Sterne, A. Champneys, and B. Gibbison. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical care*, 23:1–11, 2019.

[31] E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine*, 1(1):67, 2018.

[32] M. Smuck, C. A. Odonkor, J. K. Wilt, N. Schmidt, and M. A. Swiernik. The emerging clinical role of wearables: factors for successful implementation in healthcare. *NPJ Digital Medicine*, 4(1):45, 2021.

[33] T. Srivastava, P. Khanna, S. Pan, P. Nguyen, and S. Jain. Muteit: Jaw motion based unvoiced command recognition using earable. *Proc. ACM on IMWUT*, 6(3):1–26, 2022.

[34] M. Sun, J. D. Fernstrom, W. Jia, S. A. Hackworth, N. Yao, Y. Li, C. Li, M. H. Fernstrom, and R. J. Sclabassi. A wearable electronic system for objective dietary assessment. *Journal of the American Dietetic Association*, 110(1):45, 2010.

[35] C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[36] H. Wang, J. Gao, and M.-g. Xie. Clustering of trajectories using non-parametric conformal dbscan algorithm. In *IPSN 2022*. IEEE, 2022.

[37] X. Wang, R. Huang, C. Yang, and S. Mao. Smartphone sonar-based contact-free respiration rate monitoring. *ACM Transactions on Computing for Healthcare*, 2(2):1–26, 2021.

[38] X. Wang, C. Yang, and S. Mao. On csi-based vital sign monitoring using commodity wifi. *ACM Transactions on Computing for Healthcare*, 1(3):1–27, 2020.

[39] S. Wen, Z. Ge, D. Yuan, Y. Chen, F. Wen, J. Xu, and W. Guan. Enhanced pedestrian navigation on smartphones with vlp-assisted pdr integration. *IEEE Sensors Journal*, 2023.

[40] R. R. Wilcox and T. S. Tian. Measuring effect size: a robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics*, 38(7):1359–1368, 2011.

[41] J. Wu, X. Ye, C. Mou, and W. Dai. Fineehr: Refine clinical note representations to improve mortality prediction. *arXiv preprint arXiv:2304.11794*, 2023.

[42] X. Wu, C. Huang, P. Robles-Granda, and N. V. Chawla. Representation learning on variable length and incomplete wearable-sensory time series. *ACM TIST*, 13(6):1–21, 2022.

[43] C. Xu, T. Chen, H. Li, A. Gherardi, M. Weng, Z. Li, and W. Xu. Hearing heartbeat from voice: Towards next generation voice-user interfaces with cardiac sensing functions. In *Sensys*, pages 149–163, 2022.

[44] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web*, pages 351–360, 2017.

[45] X. Ye, J. Wu, C. Mou, and W. Dai. Medlens: Improve mortality prediction via medical signs selecting and regression interpolation. *arXiv preprint arXiv:2305.11742*, 2023.

[46] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36:2431–2448, 2012.

[47] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang. Wave2vec: Deep representation learning for clinical temporal data. *Neurocomputing*, 324:31–42, 2019.

[48] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on CVPR*, pages 7356–7365, 2018.