Adaptive multi-channel event segmentation and feature extraction for monitoring health outcomes

Xichen She[*1], Yaya Zhai[*2], Ricardo Henao[3], Christopher W. Woods[3], Christopher Chiu[4], Geoffrey S. Ginsburg[3], Peter X.K. Song[1], and Alfred O. Hero *Fellow, IEEE*[5]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109
[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109
[3]Center for Applied Genomics and Precision Medicine, Duke University, Durham, NC 27708
[4]Department of Infectious Disease, Imperial College London, London W2 1PG
[5]Departments of Electrical Engineering and Computer Science, Biomedical Engineering, and Statistics, University of Michigan, Ann Arbor, MI 48109

Abstract—*Objective*: To develop a multi-channel device event segmentation and feature extraction algorithm that is robust to changes in data distribution. *Methods*: We introduce an adaptive transfer learning algorithm to classify and segment events from non-stationary multi-channel temporal data. Using a multivariate hidden Markov model (HMM) and Fisher's linear discriminant analysis (FLDA) the algorithm adaptively adjusts to shifts in distribution over time. The proposed algorithm is unsupervised and learns to label events without requiring *a priori* information about true event states. The procedure is illustrated on experimental data collected from a cohort in a human viral challenge (HVC) study, where certain subjects have disrupted wake and sleep patterns after exposure to an H1N1 influenza pathogen. *Results*: Simulations establish that the proposed adaptive algorithm significantly outperforms other event classification methods. When applied to early time points in the HVC data, the algorithm extracts sleep/wake features that are predictive of both infection and infection onset time. *Conclusion*: The proposed transfer learning event segmentation method is robust to temporal shifts in data distribution and can be used to produce highly discriminative event-labeled features for health monitoring. *Significance*: Our integrated multisensor signal processing and transfer learning method is applicable to many ambulatory monitoring applications. [1]

Index Terms—Covariate shift, domain adaptation, wearable sensors, human viral challenge study, digital health, early detection of viral infection.

I. INTRODUCTION

MANY physiological time series involve dynamic transition among event states. For example, the transitions in human circadian cycling alternate between sleep and wake states over a 24-hour period. The sleep state can itself be a dynamic process, switching over different stages of rapid eye movement (REM) and non-REM sleep states [9]. As another example, the transitions in menstrual cycles go through follicular and luteal phase [51], transitioning to an ovulation state. With the emergence of cheap wearable multi-channel physiological monitoring devices, there has been much interest in automating the detection of physiological event states. The potential impact of automated event detection is far reaching, potentially improving a person's health awareness and aiding the management of disease.

A principal impediment to automating event detection for wearable devices is the intrinsically high variability of the measured physiological signals over time. An especially challenging situation is when the wearer of the device undergoes a strong perturbation, such as exposure to a pathogen that results in infection. Most event detection algorithms that are trained over a healthy baseline time period will have difficulty adapting when the device's signals become strongly perturbed away from baseline. Modern machine learning approaches to training that incorporate transfer learning can mitigate these difficulties. This paper introduces an adaptive transfer learning method for automating event detection for wearable devices and demonstrates its ability to adapt to strong perturbations from a healthy baseline in the context of a human viral challenge study.

Another challenge to automating event detection in time series data is the lack of available ground truth event labels. The training of the automated algorithm must thus be unsupervised, based only on the observable physiological signals measured by the device. Effective unsupervised learning is only possible when features extracted from these signals have statistical distributions that strongly depend on the event states. Selection of strong discriminating features is therefore a crucial step in designing an unsupervised learning algorithm. The adaptive transfer learning algorithm introduced here generates a set of strongly discriminating event-labeled features that are based on statistical summaries of the device signals at multiple time scales.

A state-of-the-art unsupervised algorithm for detecting unobserved event state sequences from time series data is the hidden Markov model (HMM) [60]. The HMM is a generative model in which the hidden states are latent random variables that condition the joint probability distribution of the data. HMM methods have been widely applied to many health applications [8], [36], [58], [53], [5]. However, the HMM is not well adapted to data whose statistical distribution may shift over time. When there are external factors that cause perturbations to the distribution, the training data may be mismatched to the post-training data and the HMM will fail to perform as well as expected. In the machine learning literature this phenomenon is known as *data covariate shift* [43]. Many methods have been proposed to address the covariate shift problem, principally for supervised learning [57], [35], [38], [44], [21].

The adaptive unsupervised transfer learning method for event state classification introduced in this paper is designated by the acronym HMM-FLDA. It involves learning a multivariate HMM latent state prediction model that initializes a sequential version of Fisher's linear discriminant analysis (FLDA) to adapt the initialized HMM to perturbed post-baseline data. Our adaptive transfer learning method is introduced as the key component of a complete data processing pipeline that includes pre-processing for local feature extraction and abnormal signal detection, adaptive event classification using HMM-FLDA, and post-processing for constructing strong high dimensional event-labeled features for inferring perturbation-dependent events, e.g., health outcome after exposure to a viral pathogen.

We illustrate the proposed analysis pipeline for a human viral challenge (HVC) experiment in which physiological data from multiple subjects is collected over multiple days from wearable wristband devices. On the second day of the experiment the subjects were exposed to a H1N1 flu virus that caused some to become infected, as clinically determined by viral shedding which does not occur before the 4th day. For each subject, the HMM-FLDA is trained on the first two days and nights of wearable data to generate sleep/wake features. Using these features we are able to predict if the subject will get infected and the onset time, defined as the first day that shedding is detected.

The remainder of this paper is organized as follows: we introduce in detail the proposed transfer learning algorithm in Section II. Section III presents a numerical simulation study emulating the HVC experimental data, but with ground truth event labels. Section IV applies the proposed analysis pipeline to wearable data collected in the experimental HVC study. In Section V we end the paper with discussion and conclusions. Supporting information on the algorithm, the pipeline, and the HVC data is presented in the supplementary Appendices.

II. Adaptive unsupervised event monitoring

Figure 1a summarizes the proposed adaptive event labeling procedure and Algorithm 1 provides pseudo-code for its principal steps. Figure 1b shows the entire data processing and analysis pipeline, from sensor data capture to event-labeled feature extraction to predictive models using these features.

The pre-processed data comes in the form of a $p \times N$ matrix $\boldsymbol{X} = [\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_N}]$ where $\boldsymbol{x}_t \in \mathbb{R}^p$ is a p-dimensional real valued feature vector, available at a set of time instants $\{t_i\}_{i=1}^N$. We assume a statistical time series model for $\boldsymbol{x}_t$ with hidden states that determine the joint probability distribution $p_{\boldsymbol{X}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$. Specifically, associated with each time sample $\boldsymbol{x}_t$ is a latent (hidden) variable $y_t \in \{0, 1, \ldots, K\}$, which labels the event associated with $\boldsymbol{x}_t$. Later in the paper we will specialize to the binary case where $K = 2$ and the events correspond to a person's sleep or wake state at a particular time. We denote $\boldsymbol{y} = [y_{t_1}, \ldots, y_{t_N}]$ the time sequence of a subject's hidden event states. In the time series setting, $\boldsymbol{y}$ determines the conditional probability distribution $p_{\boldsymbol{X}|\boldsymbol{y}}$ and the marginal probability distribution $p_{\boldsymbol{X}}$. We construct an event classifier function based on an HMM with a parametric (Gaussian) conditional distribution of $\boldsymbol{x_t}$ given y_t that may gradually shift over time, e.g., due to a perturbation after exposure to a pathogen. The parameters of the classifier function are estimated from the data in a two phase process involving: (1) an initial training phase where the HMM model is fitted to a batch of (healthy) baseline data, followed by (2) an adaptation phase where a naive Bayes Gaussian model is adapted to future batches of possibly perturbed data using Fisher Linear Discriminant Analysis (FLDA) in a sequential transfer learning framework.

A. Baseline training: HMM initial segmentation

For $N_0 < N$ define the *baseline* (BL) segment of the data $\boldsymbol{X}$ as the first N_0 time samples $\boldsymbol{X}_{\mathrm{BL}} = [\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_{N_0}}]$ and the *post-baseline* (PBL) segments of the data as the remaining part $\boldsymbol{X}_{\mathrm{PBL}} = [\boldsymbol{x}_{t_{N_0+1}}, \ldots, \boldsymbol{x}_{t_N}]$. The BL segment is used as a training set for learning the parameters of a multivariate HMM [56]. The HMM is a Markov model: modeling the observations as an N-length segment of a Markov random process $\{(\boldsymbol{x}_t, y_t)\}_{t=-\infty}^{\infty}$, where transitions $y_{t-1} \to y_t$ between event states have probabilities specified by a $K \times K$ state transition matrix $\boldsymbol{A}$. The Markov property implies that the joint distribution $p_{\boldsymbol{X},\boldsymbol{y}}$ factorizes as: $p_{\boldsymbol{X},\boldsymbol{y}} = p_{\boldsymbol{x}_{t_1}, y_{t_1}} \prod_{i=2}^{N_0} p_{\boldsymbol{x}_{t_i}, y_{t_i} | \boldsymbol{x}_{t_{i-1}}, y_{t_{i-1}}}$, implying an analogous factorization of the marginal $p_{\boldsymbol{X}}$ and the conditionals $p_{\boldsymbol{X}|\boldsymbol{y}}$. In the special case of a Gaussian HMM model each factor $p_{\boldsymbol{x}_t | y_t}$ is a multivariate Gaussian conditional distribution, denoted as

$$\boldsymbol{x}_t \mid (y_t = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad k = 1, \ldots, K, \quad (1)$$

where $\boldsymbol{\mu}_k$ is a p-dimensional mean parameter and $\boldsymbol{\Sigma}_k$ is a $p \times p$ covariance matrix parameter that must be learned in addition to the state transition matrix $\boldsymbol{A}$.

There are several methods available for learning multivariate HMM's [22], [60] which could be adapted to our setting. Most methods are iterative and many use a variant of the expectation-maximization (EM) algorithm to find maximum likelihood estimates of the unknown parameters. Important practical considerations are whether the HMM algorithm provides a final estimate of labeling accuracy, important for self diagnostics, if it converges sufficiently rapidly to a global maximum, and if it has fast computation time per iteration, which tends to increase as K and p increase. In some
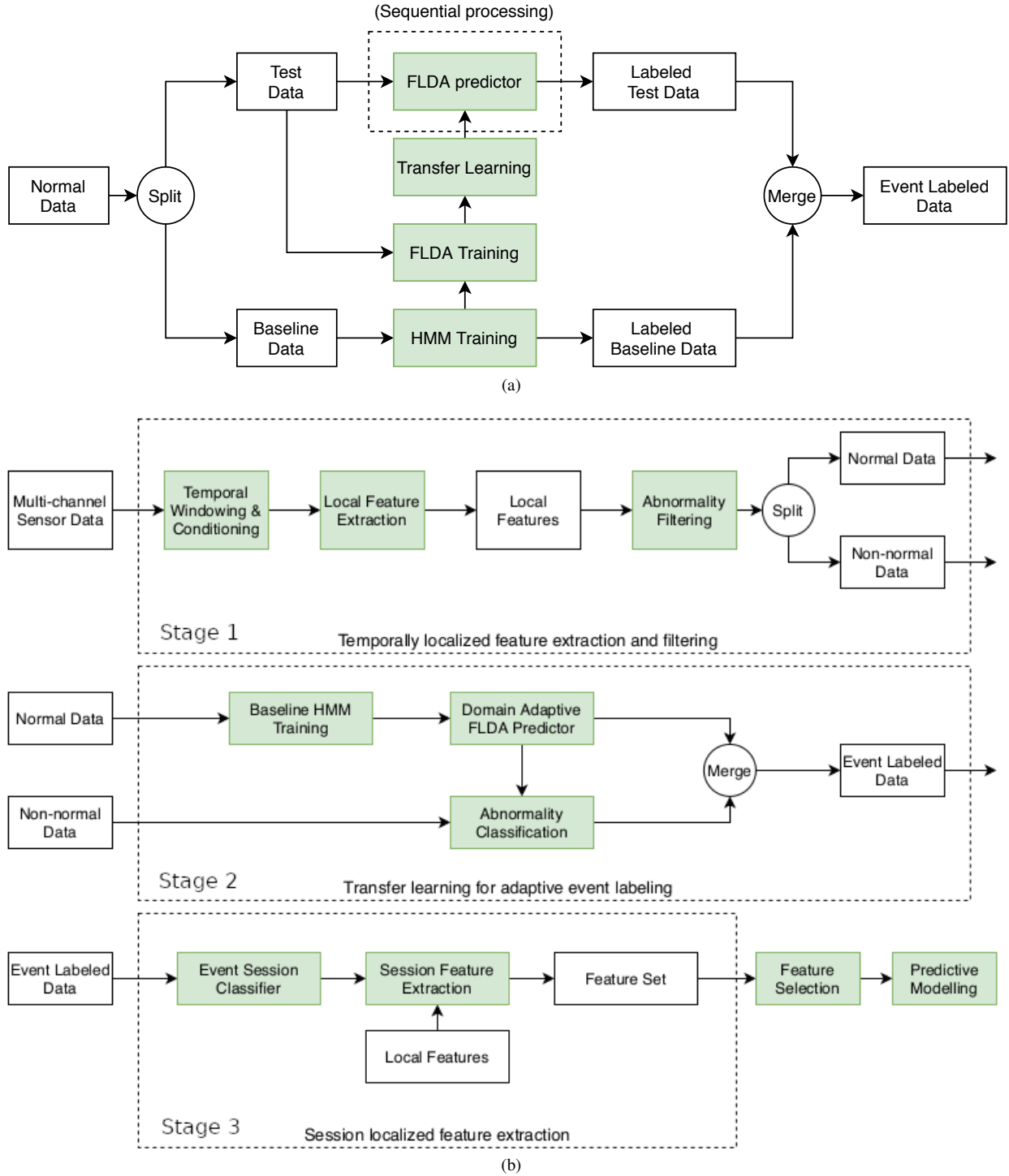
Fig. 1: **(a)** The proposed HMM-FLDA unsupervised transfer learning algorithm based on a multivariate hidden Markov model (HMM) and Fisher's linear discriminant analysis (FLDA) to segment event states; **(b)** The three stages of proposed data processing pipeline, discussed in Section IV-A. The HMM-FLDA procedure depicted in **(a)** appears in the second stage of the pipeline.

applications, it may be desirable to estimate the number of event states K, for which automated model selection versions of HMM are available. For a comprehensive review of HMM implementations the reader can refer to [60, Ch. 8]. In our pipeline we use a reasonably fast multivariate HMM algo-

rithm implemented by EM iterations that approximates the maximum likelihood estimator of the HMM parameters. In the Expectation (E) step, the marginal likelihood is computed by a variant of the forward-backward algorithm [39], which calculates the gradient of the log-likelihood (score) function

---

Algorithm 1 HMM-FLDA gradual self-training event state classifier

---

1: Initialize with: $\{\hat{y}_t, \boldsymbol{x}_t\}_{t=1}^{t_{N_0}}$ ▷ Use HMM to classify baseline data
2: **while** $t_{N_0} < t < t_N$ **do**   ▷ Continue over all batches of samples
3:    **for** l in 1 to L **do**   ▷ Iterate over candidate training sizes d_l
4:      **procedure** FISHER'S LDA(train = $(t - d_l, t]$, test = $(t, t + \Delta_t]$)
5:        $\boldsymbol{w} = \boldsymbol{S}_W^{-1}(\bar{\boldsymbol{x}}_{1\cdot} - \bar{\boldsymbol{x}}_{0\cdot})$ ▷ Update FLDA weights on train data
6:        $z_t = \boldsymbol{w}^T \boldsymbol{x}_t$
7:        $\hat{y}_t(d_l) = \mathbb{1}\{(z_t - \bar{z}_{0\cdot})^2/\hat{\sigma}_0^2 - (z_t - \bar{z}_{1\cdot})^2/\hat{\sigma}_1^2 > \log(\gamma\,\hat{\sigma}_1^2/\hat{\sigma}_0^2)\}$ ▷ Classify test data with updated weights
8:        $\mathrm{SI}(d_l) = \mathrm{SI}(\text{train}, \text{test}; d_l)$ ▷ Compute separability index (5)
9:    $d_{(1)} = \textbf{which.max}(\mathrm{SI}(d_l))$ ▷ Select optimal training window length
10:    $\hat{y}_t = \hat{y}_t(d_{(1)})$ ▷ Classified labels of current batch
11:    $t = t + \Delta_t$ ▷ Move on to the next batch

---

in a single pass.

B. Post-baseline adaptation: sequential transfer learning

If the feature distribution varies over time, the static baseline-trained HMM will have difficulty classifying and segmenting events in the post-baseline data $\boldsymbol{X}_{\mathrm{PBL}}$. To address this difficulty we introduce a transfer learning strategy that is initialized with the HMM on the initial-training data and sequentially updates the event classifier over successive batches of test data, continually adapting to changes in distribution. In transfer learning the batches are called target domains and the objective is to design a classifier that continually adapts to them, a property called domain adaptation [33], [43]. To achieve this objective, we use an unsupervised *gradual self-training* procedure. In supervised transfer learning, "self-training" means that domain adaptation is done with an unlabeled test set and "gradual" means that the adaptation is done sequentially over time, updating over successive batch pairs. As compared to direct self-training procedures, which try to adapt to all test batches at once, gradual self-training procedures are better suited to online applications and are more robust to smoothly varying shifts in distribution. Recent theory establishes that gradual procedures are provably better when the class distributions on successive pairs of batches are close in Wasserstein distance [37]. This closeness condition is satisfied when the distributions are quasi-stationary, an assumption common in time series analysis [1] and online supervised learning [50].

Our unsupervised gradual self-training procedure consists of three components: 1) an unsupervised version of Fisher's LDA that uses class labels predicted by the HMM to project the data $\boldsymbol{X}_{\mathrm{BL}}$ to a one dimensional space that maximally separates these classes. 2) application of this projection to the next batch of samples in $\boldsymbol{X}_{\mathrm{PBL}}$ followed by event classification using a naive Bayes classifier; 3) using the class labels to update the projection by reapplying Fisher's LDA to the batch filtered by a tapered sliding-window. This process is sequentially repeated for each successive batch, resulting in a continuous adaptation of the event classifier.

Fisher's LDA is most commonly applied in the context of supervised learning for dimension reduction and classification [18], [20] when a labeled sample is available for training. Here we adapt LDA to the unsupervised context of latent event classification. After applying HMM to the baseline sample $\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_{N_0}}$ we obtain estimated class labels $\hat{y}_{t_1}, \ldots, \hat{y}_{t_{N_0}}$. We then train Fisher's LDA classifier on the putatively labeled sample $\{(\boldsymbol{x}_{t_1}, \hat{y}_{t_1}), \ldots, (\boldsymbol{x}_{t_{N_0}}, \hat{y}_{t_{N_0}})\}$. Fisher's LDA uses dimensionality reduction to learn a classifier. By applying a weight $\boldsymbol{w} \in \mathbb{R}^p$ to $\boldsymbol{x}_t$ FLDA generates a projection score $z_t = \boldsymbol{w}^T \boldsymbol{x}_t$. The weight vector is optimized so that these scores attain the largest possible separation of the classes. Fisher's LDA accomplishes classification of a novel sample $\boldsymbol{x}_{t'}$ by using the trained weights to compute the projected score $\boldsymbol{w}^T \boldsymbol{x}_{t'}$ that is used in an optimal naive Bayes LDA classifier to predict its unknown label $y_{t'}$.

Specifically, for the case of $K = 2$ classes $y_t \in \{0, 1\}$, in the dimensionality reduction stage of FLDA the class separability of the projection scores $z_t = \boldsymbol{w}^T \boldsymbol{x}_t$ is measured by the following ratio of *between-class* variation and *within-class* variation:

$$J(\boldsymbol{w}) = \frac{(\bar{z}_{1\cdot} - \bar{z})^2 + (\bar{z}_{0\cdot} - \bar{z})^2}{\sum_{t \in G_1}(z_t - \bar{z}_{1\cdot})^2 + \sum_{t \in G_0}(z_t - \bar{z}_{0\cdot})^2}, \quad (2)$$

where $G_k = \{t \leq N_0 : y_t = k\}$ denotes the data indices of $\boldsymbol{x}_t$ coming from event class k, $\bar{z}_{k\cdot} = |G_k|^{-1}\sum_{t \in G_k} z_t$ for $k \in \{0, 1\}$ are the class-specific sample mean projected score and $\bar{z}$ is the overall sample mean projected score. The optimal $\boldsymbol{w}$ that maximizes (2) is [20]

$$\boldsymbol{w} = \boldsymbol{S}_W^{-1}(\bar{\boldsymbol{x}}_{1\cdot} - \bar{\boldsymbol{x}}_{0\cdot}), \quad (3)$$

where $\bar{\boldsymbol{x}}_{k\cdot} = |G_k|^{-1}\sum_{t \in G_k} \boldsymbol{x}_t$ is the class-specific sample mean and $\boldsymbol{S}_W = \sum_{k=0,1}\sum_{t \in G_k}(\boldsymbol{x}_t - \bar{\boldsymbol{x}}_{k\cdot})(\boldsymbol{x}_t - \bar{\boldsymbol{x}}_{k\cdot})^T$ is the *within-class* scatter matrix of the $\boldsymbol{x}_t$'s.

Under the multivariate normal model (1) for $\boldsymbol{x}_t$, the z_t's are conditionally independent Gaussian distributed random variables with conditional mean μ_k and conditional variance σ_k^2 given $y_t = k$. Assuming independent conjugate priors on μ_k and σ_k^2, conditioned on the training data $\boldsymbol{X}_{\mathrm{BL}}$, the maximum a posteriori (MAP) estimator of the class label $y_{t'}$ of a novel test sample $\boldsymbol{x}_{t'}$ based on its projected score $z_{t'}$ is given by the naive Bayesian classifier [20]:

$$y_{t'} = \mathbb{1}\left\{\frac{(z_{t'} - \bar{z}_{0\cdot})^2}{\hat{\sigma}_0^2} - \frac{(z_{t'} - \bar{z}_{1\cdot})^2}{\hat{\sigma}_1^2} > \log\left(\gamma\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)\right\}, \quad (4)$$

where $\mathbb{1}(\mathcal{A})$ is the binary indicator function equaling 1 when the logical preposition $\mathcal{A}$ is true. Here, as above, $\bar{z}_k$ is the

class-specific sample mean and $\hat{\sigma}_k^2$ the within class sample variance of $\{z_t : t > N_0\}$ $\hat{\sigma}_k^2 = (|G_k| - 1)^{-1} \sum_{t \in G_k}(z_t - \bar{z}_{k\cdot})^2$. The threshold parameter γ in (4) is the prior odds ratio $P(y_{t'} = 0)/(1 - P(y_{t'} = 0))$ and can be adjusted to account for class imbalance and to trade-off the two types of classification errors. The naive Bayes classifier classifies the label of the test sample based on its relative distance to each of the class centroids $\bar{z}_k$ weighted by the class dispersions $\hat{\sigma}_k^2$.

We next describe the proposed unsupervised transfer learning procedure using gradual self-training to sequentially update the HMM-FLDA algorithm as, outlined in Algorithm 1 and depicted in Fig. 1a. The post-baseline data is successively divided into disjoint test batches of length Δ_t secs., called the test window length, a tuning parameter whose selection is application dependent. The current batch is used as a test sample to update the self-trained HMM-FLDA classifier determined from the previous batch. Let $t_n^{tst} = t_{n-1}^{tst} + \Delta t$ be the start time of the n-th test window and denote the n-th test batch as $\mathcal{T}_{tst} = \{\boldsymbol{x}_{t_i}\}_{t_i \in [t_n^{tst}, t_n^{tst} + \Delta_t)}$. The update proceeds as follows. For a specified size d, define the n-th training set $\mathcal{T}_{trn} = \{\boldsymbol{x}_{t_i}\}_{t_i \in [t_n^{tst} - d, t_n^{tst})}$. Note that the feature instances in $\mathcal{T}_{trn}$ have already been assigned class labels $\{\hat{y}_{t_i}\}_{t_i \in [t_n^{tst} - d, t_n^{tst})}$ in the previous update of HMM-FLDA. Hence, the first step in updating the classifier is to compute the updated FLDA weight vector $\boldsymbol{w}$, as defined in (3), using $\mathcal{T}_{trn}$ and its previously assigned class labels. The second step is to use this updated weight vector to assign predicted labels $\{\hat{y}_{t_i}\}_{t_i \in [t_n^{tst}, t_n^{tst} + \Delta t)}$ to instances in $\mathcal{T}_{tst}$.

The size d of $\mathcal{T}_{trn}$ is adapted from batch to batch by optimizing a separability index (SI) defined over $\mathcal{T}_{trn} \cup \mathcal{T}_{tst}$. Among the many possible SI measures that could be used, we adopt Thornton's SI [52], a competitive measure for assessing class separation originally introduced for supervised classification problems [4], [10], [23]. Also called the geometric separability index (GSI), Thornton's measure has seen wide application in health, robotics, geology and other fields [42], [14], [49], [54], [47]. We use the following simple unsupervised modification of the GSI [52], computed on the merged training and test samples for the batch:

$$\text{SI}(d) = \frac{\sum_{t \in \mathcal{T}_{trn}(d) \cup \mathcal{T}_{tst}} \{\hat{y}_t + \tilde{y}_t + 1\} \textbf{ MOD } 2}{|\mathcal{T}_{trn}(d)| + |\mathcal{T}_{tst}|}, \quad (5)$$

where $\tilde{y}_t$ is the predicted class of the nearest neighbor of $\boldsymbol{x}_t$ in the set of merged training and test samples. As compared to Thornton's original SI, defined for supervised classification problems where the true labels y_t are known, the unsupervised modification (5) uses the predicted labels $\hat{y}_t$ in place of y_t. Here the nearest neighbors are determined by the "projection distance" defined, for samples at time t and t', as $d(t, t') = |\hat{\boldsymbol{w}}^T(\boldsymbol{x}_t - \boldsymbol{x}_{t'})|$. Rank ordering the SI indices $\text{SI}(d_{(1)}) > \ldots > \text{SI}(d_{(L)})$ yields $d_{(1)}$ as the optimal training window length for the batch.

III. NUMERICAL SIMULATION STUDY

We performed a simulation of HMM-FLDA that emulates the experimental study presented in Section IV, but with ground truth event states. Two dimensional data $\boldsymbol{x}_t =$

$[x_t^{(2)}, x_t^{(2)}]$ is simulated from conditional distributions given the latent event state y_t, which randomly switches between wake ($y_t = 0$) and sleep ($y_t = 1$) with a mean cadence of approximately 24 hours. To emulate the perturbation effect of viral infection, each of these conditional distributions are fixed during the baseline training period but may undergo slowly varying time shifts in the post-baseline period. Two different post-baseline scenarios are simulated, called the stable case (no shift) and the unstable case (slow shift).

For the stable case, we generate the sequence of states as a realization of a Markov process with state transition probabilities that are empirically matched to those extracted from the HVC data described in the next section. We fix the number of state transitions T and the initial state is set to $y_0 = 0$. The duration $\Delta \tau_1$, i.e., the time to the first transition, is drawn from $TN(\mu_0, \sigma_0)$, a truncated normal (TN) distribution with location and scale parameters μ_0 and σ_0. The duration $\Delta \tau_2$ of the second event is independently drawn from a TN distribution with mean $\mu_1 = 24 - \mu_0$ and variance $\sigma_1^2 = \sigma_0^2$. This process is repeated until the T-th transition variable $\Delta \tau_T$ has been drawn. The discretized transition times $\tau_m = \sum_{i=1}^{m} \Delta \tau_i$, $m = 1, \ldots, T$, specify the state sequence $\{y_t, t = 1, \ldots, N\}$. A consecutively occurring pair of (wake, sleep) periods is called a session. Over a given session m, the data channels $x_t^{(i)}$, $i = 1, 2$, were drawn independently from a truncated normal distribution $p(x_t^{(1)}|y_t = k)$ of the form $TN(u_k^{(1)}, \sigma_k^{(1)})$ and a log-normal distribution $p(x_t^{(2)}|y_t = k)$ of the form $LN(u_k^{(2)}, \sigma_k^{(2)})$, whose translation and scale parameters u and σ depend on the state $k = 0, 1$. These two distributions emulate the median heart rate feature and the accelerometer standard deviation feature used in Section IV-B2 for the full time course HVC data.

For the unstable case, the parameters of the distributions of y_t and $\boldsymbol{x}_t$ were changed from session to session to reflect perturbations during the post-baseline period. Specifically, the truncated normal parameters (μ_{mk}, σ_{mk}) of the event durations $\Delta \tau_m$ were made to be session dependent by matching these parameters to the empirical distributions of the HMM-FLDA estimated duration for the m-th session, where the empirical distribution was constructed over the sub-population of symptomatic subjects in the HVC study. Furthermore, we introduced a time-varying post-baseline conditional mean $u_{mk}^{(i)} = E[x_t^{(i)}|y_t = k]$ of the i-th channel over the m-th session for the k-th event state. In particular, the post-baseline conditional mean was modeled as a quadratically varying function of m: $u_{mk}^{(i)} = u_{1k}^{(i)} - b_k^{(i)}(m - m_o)^2 / (1 - m_o)^2 + b_k^{(i)}$, for $m = 1, \ldots M$, $i = 1, 2$, $k \in \{0, 1\}$. Here m_o is the index of the session where the mean $u_{mk}^{(i)}$ reaches a positive or negative apex, and $b_k^{(i)} = u_{m_o j}^{(i)} - u_{1k}^{(i)}$ is the difference between the apex value $u_{m_o k}^{(i)}$ and the initial value $u_{mk}^{(i)}$. The pair $(m_o, b_k^{(i)})$ controls how the mean values of $(x_t^{(1)}, x_t^{(2)})$ change over sessions. In the simulation, m_o was randomly drawn from $\{5, 6, 7\}$ with equal probability, and we considered 2 sets of values for the $b_k^{(i)}$'s. To achieve a close facsimile to the experimental data analyzed in Section IV, we simulated $M = 11$ sessions, and the time between samples was set to $\delta_t = 1/6$ hr, i.e., the 10 minute aggregated sample period used

in that section. Furthermore, all of the translation and scale parameters used in the simulation were matched to summary statistics obtained from the HMM-FLDA analysis described in Section IV-B2. Several realizations of the simulated data are shown in Fig. 5 in the Appendix.

Table I compares the empirical performance (a total of 1000 independent trials on 100 different realizations of the model) achieved by the proposed HMM-FLDA procedure to other procedures (HMM and dHMM) for the stable and unstable cases described above. Each of the stable and unstable cases is characterized by the value of the four coefficients $(b_0^{(1)}, b_1^{(1)}, b_0^{(2)}, b_1^{(2)})$ that define the trend in the post-baseline mean $u_{mk}^{(i)}$. For a stable subject there is no mean trend for either event class and all of these coefficients are zero. Two unstable cases are considered: a case called *unstable++*, where the trends for both event classes are concave with a randomly located peak, and a case called *unstable+-*, where for each event class the trends go in opposite directions, i.e., one trend is concave while the other is convex. The table shows the value of 5 performance criteria (averaged over 1000 independent trials) when the method is trained on a single realization of the two signals and tested on an independent out-of-sample realization drawn from the same distribution. The performance criteria are:

- `Accuracy`: the number of samples with correctly classified event states divided by the total number of samples,
- `F1`: the harmonic mean of the precision and recall classification criteria,
- `Cosine dis.`: the cosine distance between the vector of classified event states and the ground truth states,
- `Onset diff.` the absolute difference between estimated onset time of a sleep session and the nearest onset time of a true sleep period, averaged over sessions,
- `Duration diff.` the absolute difference between estimated duration of a sleep session and the duration of a true sleep period that overlaps with the predicted one, averaged over sessions. (If the predicted period doesn't overlap with any true period, the difference is taken as the length of the predicted period)

Table I compares the mean performance of the proposed method to a standard off-the-shelf two-state HMM state estimator and a variant of HMM, called dHMM, that uses locally estimated scatterplot smoothing (LOESS) [13] to detrend each of the two data channels as a preprocessing step before training the HMM. The HMM and LOESS were implemented with R packages `depmixS4` [56] and `fANCOVA`. The results shown in Table I establish the benefit of our proposed HMM-FLDA adaptive procedure. In terms of event classification performance (Accuracy, F1, and Cosine distance), except for cosine distance in the unstable++ case, where the performances are statistically equivalent, the proposed method significantly outperforms the competing HMM and dHMM procedures for both unstable++ and unstable+- models (p-value ≪0.01 according to one-sided paired t-test). It is also worth noting that, in the unstable++ case, LOESS detrending of dHMM improves on HMM across all criteria except for onset diff., which is the r.m.s. estimation error for sleep onset time. The Supplementary

Appendix C provides additional details on this simulation.

IV. APPLICATION TO HVC EXPERIMENTAL DATA

We apply the proposed adaptive transfer learning method to sleep/wake event classification for an experimental dataset undergoing a perturbation after baseline. This dataset of wearable data was collected as part of a human viral challenge (HVC) study where data from a cohort of participants was collected before and after exposure to a viral pathogen. More details on the HVC study are provided in Appendix A. We will show that the proposed FLDA-HMM algorithm, trained individually on each participant without clinical outcome data, is capable of segmenting sleep events and extracting features that can be used to accurately predict clinical outcomes, specifically whether a subject is infected or not, and the onset time of infection.

Sleep has repeatedly been found to be associated with immune, cardiovascular, and neuro-cognitive function [41], [30], among other functions. Many studies have revealed that changes in sleep pattern can be an important modulator of human response to diseases. For example, in a human viral challenge (HVC) study [16], researchers found that nasal inoculation with rhinovirus type 23 significantly reduced total sleep time among symptomatic individuals during the initial active phase of the illness. In another study [15], shorter sleep duration in the weeks preceding an exposure to a rhinovirus was found associated with lower resistance to illness. Since physiological signals such as instantaneous heart rate, physical activity, and skin temperature differ substantially during wake and sleep periods, automated sleep/wake labeling is possible. Therefore, the development of effective sleep monitoring methods has been of increasing interest.

While polysomnography (PSG) is the gold standard for sleep monitoring in sleep-related studies [32], [17], it is often cumbersome outside of a lab setting as it uses electroencephalogram (EEG), electrocardiogram (ECG), and electromyogram (EMG), and requires a registered technician to perform sleep scoring [46]. Recently much effort has been made to remedy the inconvenience of manual scoring and, alternatively, to provide automated analysis of the PSG signals [19], [40], [31]. Thus there has been growing interest in low-cost alternatives to PSG using small package wearable devices. Signals captured by wearable sensors have included movement induced accelerometer signals, circulating blood volume pulse (BVP), heart rate (HR), electrodermal activity (EDA), temperature, respiration effort (RSP), ambient light and sound. These devices can be wrist-worn, ankle-worn, arm-worn, lapel-worn, chest-worn or embedded in mobile phones [34],[48], [17]. The Empatica E4 device (Empatica Inc. USA), used in the HVC study discussed below, captures 4 such signals.

A. HVC data and processing pipeline

Data was collected from E4 wristband devices worn by 25 participants enrolled in a longitudinal human viral challenge (HVC) study over the time period: 0 hours to 270 hours (11 days). The study participants were exposed to a pathogen (Inf

TABLE I: Out of sample performance of the proposed adaptive transfer learning algorithm as compared with standard HMM operating on the original data (HMM) and operating on LOESS detrended data (dHMM).

Setting	Methods	Accuracy	F1	Cosine dis.	Onset diff.	Duration diff.
stable	HMM	**0.9981**	**0.9987**	**0.9968**	**0.0165**	**0.0307**
	dHMM	0.9981	0.9986	0.9967	0.0253	0.0373
	Proposed	0.9980	0.9986	0.9964	0.0168	0.0313
unstable++	HMM	0.8959	0.9209	0.8559	1.6704	2.9169
	dHMM	0.9127	0.9231	**0.8957**	2.4709	2.2653
	Proposed	**0.9356**	**0.9544**	0.8904	**1.0679**	**1.5756**
unstable+-	HMM	0.9369	0.9510	0.9187	**0.2763**	1.9750
	dHMM	0.9271	0.9370	0.9114	1.0170	2.0131
	Proposed	**0.9758**	**0.9833**	**0.9564**	0.3712	**0.3817**

TABLE II: Parameter settings for the HVC data processing pipeline

Symbol	Value	Description	Loc. in pipeline
δ_t	10 minutes	epoch length	local feature extraction (pre-processing)
t_1	Hour 0	starting time of BL segment	Baseline HMM training
t_{N_0}	Hour 36	end time of BL segment	Baseline HMM training
Δ_t	3 hours	test window length	Domain adaptive FLDA training
d_l	$\{12, 13, \ldots, 59, 60\}$ hours	training window lengths	Domain adaptive FLDA training
γ	1	prior odds ratio	Domain adaptive FLDA training

H1N2) on the second day of the study at 36 hours (noon on day 2). Over subsequent days some participants developed symptomatic infection, confirmed by viral shedding from PCR assay on nasal lavage once per day. Such tests are often used to detect or confirm viral infection and to identify asymptomatic spreaders [45], [29], [59], [24]. In the HVC study, shedding in infected subjects was first detected either on day 4 at 81 hours, designating early onset shedders, or on day 5 at 106 hours, designating late onset shedders.

Four channels of physiological signals, including 3-axis accelerometry (ACC sampled at 32 sa/sec), heart rate (HR at 1 sa/sec), skin temperature (TEMP at 4 sa/sec) and electrodermal activity (EDA at 4 sa/sec), were measured by the E4 devices. See Fig. 4a in Appendix B for representative signals from an infected and a non-infected subject.

We implemented the three-stage data processing pipeline in Fig. 1b for sleep segmentation and high dimensional feature extraction from the HVC data. The first stage (top branch in Fig. 1b) consists of a pre-processing module, performing data conditioning, fine grain temporally-localized feature extraction and abnormality filtering. The second stage (middle branch) implements the HMM-FLDA module, performing adaptive sleep segmentation on non-abnormal data. The third stage (bottom branch) is a post-processing module, aggregating the data into sleep and wake sessions and performing coarse grain session-localized feature extraction.

Table II summarizes the parameter settings we used to implement the pipeline for the HVC data. Figure 2 depicts how the pipeline constructs sleep/wake sessions and session-localized features from the fine grained features.

1) First stage: local features and abnormality filtering: Like other wearable sensors, the Empatica E4 captures data at high frequencies and, because of the fact that they are worn by subjects in non-laboratory situations, the raw data collected are often voluminous and noisy. To mitigate the impact of occasional poor readings and reduce computational burden, the pre-processing stage performs data conditioning, local feature extraction and abnormality filtering. First the missingness of a subject's available data is evaluated, resulting in rejection of any subject with more than 40% missing time points. For each of the remaining subjects, their data is segmented into non-overlapping time intervals, which we call *epochs*, of equal length δ_t secs. The epoch length was set to $\delta_t = 10$ mins to achieve a tradeoff between oversmoothing the sleep/wake transitions (excessive length) and noise sensitivity (insufficient length). See Appendix D for additional discussion. Shorter epoch lengths may be more appropriate for classifying other types of event states, e.g., for capturing quality of sleep or detecting transitions between different sleep stages. For each of the 4 E4 signals, the module in Fig. 1b labeled "Local feature extraction" computes statistical features over each epoch corresponding to the signal mean (MEAN), signal median (MED) and signal standard deviation (SD). If a particular epoch has fewer than 90% available samples, e.g., due to the

TABLE III: Features (**196** in total) extracted from sleep and wake sessions for each day from subjects in the HVC study.

Name	Number	Description
Duration	2	total sleep, night sleep
Onset/offset	2	night sleep only
HR summary	9×2	3 (mean, median, s.d.) $\times$ 3 (mean, median, sd within session) $\times$ 2 (sleep, wake)
HR linear coef.	6×2	3 (mean, median, sd) $\times$ 2 (coef.0, coef.1) $\times$ 2 (sleep, wake)
HR quadratic coef.	9×2	3 (mean, median, sd) $\times$ 3 (coef.0, coef.1, coef.2) $\times$ 2 (sleep, wake)
TEMP	24×2	same as HR
ACC	24×2	same as HR
EDA	24×2	same as HR
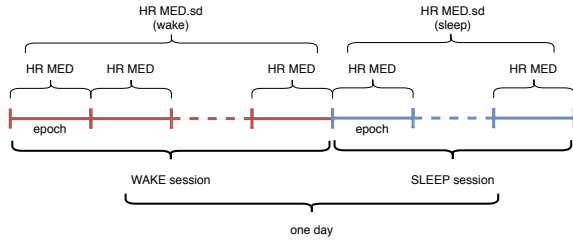
Fig. 2: Labeling of events and event sessions from wearable data, illustrated for one of the features (HR MED = heart rate median) and one of the meta-features (HR MED.sd = standard deviation of HR MED) for Empatica E4 device in the Human Viral Challenge (HVC) study. In the first stage of the pipeline in Fig. 1b, epochs of 10 min duration (δ_t in Table II) are used as windows over which temporally localized statistical features (mean, median, standard deviation) of the E4 device signals are extracted. These features are used in the second stage of the pipeline that extracts event labels, wake and sleep in the case of the HVC, and organizes them into contiguous labeled time segments, called sessions. In the third stage of the pipeline event-labeled meta-features (HR MED.sd (wake) and HR MED.sd (sleep)) are extracted as statistical summaries over the wake and sleep sessions.

E4 not being worn, the epoch is discarded.

These local features are then processed by the "Abnormality filtering" module to identify time points at which outliers occur, labeled as non-normal data in Fig. 1b. Similarly to methods used in network intrusion detection [6], the module uses marginal k-means clustering and quantile filtering to label time points as outliers (See Appendix D-A1). Subjects having less than 60% normal data cannot be reliably segmented and are omitted from subsequent analysis.

The last pre-processing step consists of selecting a subset of the 12 local features to train the HMM-FLDA sleep segmentation procedure. For this we select putative wake and sleep time intervals in the early evening around 21:00 and in the twilight hours around 4:00. These time periods were chosen since it is expected that most people would be awake at 9pm and sleeping at 4am. The sleep/wake discrimination capability of each feature over these two intervals was measured by a

sleep/wake separability index (SWSI) computed over all non-abnormal subjects and over all available days. The SWSI is defined similar to the geometric separability index (5):

$$\text{SWSI}(k) = \frac{\sum_{t\in\mathcal{T}_{wake}\cup\mathcal{T}_{sleep}} \{\hat{y}_t + \tilde{y}_t(k) + 1\}\ \textbf{MOD}\ 2}{|\mathcal{T}_{wake}| + |\mathcal{T}_{sleep}|}, \quad (6)$$

where $k = 1,\ldots,p$ indexes the features, $\mathcal{T}_{wake}$ and $\mathcal{T}_{sleep}$ are respectively the time intervals around 21:00 and 4:00 over all available days of data, $\hat{y}_t$ is the putative label assigned to the k-th feature $x_t(k)$, defined as $\hat{y}_t = 1$ if $t \in \mathcal{T}_{sleep}$ and $\hat{y}_t = 0$ otherwise, and $\tilde{y}_t(k)$ is the label of the nearest neighbor of $x_t(k)$ among $\{x_\tau(k) : \tau \in \mathcal{T}_{wake} \cup \mathcal{T}_{sleep}\}$. The criterion for including the k-th feature among those that are used to train the HMM-FLDA is that SWSI(k) be greater than 0.7 for at least 75% of the subjects.

2) Second stage: adaptive sleep detection: Adaptive sleep segmentation was implemented using the HMM-FLDA procedure described in Section II. The initial training window (baseline) for training the HMM-FLDA (Algorithm 1) was defined as the period from 0 to 36 hours, the time of viral inoculation. The test window length was chosen to be $\Delta_t = 3$ hours, which is based on our expectations about the time scale of immune response induced changes in the E4 signal distributions subsequent to pathogen exposure. The most rapid time scales of immune response are on the order of a few hours, e.g., inflammatory monocyte recruitment [27]. Such a choice of Δt will minimize the likelihood that the HMM-FLDA algorithm loses track of covariate shifts from batch to batch. The range of training window lengths $d_1,\ldots,d_L$ used by the optimization loop in Algorithm 1 was restricted to $d_1 = 12$ to $d_L = 60$ hours to guarantee a good balance of wake and sleep samples over the window.

After the HMM-FLDA procedure labels the normal data, an "Abnormality classifier" module identifies any abnormal samples (identified in Stage 1) that is physiologically meaningful and reinserts it into the data stream, labeling it with the label of the sample immediately preceding it (See Fig. 12 in Appendix D-A2). Finally, a modified median filter with a 90-min smoothing window is applied to remove short bursty sleep periods of less than 60 mins. Such short sleep periods are likely to lack deep sleep stages 3 and 4, which typically

start 30 mins. after sleep onset and can last 20 to 40 mins. [9].

3) Third stage: session-level feature set: In the third stage of the pipeline, the sleep/wake sessions produced by the second stage are aligned to particular calendar days. A full day session is composed of wake sessions and sleep sessions. Any sleep session overlapping with a given day whose onset occurs before 5am (5:00) is associated with the previous day session. The binary sleep/wake label is then appended to the local features generated in the the first stage of the pipeline, doubling the total number of local features. Then a set of session-level features are extracted by computing several statistical summaries over daily wake and sleep sessions. This results in 196 session-level features that include the mean, median, and standard deviation (sd) of the event labeled local features over a sleep or wake session. All features related to standard deviation are log-transformed. The session-level features also include the coefficients of linear and quadratic fits to the time course of these features over a sleep or wake session. In addition session timing features, such as, sleep duration, sleep onset, and sleep offset are included in the session-level feature set. See Table III.

B. Online sleep segmentation

We emulate an online implementation of the pipeline in Fig. 1b by reapplying it successively to all available data at the end of each day. Daily updating of the segmentation corresponds to the real-world scenario where data is uploaded from the E4 device once per day, corresponding to the data acquisition rate in the HVC study. Continuous (real-time) updating will only become practical when devices have sufficient power for continuous data transfer or onboard processing. Below we illustrate the online implementation for a case of pre-infection segmentation before shedding occurs, and a full time course segmentation at the end of the study. For additional details see Appendix E.

1) Illustration: pre-infection segmentation: Here only the first 60 hours (up to day 3 at 12pm) of the data are available to the pipeline. The first 60 hours include the inoculation time (36 hours) and the first two nights of sleep. Note that no infection (shedding) is detected before day 4 at 81 hours. The preprocessing stage of the pipeline removed 5 subjects with excessive missing or abnormal data: 2 of these subjects had more than 40% of their time points missing and 3 subjects had more than 40% abnormal time points. For the remaining 20 subjects local feature selection was accomplished by applying SWSI (6) to all 12 variables to contrast early evening periods (19:00-22:00) to twilight periods (2:00-5:00). Only three features were found to have SWSI above 0.7 for at least 75% of the subjects: the mean and median heart rate (HR MEAN and HR MEDIAN) and the standard deviation of the magnitude accelerometer (ACC SD). As the mean and median heart rate are highly correlated and the SWSI of HR MEAN has lower 25% quantile than does the HR MEDIAN, only the HR MEDIAN and ACC SD were selected. Three hour long sleep and wake periods were necessary in order to obtain a sufficient number of representative time samples since only two days and nights are available.

2) Illustration: full time course segmentation: The full time course (0-270 hours) of a subject's data was made available to the pipeline. No subjects had more than 40% missing time points over this period. Stage 1 abnormality filtering resulted in elimination of 5 subjects as excessively abnormal, who were removed. To compute SWSI for variable selection on each of the remaining 20 subjects, we used shorter putative wake periods (21:00-22:00) and sleep periods (4:00-5:00) since there are many more available days and nights than in the 60 hour pre-infection time period. Applying the same SWSI variable selection criterion three features were selected to train the HMM-FLDA: HR MED, HR SD, and ACC SD.

3) Clinical outcome prediction: We used the feature set generated by the pre-infection segmentation to perform early detection of infection. Two clinical outcomes were considered: 1) a binary infection state (whether or not the subject will shed virus anytime after 60 hours) designating the subject as infected or non-infected; and 2) the ternary infection state corresponding to onset time of viral shedding. The three onset times are defined as: early onset shedding (first detected by PCR at 81 hours), late onset shedding (first detected by PCR at 106 hours), and no onset shedding (subject never sheds virus). Among the 20 subjects whose 0-60 hour data passed the abnormality filtering test, 9 of them were infected and 11 were non-infected. Among those infected, 3 were early shedders and 6 were late shedders.

Two predictor models were used for each of these outcomes, which were trained only on the sleep/wake features collected on inoculation day (wake/sleep sessions for day 2). For prediction of whether a subject will shed or not we applied logistic regression (LR) [20] and for prediction of onset time we used continuation ratio (CR) regression [2], which can be interpreted as a discrete version of the Cox regression model [26].

For both the logistic and the continuation-ratio models the classification capability of individual features is evaluated by leave-one-out cross-validation. The top 4 discriminative features are reported in Table IV, together with their corresponding regression coefficients and a detection performance criterion: the area under the ROC curve (AUC). AUC is a measure of the accuracy of a binary classifier and a value of AUC near 1 means very high accuracy. For CR regression, since there exist three classes, three different AUCs are calculated using the "one-versus-the-rest" strategy [7]. The CR performance is rank ordered in decreasing order of the minimum of the AUC's for Early, Late and No onset. See Appendix E for boxplots of these top 4 features.

The top feature for both LR and CR models is a sleep feature: the standard deviation of the median heart rate (HR MED.sd (sleep)). Interestingly, most of the 4 top ranked features come from sleep sessions, with the exception of the linear coefficient to a linear fit to the time course of the standard deviation of ACC, denoted ACC SD.linear.coef1 (wake). Features related to sleep heart rate variation (HR MED.sd, HR MEAN.sd) and sleep duration (total duration, offset, night duration) are most discriminative.

The results of running full timecourse sleep/wake segmentation reveals striking temporal differences between shedders

TABLE IV: Top 4 E4 wake/sleep features for predicting clinical outcome within 24 hours (later than day 3) of inoculation day (day 2) based on device data collected over the range 0-60 hours (day 1 and 2).

logistic regression model			continuation-ratio regression model				
Feature	Coef.	AUC	Feature	Coef.	AUC		
					Early	Late	No onset
HR MED.sd (sleep)	-3.921	0.758	HR MED.sd (sleep)	-5.073	0.882	0.718	0.864
ACC SD.linear.coef1 (wake)	-14.468	0.737	HR MEAN.sd (sleep)	-4.466	0.706	0.628	0.773
HR MED.quad.coef2 (sleep)	4.318	0.707	Total duration	-0.599	0.676	0.551	0.750
Offset	-1.452	0.697	Night duration	-0.616	0.647	0.500	0.705
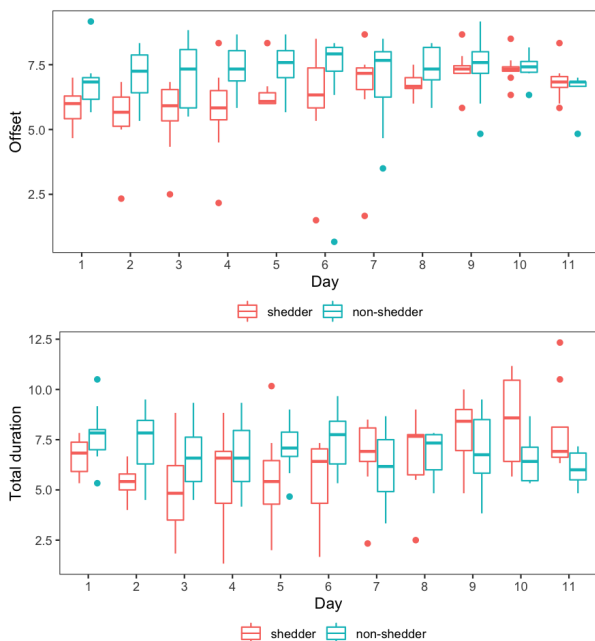
Fig. 3: Offset (wake-up time) and total sleep duration (night and day sleeping) features extracted by the proposed HMM-FLDA pipeline applied to the experimental human viral challenge study (HVC) when all time points (0-270 hours) are available. Viral inoculation took place on day 2 and all infected people (shedders) started shedding virus on day 4 or day 5.

and non-shedder features. Figure 3 indicates qualitatively different total sleep duration and sleep offset behaviors in these two groups over the full 11 day duration of the study. The non-shedder group sleeps on the average 2.041 hours longer than does the shedder group (p-value < 0.005 using two-sided t-test) over this time period. This is consistent with studies of the effects of sleep on the course of respiratory infection [16]. The mean sleep duration deficit among the shedders gradually decreases over time, and the trend reverses after day 6. The offset feature illustrates that non-shedders tend to wake-up later than the shedders until day 9 of the study.

V. DISCUSSION AND CONCLUSIONS

We have developed an unsupervised adaptive algorithm for classification of latent event states from multivariate physio-logical data collected from a wearable device. The algorithm adapts to perturbations of the initial training distributions using a sequential transfer learning model to mitigate covariate shift. The proposed algorithm operates without requiring *a priori* information about true sleep/wake states and is capable of automatically detecting anomalies and abnormal data records. Numerical simulations established significant advantages of our model relative to hidden Markov approaches to hidden event classification.

The results presented in this paper are not without limitations. The HVC experiment has the limitation of small sample size and the classes are imbalanced. The negative effect of class imbalance can be compensated, to an extent, using methods such as the synthetic minority over-sampling technique (SMOTE) [11]. While no substitute for increasing actual sample size, we have demonstrated that SMOTE can improve clinical outcome prediction accuracy as measured by AUC (Table XI in Appendix E). Another limitation is that the HVC experiment lacks ground truth information about the true sleep/wake states of the subjects. The simulation study we presented emulating a similar HVC experiment with ground truth is an in-silico validation but a controlled experiment, e.g., performed in a sleep lab, would provide better confirmation. It would also be worthwhile to test the algorithm in a larger scale experiment that collects self-reported sleep diaries in addition to clinical data.

There are also limitations of the proposed HMM-FLDA adaptive event segmentation algorithm. First, the algorithm may fail if there is an abrupt and excessively large shift in the event class distributions from time to time. More generally, loss of track due to abrupt changes is a limitation of the gradual self-training approach [37],[12] commonly adopted in transfer learning. In extreme cases this limitation may be insurmountable as there are fundamental theoretical limits that limit tracking ability of any adaptive algorithm [25]. If the abrupt shift in distributions persists over time, a possible remedy would be to episodically re-initialize the HMM-FLDA algorithm during the adaptation phase. Secondly, the classification accuracy of FLDA may be poor if the event classes are not linearly separable. At the possible cost of reduced simplicity of implementation, use of a non-linear classifier in place of FLDA would overcome this limitation, e.g., using kernelized FLDA or a Support Vector Machine (SVM) [20].

We conclude by pointing out that the proposed framework for adaptive multi-channel event segmentation and feature extraction easily generalizes beyond the setting of the binary sleep/wake segmentation illustrated in this paper. Monitoring different stages of sleep or different wake activity types would be a natural non-binary extension. With the continuing advances in the capabilities of wearable devices for digital health, many new applications will be enabled by continuous multi-event tracking. These could include accurate behavioral and health assessment tools that will advance personalized health care.

The R code and HVC data used for this paper has been made publicly available at GitLab (gitlab.eecs.umich.edu/yayazhai/ shezhai_bme2020).

VI. Disclosures

Human Subjects: The data analysis reported in Section IV of this paper followed a protocol that was submitted to the Internal Review Boards of Duke University and the University of Michigan. On 4/25/2017 the Internal Review Board of Duke University Health Sciences (DUHS IRB) determined that the protocol meets the definition of research not involving human subjects, as described in 45CFR46.102(f), 21 CFR56.102(e) and 21CFR812.3(p), and that the protocol satisfies the Privacy Rule, as described in 45CFR164.514. On 8/4/2017 the Internal Review Board of the University of Michigan (UM IRB) determined that the research performed under this protocol was not regulated.

Conflict-of-interest: The authors have no conflicts of interest to disclose.

References

[1] S. Adak, "Time-dependent spectral analysis of nonstationary time series," *Journal of the American Statistical Association*, vol. 93, no. 444, pp. 1488–1501, 1998.

[2] A. Agresti, *Analysis of ordinal categorical data*. John Wiley & Sons, 2010, vol. 656.

[3] V. V. Anisimov, *Switching processes in queueing models*. John Wiley & Sons, 2008.

[4] G. Anthony and H. Ruther, "Comparison of feature selection techniques for svm classification," in *10th International Symposium on Physical Measurements and Signatures in Remote Sensing*, 2007, pp. 1–6.

[5] C. Bazot, N. Dobigeon, J.-Y. Tourneret, and A. O. Hero, "Bayesian linear unmixing of time-evolving gene expression data using a hidden markov model," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 944–948.

[6] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: methods, systems and tools," *Ieee communications surveys & tutorials*, vol. 16, no. 1, pp. 303–336, 2013.

[7] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[8] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Computer methods and programs in biomedicine*, vol. 140, pp. 77–91, 2017.

[9] M. A. Carskadon, W. C. Dement *et al.*, "Normal human sleep: an overview," *Principles and practice of sleep medicine*, vol. 4, pp. 13–23, 2005.

[10] J. O. Cerdeira, M. J. Martins, and P. C. Silva, "A combinatorial approach to assess the separability of clusters," *Journal of classification*, vol. 29, no. 1, pp. 7–22, 2012.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[12] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Advances in neural information processing systems*, 2011, pp. 2456–2464.

[13] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979.

[14] A. L. Coelho and C. A. Lima, "Assessing fractal dimension methods as feature extractors for emg signal classification," *Engineering Applications of Artificial Intelligence*, vol. 36, pp. 81–98, 2014.

[15] S. Cohen, W. J. Doyle, C. M. Alper, D. Janicki-Deverts, and R. B. Turner, "Sleep habits and susceptibility to the common cold," *Archives of internal medicine*, vol. 169, no. 1, pp. 62–67, 2009.

[16] C. L. Drake, T. A. Roehrs, H. Royer, G. Koshorek, R. B. Turner, and T. Roth, "Effects of an experimentally induced rhinovirus cold on sleep, performance, and daytime alertness," *Physiology & behavior*, vol. 71, no. 1-2, pp. 75–81, 2000.

[17] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution," *Personalized medicine*, vol. 15, no. 5, pp. 429–448, 2018.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[19] A. Flexerand, G. Dorffner, P. Sykacekand, and I. Rezek, "An automatic, continuous and probabilistic sleep stager based on a hidden markov model," *Applied Artificial Intelligence*, vol. 16, no. 3, pp. 199–207, 2002.

[20] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, NY, 2001.

[21] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys*, vol. 46, no. 4, p. 44, 2014.

[22] X. Gao and P. X.-K. Song, "Composite likelihood em algorithm with applications to multivariate hidden markov model," *Statistica Sinica*, pp. 165–185, 2011.

[23] J. Greene, "Feature subset selection using Thornton's separability index and its applicability to a number of sparse proximity-based classifiers," in *Proceedings of Annual Symposium of the Pattern Recognition Association of South Africa*, 2001.

[24] X. He, E. H. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan *et al.*, "Temporal dynamics in viral shedding and transmissibility of covid-19," *Nature medicine*, vol. 26, no. 5, pp. 672–675, 2020.

[25] D. P. Helmbold and P. M. Long, "Tracking drifting concepts by minimizing disagreements," *Machine learning*, vol. 14, no. 1, pp. 27–45, 1994.

[26] B. T. Hemker, L. A. van der Ark, and K. Sijtsma, "On measurement properties of continuation ratio models," *Psychometrika*, vol. 66, no. 4, pp. 487–506, 2001.

[27] R. B. Henderson, J. A. Hobbs, M. Mathies, and N. Hogg, "Rapid recruitment of inflammatory monocytes is independent of neutrophil migration," *Blood*, vol. 102, no. 1, pp. 328–335, 2003.

[28] D. C. Hoaglin, B. Iglewicz, and J. W. Tukey, "Performance of some resistant rules for outlier labeling," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 991–999, 1986.

[29] Y. Huang, A. K. Zaas, A. Rao, N. Dobigeon, P. J. Woolf, T. Veldman, N. C. Øien, M. T. McClain, J. B. Varkey, B. Nicholson *et al.*, "Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection," *PLoS Genet*, vol. 7, no. 8, p. e1002234, 2011.

[30] M. R. Irwin, "Why sleep is important for health: a psychoneuroimmunology perspective," *Annual review of psychology*, vol. 66, 2015.

[31] D. Y. Kang, P. N. DeYoung, A. Malhotra, R. L. Owens, and T. P. Coleman, "A state space and density estimation framework for sleep staging in obstructive sleep apnea," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 6, pp. 1201–1212, 2018.

[32] K. A. Kaplan, J. Hirshman, B. Hernandez, M. L. Stefanick, A. R. Hoffman, S. Redline, S. Ancoli-Israel, K. Stone, L. Friedman, J. M. Zeitzer *et al.*, "When a gold standard isnâĂŹt so golden: Lack of prediction of subjective sleep quality from sleep polysomnography," *Biological psychology*, vol. 123, pp. 37–46, 2017.

[33] W. Karlen and D. Floreano, "Adaptive sleep–wake discrimination for wearable devices," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 4, pp. 920–926, 2011.

[34] W. Karlen, C. Mattiussi, and D. Floreano, "Improving actigraph sleep/wake classification with cardio-respiratory signals," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008, pp. 5262–5265.

[35] R. Klinkenberg and T. Joachims, "Detecting concept drift with support vector machines," in *Proceedings of the 17th International Conference on Machine Learning*, 2000, pp. 487–494.

[36] J. M. Kortelainen, M. O. Mendez, A. M. Bianchi, M. Matteucci, and S. Cerutti, "Sleep staging based on signals acquired through bed sensor," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 776–785, 2010.

[37] A. Kumar, T. Ma, and P. Liang, "Understanding self-training for gradual domain adaptation," *arXiv preprint arXiv:2002.11361*, 2020.

[38] L. I. Kuncheva and I. Žliobaitė, "On the window size for classification in changing environments," *Intelligent Data Analysis*, vol. 13, no. 6, pp. 861–872, 2009.

[39] T. C. Lystig and J. P. Hughes, "Exact computation of the observed information matrix for hidden markov models," *Journal of Computational and Graphical Statistics*, vol. 11, no. 3, pp. 678–689, 2002.

[40] A. Malhotra, M. Younes, S. T. Kuna, R. Benca, C. A. Kushida, J. Walsh, A. Hanlon, B. Staley, A. I. Pack, and G. W. Pien, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *Sleep*, vol. 36, no. 4, pp. 573–582, 2013.

[41] M. A. Miller, "The role of sleep and sleep disorders in the development, diagnosis, and management of neurocognitive disorders," *Frontiers in neurology*, vol. 6, p. 224, 2015.

[42] A. K. Mishra and S. Raghav, "Local fractal dimension based ecg arrhythmia classification," *Biomedical Signal Processing and Control*, vol. 5, no. 2, pp. 114–123, 2010.

[43] J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[44] S. J. Pan, Q. Yang *et al.*, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[45] J. S. M. Peiris, C.-M. Chu, V. C.-C. Cheng, K. Chan, I. Hung, L. L. Poon, K.-I. Law, B. Tang, T. Hon, C. Chan *et al.*, "Clinical progression and viral load in a community outbreak of coronavirus-associated sars pneumonia: a prospective study," *The Lancet*, vol. 361, no. 9371, pp. 1767–1772, 2003.

[46] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system of sleep stages in human subjects," Brain Information Service, University of California, Los Angeles.

[47] L. Rossi, J. Walker, and M. Musolesi, "Spatio-temporal techniques for user identification by means of gps mobility data," *EPJ Data Science*, vol. 4, no. 1, pp. 1–16, 2015.

[48] S. Saeb, T. R. Cybulski, S. M. Schueller, K. P. Kording, and D. C. Mohr, "Scalable passive sleep monitoring using mobile phones: opportunities and obstacles," *Journal of medical Internet research*, vol. 19, no. 4, 2017.

[49] F. C. Schadd and N. Roos, "A feature selection approach for anchor evaluation in ontology mapping," in *International Conference on Knowledge Engineering and the Semantic Web*. Springer, 2014, pp. 160–174.

[50] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[51] D. U. Silverthorn, W. C. Ober, C. W. Garrison, A. C. Silverthorn, and B. R. Johnson, *Human physiology: an integrated approach*. Pearson/Benjamin Cummings San Francisco, 2010.

[52] C. Thornton, "Separability is a learner's best friend," in *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*. Springer, 1998, pp. 40–46.

[53] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "An unsupervised approach for automatic activity recognition based on hidden markov model regression," *IEEE Transactions on automation science and engineering*, vol. 10, no. 3, pp. 829–835, 2013.

[54] E. Tuci, G. Massera, and S. Nolfi, "Active categorical perception of object shapes in a simulated anthropomorphic robotic arm," *IEEE transactions on evolutionary computation*, vol. 14, no. 6, pp. 885–899, 2010.

[55] J. W. Tukey, *Exploratory data analysis*. Addison-Wesley, Reading, MA, 1977.

[56] I. Visser, M. Speekenbrink *et al.*, "depmixs4: an r package for hidden markov models," *Journal of Statistical Software*, vol. 36, no. 7, pp. 1–21, 2010.

[57] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, no. 1, pp. 69–101, 1996.

[58] J. Yu, "Health condition monitoring of machines based on hidden markov model and contribution analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 8, pp. 2200–2211, 2012.

[59] A. K. Zaas, M. Chen, J. Varkey, T. Veldman, A. O. Hero III, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams *et al.*, "Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans," *Cell host & microbe*, vol. 6, no. 3, pp. 207–217, 2009.

[60] W. Zucchini, I. L. MacDonald, and R. Langrock, *Hidden Markov models for time series: an introduction using R*. CRC press, 2017.

Adaptive multi-channel event segmentation and feature extraction for monitoring health outcomes: Appendices

Xichen She[*1], Yaya Zhai[*2], Ricardo Henao[3], Christopher W. Woods[3], Christopher Chiu[4], Geoffrey S. Ginsburg[3], Peter X.K. Song[1], and Alfred O. Hero *Fellow, IEEE*[5]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109
[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109
[3]Center for Applied Genomics and Precision Medicine, Duke University, Durham, NC 27708
[4]Department of Infectious Disease, Imperial College London, London W2 1PG
[5]Departments of Electrical Engineering and Computer Science, Biomedical Engineering, and Statistics, University of Michigan, Ann Arbor, MI 48109

APPENDIX A
HUMAN VIRAL CHALLENGE (HVC) STUDY

A human viral challenge (HVC) study was conducted in 2018 as a collaborative effort between Duke University and University College London under a grant from the Defense Advanced Research Projects Agency (DARPA) under the PROMETHEUS program. Thirty nine healthy volunteers between the ages of 18 and 55 were enrolled as participants in the study, which took place in the United Kingdom.

The HVC was divided into outpatient and confinement phases. During the confinement phase participants stayed overnight for a period of 8-11 days in total from the morning before the day the viral challenge was administered to the end of confinement. During the outpatient phase subjects were evaluated for health conditions by tests including ear, nose and throat (ENT), respiratory and cardiac assessment. In addition to other data types not relevant to this paper, wearable device data (Empatica E4) and clinical infection status data (viral shedding) were collected from the participants. All data was anonymized prior to transfer to Duke and Michigan for the analysis described in this paper.

Participant exclusion criteria: Chronic respiratory disease (asthma, COPD, rhinitis, sinusitis) in adulthood. Inhaled bronchodilator or steroid use within the last 12 months. Use of any medication or other product (prescription or over-the-counter) for symptoms of rhinitis or nasal congestion within the last 3 months. Acute upper respiratory infection (URI or sinusitis) in the past 6 weeks. Smoking in the past 6 months or >5 pack-year lifetime history. Subjects with allergic symptoms present at baseline. Clinically relevant abnormality on chest X-ray. Any ECG abnormality. Those in close domestic contact (i.e. sharing a household with, caring for, or daily face to face contact) with children under 3 years, the elderly (>65 years), immunosuppressed persons, or those with chronic respiratory disease. Subjects with known or suspected immune deficiency. Receipt of systemic glucocorticoids (in a dose $\geq$ 5 mg prednisone daily or equivalent) within one month, or any other cytotoxic or immunosuppressive drug within 6 months prior to challenge. Known IgA deficiency, immotile cilia syndrome, or KartagenerâĂŹs syndrome. History of frequent nose bleeds. Any significant medical condition or prescribed drug deemed by the study doctor to make the participant unsuitable for the study. Pregnant or breastfeeding women. Positive urine drug screen. Detectable baseline antibody titres against influenza challenge strains. History of hypersensitivity to eggs, egg proteins, gentamicin, gelatin or arginine, or with life-threatening reactions to previous influenza vaccinations.

Confinement phase study: The eight days of the confinement phase consisted of a 36 hour healthy reference time period (baseline), inoculation at 36 hours (exposure), and a post-baseline time period. The E4 data of the participants was collected over the entire period and viral shedding was measured once per day in the morning. On the second day (day 2) at approximately noon (36 hours from start of study) each participant was challenged by a GMP influenza A/California/04/2009-like (H1N1) virus strain. The inoculation was administered by inserting intra-nasal drops on a single occasion with diluted inoculum with an average dose of 10^6 TCID50 in 1mL PBS divided equally between the two nostrils. This resulted in an average attack rate of 44%. Following inoculation, advice regarding hand hygiene was given and subjects were provided with alcohol hand gel and face-masks if they moved between the inoculation room and the quarantine ward.

Viral shedding assay: Over the confinment phase of the study, viral shedding was measured once per day through a nasal lavage. The collected fluid was aliquoted into sterile microfuge tubes and centrifuged for analysis of cells, and lavage fluid was later analysed by singleplex PCR to quantify the degree of viral shedding of the inoculated strain. Multiplex PCR was performed on the pre-inoculation lavage and post-inoculation lavage collected during the study to exclude the presence of other respiratory viruses.

Wearable device and protocol: Over the confinement phase of the study, subjects agreed to comply with the following wearable device protocol: 1) they wear the Empatica E4 device properly, i.e., comfortably tight on the wrist of their dominant hand; 2) they take care to maintain the device and protect it from shocks, water immersion, and other damage; 3) they wear the Empatica continuously without interruption except for periods that they were showering, recharging, or uploading data. The E4 has several sensors that measure physiological parameters including blood volume pulse, skin conductance, temperature, and movement. The E4 was recharged once per day during which time each participant's data were uploaded to a cloud server for processing using Empatica proprietary software.

The result of this processing was reported at sub-second temporal resolution as the following variables: heart rate (1sa/sec), skin temperature (4sa/sec), electrodermal activity (4sa/sec), and 3 axis accelerometer (32sa/sec). These were mapped to a vector of four variables at each time point: HR, TEMP, EDA and ACC, respectively, where ACC was computed as the Euclidean norm (magnitude) of the 3 dimensional acceleration vector. Subjects were trained on best practices for wearing and maintaining the E4 devices over the course of the confinement phase of the study.

Of the 39 participants enrolled in the HVC, only 25 had sufficient quality E4 data to be included in the analysis presented in Section IV. A histogram of the demographic data of these 25 participants is shown in the figure below. The other participants had wearable data that suffered from factors such as device failure, excessive missingness or data corruption, making their data unusable for our analysis. Non-compliance with wearable device protocols, data upload errors, and device malfunction were the cause of most of these problems.
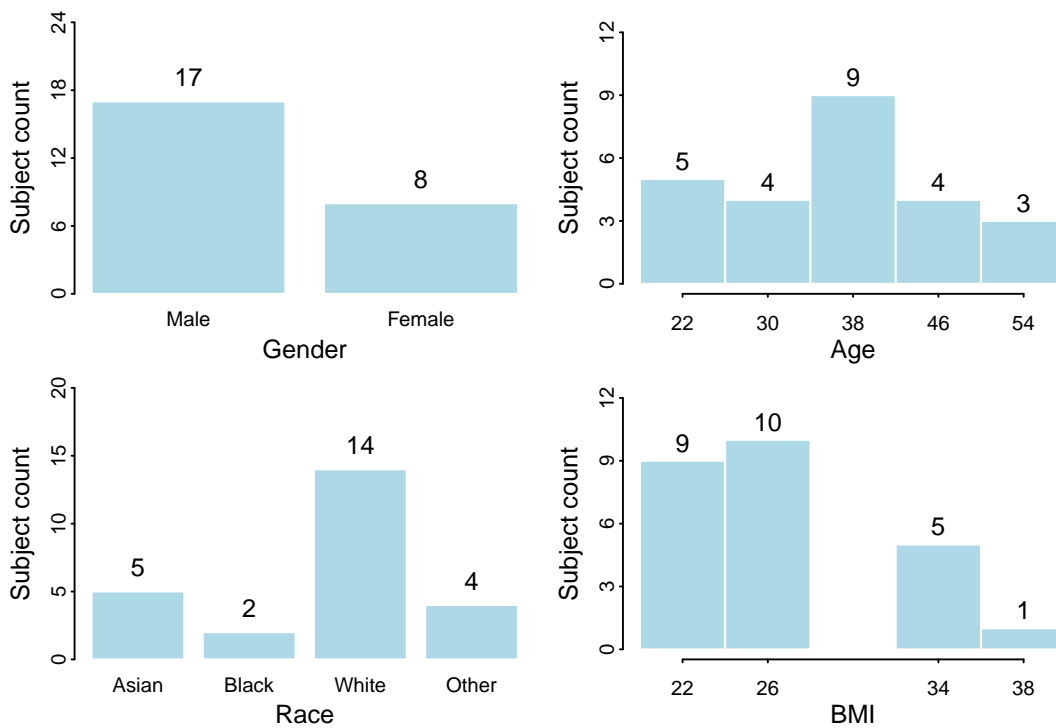
Demographics of 25 participants in HVC study.

APPENDIX B
HVC DATA

The raw time course E4 data of two representative subjects in the HVC study are shown in Fig. 4. Subject 1 (Fig. 4a) did not become infected after inoculation on the second day of the study (day 2), i.e., this subject had no detectable level of viral shedding at any time over the 8-11 days of the study, while Subject 2 (Fig. 4b) became infected. The four E4 signals shown in the figure are heart rate (HR), accelerometer (ACC), temperature (TEMP), and electrodermal activity (EDA). The data clearly shows diurnal differences in signal behavior, corresponding to the cycling of sleep and wake states of these subjects. A quadratic trend in heart rate and temperature is clearly visible in Subject 2, the trend peaking at around 144 hours. Robustness to this trend is desirable and motivated the proposed HMM-FLDA sequential adaptive sleep/wake segmentation algorithm.

APPENDIX C
SIMULATION STUDY

In this section we include additional figures supporting the simulation study reported in Table I and described in Section III of the main text. We also report on in-sample performance comparisons between the proposed HMM-FLDA and the competing
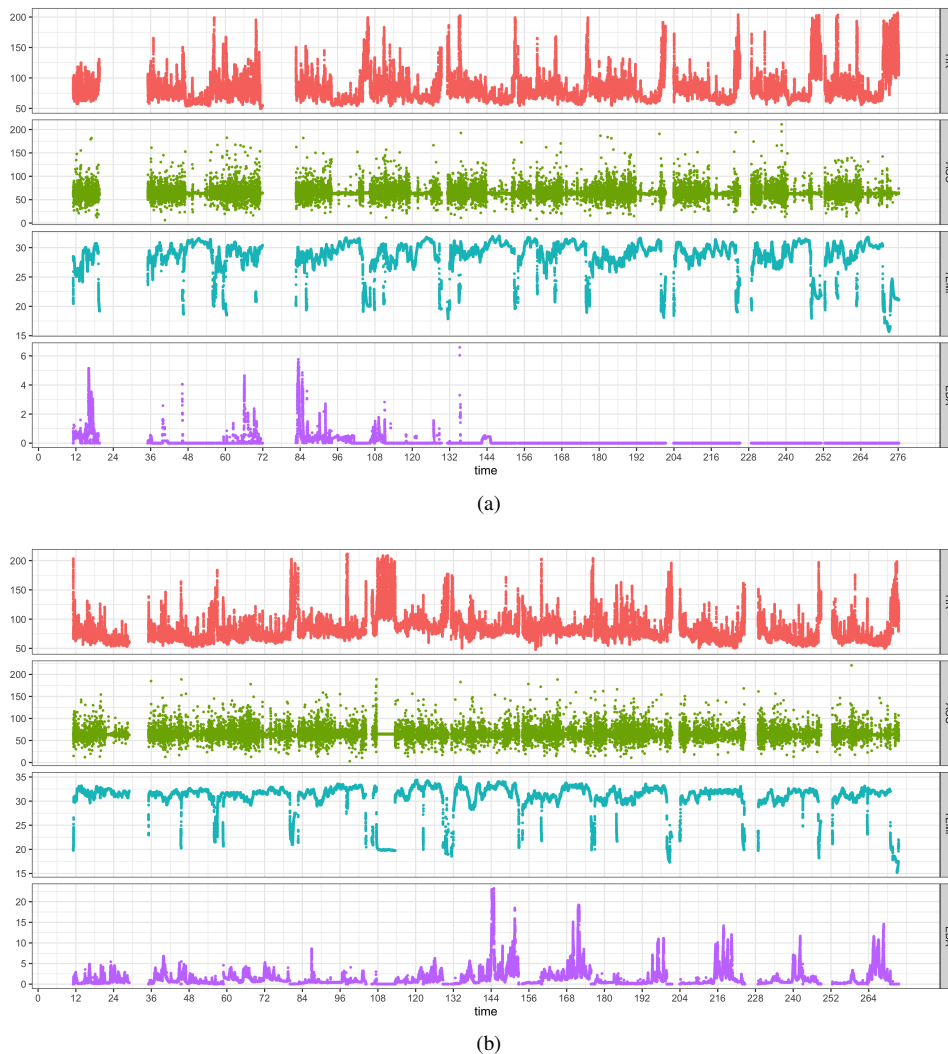
(a)

(b)

Fig. 4: Four channel device data (Empatica E4) from human viral challenge (HVC) study for two subjects (Subjects 1 and 2). The four channels are: HR (red), ACC (green), TEMP (blue) and EDA (purple). Time 0hrs corresponds to 12am local time on the first day (day 1) of the study. Viral inoculations were administered to all subjects on the morning of the second day (day 2), i.e., between 32hrs and 36hrs. **(a)** data from a Subject 1 who had no detected shedding (Non-infected class); **(b)** data from Subject 2 for whom shedding was detected (Infected class).

HMM and dHMM algorithms.

As mentioned in Sec. III these simulations were intended to emulate the experimental HVC data used in Sec IV. In Fig. 5 we show realizations of two simulated device channels, $X1$ and $X2$, emulating the HR MED and ACC SD features in the HVC study, under the unstable++, and unstable+- models for mean trends occurring after 36 hours. See Figs. 4a and 4b for comparison to real E4 data.

Table V provides a more comprehensive report of the results of our simulations of the proposed HMM-FLDA event segmentation than does Table I in the main text. Three segmentation algorithms are compared: the proposed HMM-FLDA, the HMM and a detrended HMM (dHMM) using LOESS pre-filter to detrend prior to applying HMM. The simulation parameters for the random wake/sleep event transitions, event conditioned data distributions, and the covariate shift are the same as were used for Table I.

The top part of the table shows in-sample performance and the bottom part shows out-of-sample performance showing both mean and standard deviation (SD) of the performance measured over 100 simulation trials. Only the means of the OoS part of this table were shown in Table I. The performance of the best performing method in terms of the MEAN performance is bolded.

For the in-sample results (top part of table), event segmentation performance on the training set is shown. For the out-of-sample results (bottom part of table), an hold-out dataset was created from an independent simulation using the same model as the training set. The performance of each segmentation algorithm, trained only on the training set, was evaluated on the
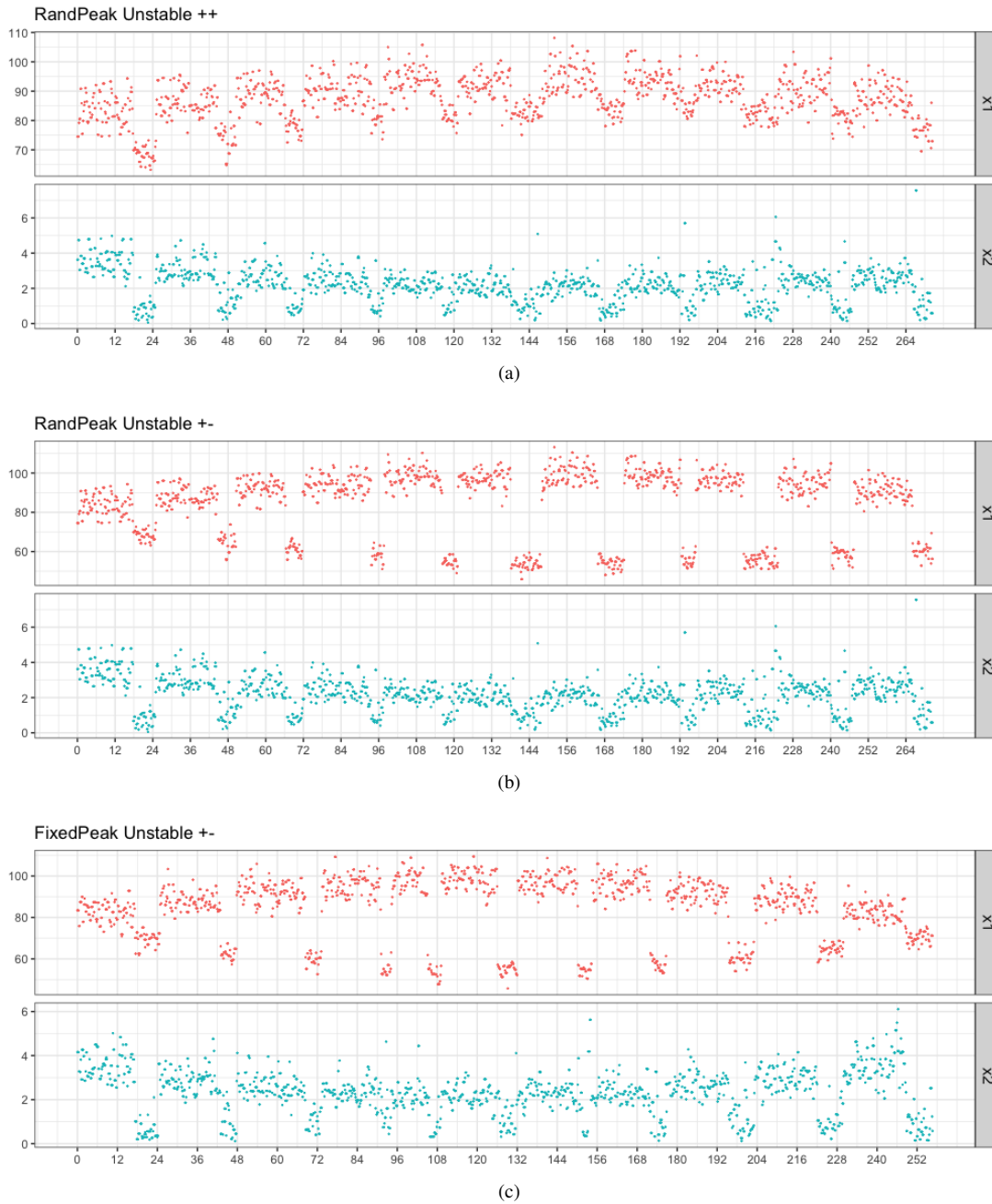
Fig. 5: Realizations of simulated signals for: (**a**) Unstable++ case for which the sleep and wake session both have concave trends after 36 hours in channel $X1$; (**b**) Unstable+- case where the sleep and wake sessions have convex and concave trends, respectively, in channel $X1$; (**c**) another realization of Unstable+- case with sleep and wake sessions whose trends hit their apogee at a different time.

hold-out set. Note that the proposed HMM-FLDA would never be implemented in this way as such an implementation would turn off the adaptation mechanism on future data. The bottom part (OoS) of Table V provides evidence for the accuracy and robustness of the proposed approach.

The bolded entries in Table V denote the best mean performance per model (stable, unstable++, unstable+-) and performance criterion (Accuracy, F1, Cosine dist, Onset diff, and Duration diff). The asterisk on an MEAN entry in the OoS part of Table V indicates the best performing method for each of the five criteria where best was determined using a one-sided paired t-test of significance applied to the set of 1000 simulations thresholded at a 0.01 level of significance. The OoS part of Table V has an additional row, P-VALUE, that is the p-value of the one-sided paired t-test that the proposed HMM-FLDA has better performance than both the HMM and dHMM. Only two P-VALUE entries are not significant, the case unstable++ for cosine distance and the case unstable+- for onset diff. For the former case, dHMM is better than HMM but dHMM cannot

be characterized as better than the proposed HMM-FLDA (0.01 level of significance). For the latter, HMM is better than the proposed for onset diff estimation. Hence, except for the onset diff criterion in the unstable+- case, the proposed method is equivalent or better than the other methods at a 0.01 level of significance.

TABLE V: Performance of the proposed adaptive transfer learning algorithm as compared with standard HMM operating on the original data (HMM) and operating on LOESS detrended data (dHMM).

Setting	Methods		Accuracy	F1	Cosine dis.	Onset diff.	Duration diff.
(0, 0, 0, 0)	HMM	MEAN	**0.9982**	**0.9987**	**0.9968**	**0.0236**	0.0427
stable		SD	**0.0012**	**0.0009**	**0.0021**	**0.0199**	0.0287
	dHMM	MEAN	0.9981	0.9986	0.9966	0.0271	0.0450
		SD	0.0012	0.0009	0.0022	0.0209	0.0288
	Proposed	MEAN	0.9981	0.9987	0.9967	0.0259	**0.0420**
		SD	0.0012	0.0009	0.0021	0.0218	**0.0278**
(15, 10, 0.5, -0.5)	HMM	MEAN	0.8744	0.9052	0.8234	2.4786	3.6780
unstable++		SD	0.0642	0.0520	0.0754	1.5541	2.2121
	dHMM	MEAN	**0.9558**	**0.9680**	**0.9306**	0.9912	**1.1574**
		SD	**0.0502**	**0.0375**	**0.0713**	1.6795	**1.4117**
	Proposed	MEAN	0.9487	0.9636	0.9146	**0.9813**	1.4605
		SD	0.0498	0.0363	0.0781	**0.9889**	1.3812
(-15, 15, 0.5, -0.5)	HMM	MEAN	0.9371	0.9513	0.9185	0.4329	2.0992
unstable+-		SD	0.0836	0.0660	0.1043	0.7836	2.7722
	dHMM	MEAN	0.9483	0.9615	0.9253	0.4909	1.4544
		SD	0.0666	0.0505	0.0928	0.9295	1.8854
	Proposed	MEAN	**0.9923**	**0.9946**	**0.9868**	**0.1631**	**0.1699**
		SD	**0.0221**	**0.0157**	**0.0365**	**0.4648**	**0.4651**
stable	HMM (OoS)	MEAN	**0.9981**	**0.9987**	**0.9968**	**0.0165**	**0.0307**
		SD	**0.0005**	**0.0004**	**0.0009**	**0.0051**	**0.0079**
	dHMM (OoS)	MEAN	0.9981	0.9986	0.9967	0.0253	0.0373
		SD	0.0015	0.0011	0.0027	0.0163	0.0211
	Proposed (OoS)	MEAN	0.9980	0.9986	0.9964	0.0168	0.0313
		SD	0.0015	0.0011	0.0027	0.0163	0.0211
		P-VALUE	1.0000	1.0000	1.0000	0.7850	0.8735
unstable++	HMM (OoS)	MEAN	0.8959	0.9209	0.8559	1.6704	2.9169
		SD	0.0679	0.0547	0.0792	1.0972	2.2120
	dHMM (OoS)	MEAN	0.9127	0.9231	**0.8957***	2.4709	2.2653
		SD	0.1284	0.1377	**0.1089**	2.6125	3.1967
	Proposed (OoS)	MEAN	**0.9356***	**0.9544***	0.8904*	**1.0679***	**1.5756***
		SD	**0.0353**	**0.0263**	0.0533	**0.7614**	**1.0664**
		P-VALUE	9.94E-07	1.07E-10	0.9521	4.04E-28	3.73E-08
unstable+-	HMM (OoS)	MEAN	0.9369	0.9510	0.9187	**0.2763***	1.9750
		SD	0.0810	0.0635	0.1019	**0.4079**	2.5778
	dHMM (OoS)	MEAN	0.9271	0.9370	0.9114	1.0170	2.0131
		SD	0.1151	0.1298	0.1112	2.1413	3.2464
	Proposed (OoS)	MEAN	**0.9758***	**0.9833***	**0.9564***	0.3712	**0.3817***
		SD	**0.0326**	**0.0226**	**0.0591**	0.6225	**0.5472**
		P-VALUE	1.63E-31	9.69E-37	5.80E-16	0.9999	1.34E-61

In Fig. 6 is shown boxplots and density plots of the out-of-sample (OoS) simulation of all three scenarios, whose mean is shown in Table I. These plots show the distribution of errors committed by the various methods compared in terms of event labeling accuracy, onset estimation error, and duration estimation error when the trained methods are applied to independent sample trajectories drawn from the same distribution. As compared to the others the proposed method has a distribution that is more highly concentrated and has fewer outliers.
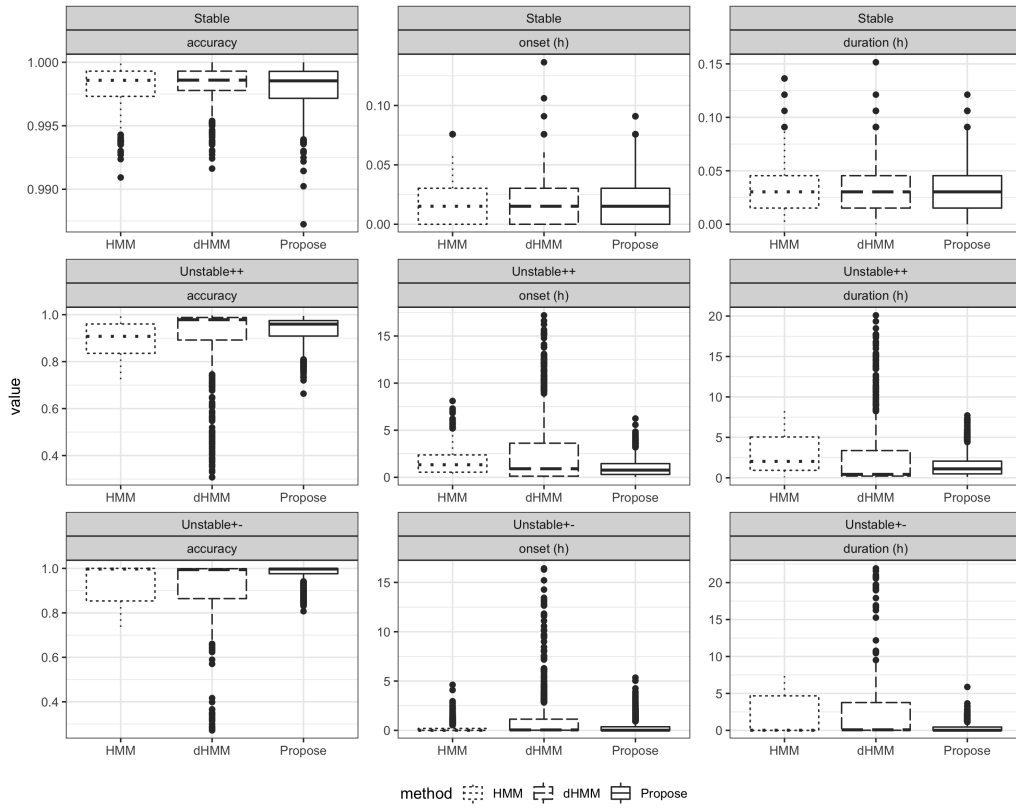
(a)
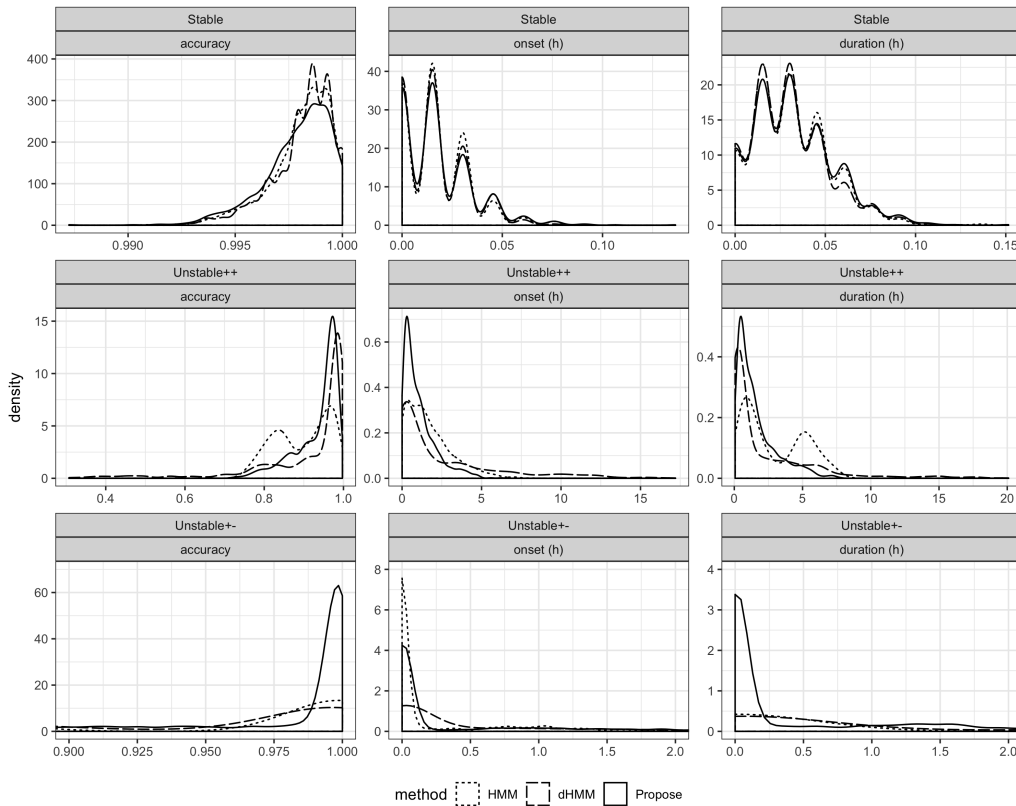
(b)

Fig. 6: (**a**) Boxplots for label accuracy, onset estimation error, duration estimation error for simulations shown in Table I for out-of-sample (OoS) simulation performance (excluding two outliers by dHMM method in the stable case). (**b**) corresponding interpolated densities associated with (**a**) (for unstable+- case, the density plots are shown for a zoomed in region to better compare density concentration).

Table VI shows runtime comparisons. The proposed method runs slower than HMM and dHMM. However, 1 minute run time of HMM-FLDA is still small as compared to the 8-11 days time period of HVC data. Furthermore, the method is not optimized in terms of runtime and can likely be accelerated.

TABLE VI: Average run time (seconds) per replication of the proposed adaptive transfer learning algorithm as compared with standard HMM operating on the original data (HMM) and operating on LOESS detrended data (dHMM) on a 2.3 GHz Dual-Core Intel Core i5 processor with 16 GB RAM.

	stable	unstable++	unstable+-
HMM	0.4347	0.5612	0.4913
dHMM	0.7396	0.8351	0.8347
Proposed	53.5126	54.3654	55.2601

APPENDIX D
ANALYSIS PIPELINE IMPLEMENTATION FOR HVC STUDY

Here we provide additional details on our implementation of the pipeline of Fig. 1b, in the context of the Human Viral Challenge Study Empatica E4 data. We illustrate the implementation for both the pre-infection time-line (0-60 hours) and the full time-line (0-270 hours). First we describe the features used for abnormality detection.

Wearable devices are subject to outliers, anomalies and other abnormal sensor readings. Some types of outliers are physiological and are important to include in the final event-segmented data stream. Other types of outliers are technical and can be due to device malfunction or improper wearing of the device. These technical outliers must be removed early in the analysis pipeline so as to not compromise downstream event detection and labeling performance. We use a two stage procedure for isolating such anomalies in the pre-processing stage (Stage 1) of the pipeline in Fig. 1b and classifying them as physiological vs technical outliers for possible re-insertion in the transfer learning stage (Stage 2). The first stage of the procedure is called *abnormality filtering* and the second stage is called *abnormality classification*, which will be discussed below. Both procedures apply standard outlier detection methods to a set of predefined features. However, since the purposes of these procedures are different, the outlier detection thresholds are different.

Features for abnormality filtering (Stage 1) and classification (Stage 2): The best features to use for detection of abnormal samples will be experiment dependent and device dependent. Here we explain how the abnormality filtering features were selected for the HVC experiment where participants wore the Empatica E4 device, as discussed in Section IV. Based on experiments on an Empatica E4 performed in our laboratory we determined that there are three principal causes for abnormal measurements, each manifesting abnormality in different combinations of channels. See Fig. 7. For abnormality filtering and classification physical intuition motivated us to select three temporally localized features, HR MED, TEMP MED and ACC SD, as they are especially affected by the types of abnormalities described below.

- Device not worn (NW). *Effect*: Skin temperature sensor (TEMP) reads ambient temperature and activity sensor (ACC) records little or no physical movement. *Primary features affected*: median of TEMP and standard deviation of ACC magnitudes are abnormal.
- Device loss of contact (LOC). *Effect*: Intermittent skin contact causes spurious signal dropout. *Primary features affected*: median of TEMP and HR and standard deviation of ACC magnitudes are abnormal.
- Subject engages in intense activity (Active). *Effect*: heart rate (HR) and activity sensor (ACC) readings increase significantly over burst of physical activity. *Features primarily affected*: median of HR and standard deviation of ACC magnitudes are abnormal.

In the HVC study we used HR MED and TEMP MED features for abnormality filtering while we used HR MED, TEMP MED and ACC MED for abnormality classification. The combination of HR MED and ACC MED is especially important for classifying abnormalities that are due to physiological causes,e.g., when a subject is engaged in intense exercise which is legitimate wake session activity that should be reinserted in the final segmented data stream in Stage 2 of the pipeline.

Figure 8 illustrates three simulated cases with different levels of separability. Two variables (X_1, X_2) are generated from $\mathrm{BVN}(\mu_1, \mu_2, 1, 1, 0)$. For samples in Class 0 (indicated by red dots), $\mu_1 = \mu_2 = 0$, while for samples in Class 1 (indicated by green triangles), we considered three settings: (1) $\mu_1 = \mu_2 = 0$; (2) $\mu_1 = \mu_2 = 1.5$; (3) $\mu_1 = \mu_2 = 3$, corresponding to non-, weakly and strongly separable scenarios. SI values based on both projection distance and Euclidean distance are reported for each case. We observe that SI is indeed able to effectively characterize separability, and that the projection distance is preferred since it gives value closer to 0.5 in the non-separable case and value closer to 1 in the strongly separable case.
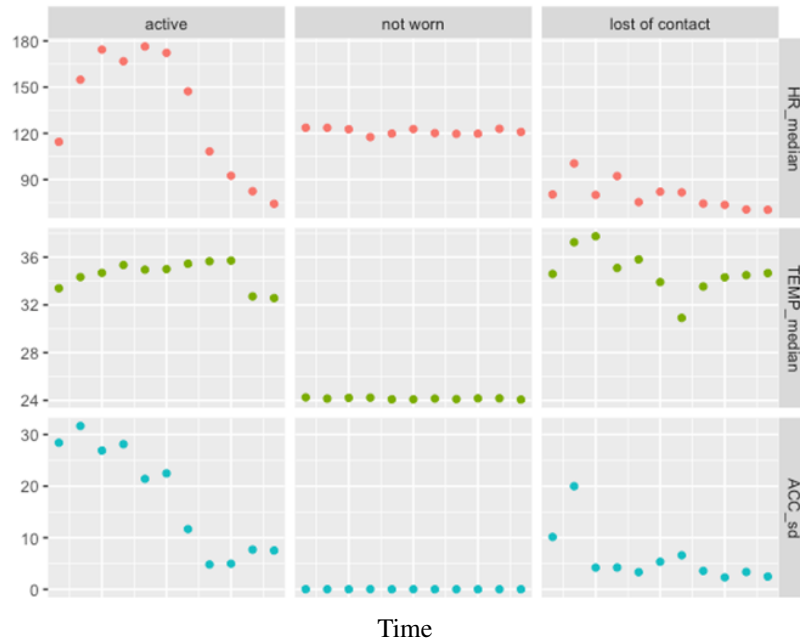
Fig. 7: Experimental Empatica E4 data collected from one of the co-authors of this paper under three different conditions: active, not worn and loss of contact. Shown are measured values of the variables used in our selected abnormality feature set: HR median, TEMP median and ACC sd. Note the very different values of these three parameters between the active, not worn and loss of contact classes.
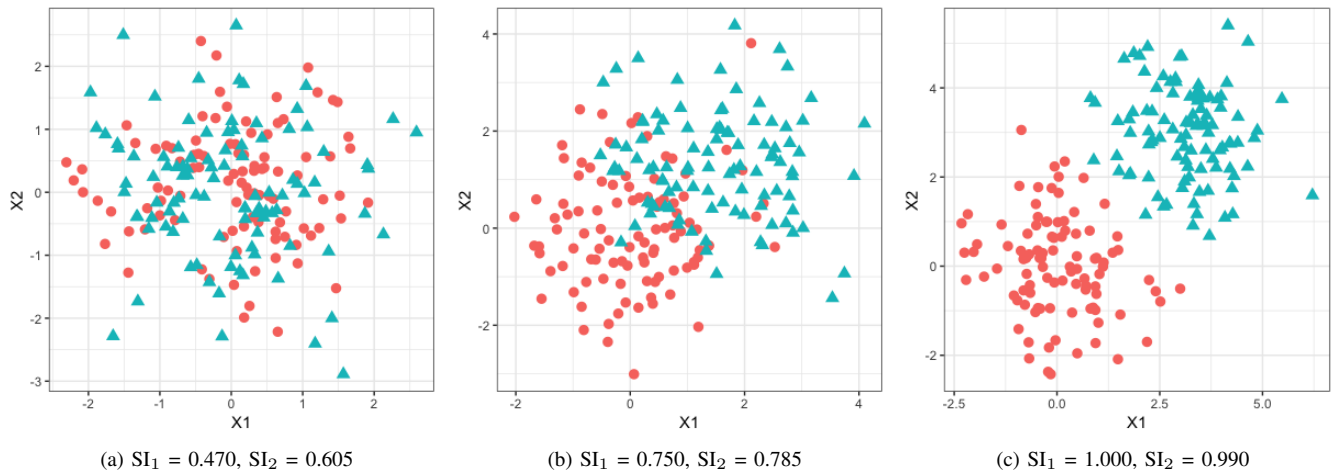

(a) $SI_1 = 0.470$, $SI_2 = 0.605$ (b) $SI_1 = 0.750$, $SI_2 = 0.785$ (c) $SI_1 = 1.000$, $SI_2 = 0.990$

Fig. 8: Separability indices under three simulated scenarios: **(a)** non-separable, **(b)** weakly separable, **(c)** strongly separable, where SI_1 is based on projection distance and SI_2 is based on Euclidean distance.

A. Online feature extraction pipeline: applied to 0-60 hours of data

Here we describe details of the proposed pipeline, illustrating with the pre-infection (0-60 hours) data. The result of applying the pipeline to the full data (0-270 hours) is described afterward.

1) Stage 1: pre-processing: Pre-processing in stage 1 of the pipeline accomplishes three tasks: temporal windowing and conditioning; local feature extraction, and abnormality filtering. Here we illustrate each of these tasks when the pipeline is applied to the pre-infection data, i.e., only the first 60 hours (2 days and two nights) are available. Results of applying the pipeline to the full data from 0 to 270 hours are presented for comparison in Section D-B.

Stage 1: temporal windowing and conditioning

The tuning parameters and their settings for Stage 1 of the pipeline are

- Temporal window (epoch) length: 10 mins
- Subject availability threshold: 60%
- Subject abnormality threshold: 40%

The subject availability threshold of 60% is applied to filter out subjects that have E4 data missingness of greater than 40%. Subject availability is quantified the proportion of time points not missing among the full set of sampling times (number of seconds) over the 0-60 hour period. Subjects with less than 60% data availability are not further processed. Figure 9 shows that 2 subjects in the HVC cohort (Subjects with identified 13 and 17) have insufficient data availability over this time period.
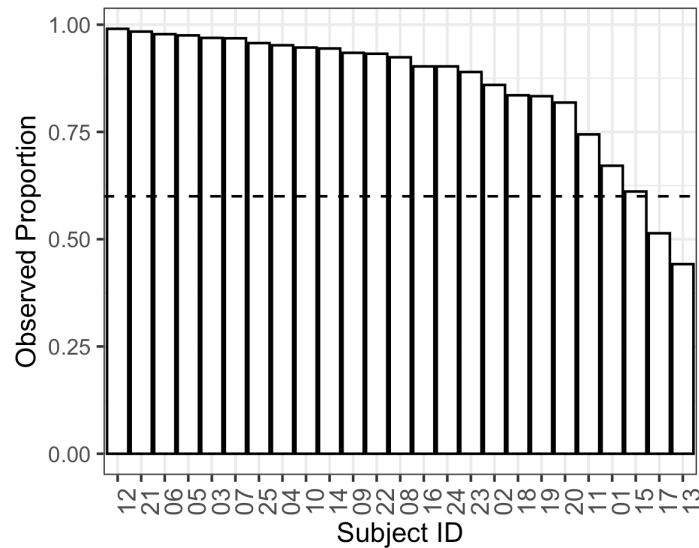

Fig. 9: Subject availability for the sleep/wake segmentation when the first 60 hours of data is available for subjects in the HVC study. Two subjects (13 and 17) have excessive ($> 40\%$) missingness. The observed proportion is defined as the ratio of the number of time samples in the subject's data record and the total number of sampling times that should be available over the 60 hour period.
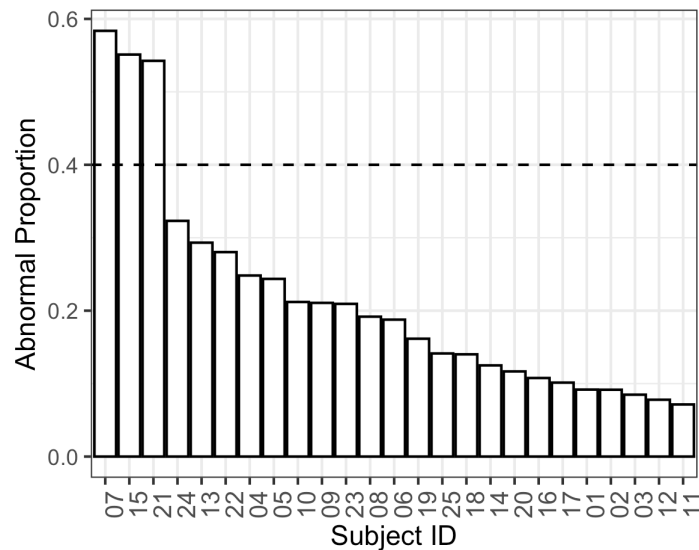
Stage 1: abnormality filtering

Fig. 10: Abnormal proportion for the sleep/wake segmentation over 0-60 hours for subjects in the HVC study. Three subjects have an excessive number of abnormal time samples (greater than 40%).

The abnormality filtering module operates as follows. After temporally localized summary statistic features are extracted from the device for each subject, these features are tested for abnormality using a clustering-based anomaly detection procedure. The procedure labels the abnormal samples as *non-normal data* and these are temporarily removed from the data stream.

The clustering-based anomaly detection procedure is as follows. Define the vector-valued local features for the i-th subject at the t-th time instant $\tilde{\boldsymbol{X}}_{i,t} = (\tilde{X}_{1i,t}, \ldots, \tilde{X}_{Mi,t})$, where here the feature dimension is $M = 12$. As explained above, only two of these features was used for abnormality filtering: HR MED and TEMP MED. For each subject i, a combination of k-means clustering and quantile thresholding is used to determined a normal region $\mathcal{C}_i$ in this two dimensional feature space. Subject i's feature instance at time t is declared non-normal if it is not in $\mathcal{C}_i$. While we also investigated abnormality using

k-means clustering in the full 2 dimensional feature space of HR MED and TEMP MED (See Fig. 15b), the abnormality filtering method we adopted in the pipeline of Fig. 1b performs k-means clustering separately on each each of the feature dimensions.

For the HVC data, where the events of interest are sleep and wake, the k-means algorithm is set to extract 3 clusters corresponding to sleep and wake (normal) and non-normal classes. Let $\{S_{im1}, S_{im2}, S_{im3}\}$ be the resulting clusters with centers (centroids) $\{\mu_{im1}, \mu_{im2}, \mu_{im3}\}$. When the normal and non-normal cluster classes are well separated it is easy to construct a normal region $\mathcal{C}_i$, e.g., all points in the feature space having a majority of k-nearest neighbors in $\{(\tilde{X})_{i,t}\}_t$ outside of the non-normal cluster. More often, however, the separation between the normal and non-normal cless is not sufficient and a different method is needed. We determine $\mathcal{C}_i$ as the set complement of the hyper-rectangle of minimal volume whose empirical coverage probability is 95%. The rectangular edge lengths and position are thus determined by the marginal sample quantiles along each feature dimension.

Specifically, we define a sequence of subject-dependent cutoff values $\{c_{i1}, \ldots, c_{iM}\}$ for abnormality. The rectangular normal region is designated as the Cartesian product $S_i^{nor} = S_{i1}^{nor} \times \ldots \times S_{iM}^{nor}$, where

$$S_{im}^{nor} = \begin{cases} S_{im1} \cup S_{im2} & \text{if } |\mu_{im2} - \mu_{im1}| < |\mu_{im3} - \mu_{im2}|, \\ S_{im1} & \text{o.w..} \end{cases}$$

Cutoff values defining the normal set are defined as

$$c_{im} = \begin{cases} Q_{0.025}(\{\tilde{X}_{mi,t} \in S_{im}^{nor}\}) & \text{if } \mu_{i,nor} > \mu_{i,abn}, \\ Q_{0.975}(\{\tilde{X}_{mi,t} \in S_{im}^{nor}\}) & \text{o.w.,} \end{cases}$$

where $Q_q(\cdot)$ is the $q-$th quantile of the samples, and $\mu_{i,nor}, \mu_{i,abn}$ are the centroids of the normal and non-normal set, respectively. For the experimental HVC data we used the 95% outlier rule to define the lower and upper quantiles as 2.5% and 97.5%, but less stringent threshold values may be adequate for other datasets.

For the subjects in HVC study with the first 60 hours of data available, Figure 10 shows a bar-plot of the resulting non-normal data proportion for each subject. Three subjects (7, 15 and 21) had excessive abnormal data and were eliminated from further processing.

Thus, when taken with subjects who had excessive missingness, a total of five subjects were excluded from further analyses.Some of these subjects may come back into the analysis when a greater amount of data is available (see subsection C for the case where 0-270 hours are available).

Feature selection for training adaptive segmentation algorithm: As explained in the main text, we extracted a subset of the 12 local features using the sleep/wake separability index (SWSI) (6). The SWSI is computed over all subjects, to contrast each feature over the putative sleep periods 2:00-5:00 and the wake periods 19:00-22:00 in the 0-60 hour time interval of the pre-infection data window. Figure 11 shows the resulting SWSI for each of the 12 features, in descending order of median. Applying the selection rule that a selected feature must be greater than 0.7 for at least 75% of the subjects, three features are above threshold: HR MEAN, HR MED and HR SD and, as HR MEAN and HR MED are highly correlated, we chose eliminated HR MEAN since it has a lower 75% quantile than does HR MED.

2) Stage 2: adaptive segmentation: The HMM-FLDA procedure is implemented for a given subject on all the time points that pass through the abnormality filter. The details of the HMM-FLDA are given in Sec IV-A2.

Stage 2: Separability index

By definition (5), SI is the proportion of samples that share the same label with their nearest neighbors, and hence SI $\in [0, 1]$. Intuitively, when samples from two classes form two tight, well-separated clusters with little overlap, the nearest neighbor of one sample from, say, Class 0, will most likely belong to Class 0 as well, which will result in a large SI value close to 1. In contrast, when samples from two classes follow exactly the same distribution, i.e., completely non-separable, then the nearest neighbor of one sample will have equal probability of being Class 0 or Class 1, and thus, the SI of these samples is close to 0.5. A large SI value implies strong separability of classes, and is usually an indication of reliable prediction. SI also depends on the measure of distance used to determine the nearest neighbors. The projection distance captures the difference among samples on the optimal direction w that is most relevant to distinguishing between the two classes, and thus is better than Euclidean distance in regard to characterizing separability, as demonstrated by Figure 8.

Once the HMM-FLDA event classification procedure terminates the initial sleep/wake segmentation will have missing time points that have been removed by the abnormality filter. Some of these abnormal time points, e.g., those due to rare physiological events like an exercise session, are reinserted into the data stream. The method to do this is based on an abnormality classification procedure discussed in the next subsection.

Stage 2: Abnormality classification and reinsertion

The purpose of the abnormality classification module is to identify, re-insert and assign an event label to physiologically meaningful non-normal data, i.e., abnormalities that are not due to technical issues associated with device failure or improper
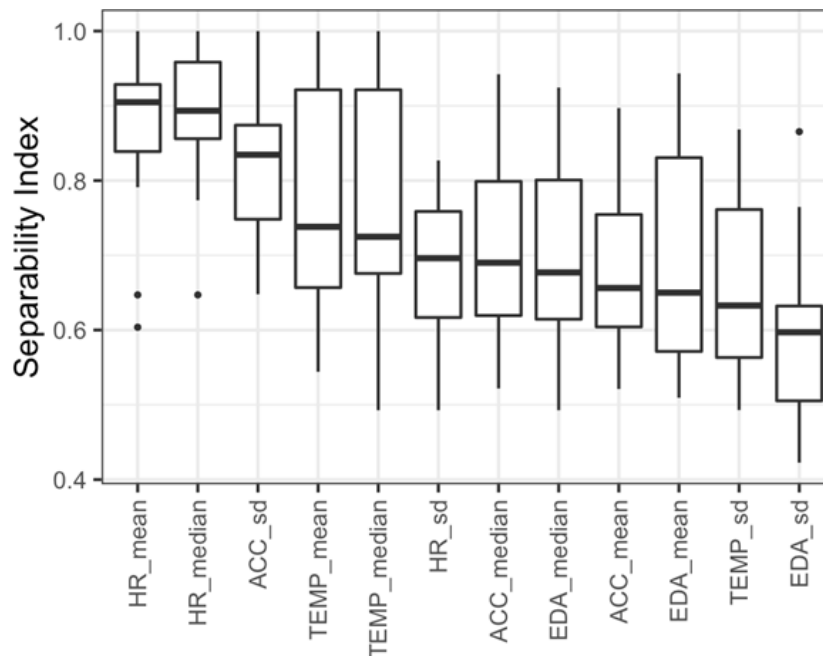
Fig. 11: Box-plot of sleep/wake separability index (6) for pre-infection (0-60 hour) data, showing the spread of the 12 local features extracted from wearable data in the first stage of the processing pipeline. Excluding HR MEAN, which is highly correlated with HR MED and has a smaller 25% quantile, two local features (HR MED, ACC SD) are selected.

wearing of the device. Any such non-normal samples re-inserted into the data stream inherit the event class label of the session into which the sample's time stamp falls.

Using predefined features HR MED, TEMP MED, and ACC SD, abnormality classification was performed using a decision tree using marginal quantile thresholding. For each non-normal sample, we applied a standard 1.5 interquartile range (IQR) test [55], [28] to each feature independently. If a feature falls outside of this range it is declared an outlier and further classified. The lower and upper endpoints of the 1.5 IQR interval are defined as:

$$
\begin{aligned}
\text{LOWER} &= Q_{0.25} - 1.5 * (Q_{0.75} - Q_{0.25}), \\
\text{UPPER} &= Q_{0.75} + 1.5 * (Q_{0.75} - Q_{0.25}).
\end{aligned}
\tag{7}
$$

Here $Q_{0.25}$ and $Q_{0.75}$ are the 25% and 75% sample quantiles of the feature empirical distributions computed over the detected sessions. All samples that were categorized as "Wake" or "Active" were incorporated back into the corresponding session. Abnormal samples assigned to other categories were discarded. The decision tree for this procedure is shown in Figure 12.

Stage 2: Median filtering

After processing by HMM-FLDA and reinsertion of physiologically meaningful abnormalities, there commonly exists short bursts of sleep sessions. We expect that some of these are actually sleep while others correspond to resting without sleep. While such bursty behavior could possibly be directly incorporated into an HMM model, e.g., using a semi-Markov switching process [3], we took a simpler approach that applies a modified median filter with 90 min smoothing window (median filter of order 9) that has the effect of merging sleep sessions shorter than 60 mins into a wake session. We choose to eliminate such short sessions in order to eliminate disambiguate resting and light naps from sleep sessions, defined as a session having deep sleep stages, i.e., Stage 3 & 4 of non-rapid eye movement (NREM), usually starting 30 minutes after sleep onset and lasting approximately 20 to 40 minutes in the first sleep cycle [9]. Hence, the 60-minute threshold is designed to eliminate putative sleep sessions that had no deep sleep stages.

B. Offline feature extraction pipeline: applied to 0-270 hours of data

1) Stage 1: pre-processing: When evaluated over the time period 0-270 hours, no subjects had missingness greater than 40% and thus none were filtered out due to inadequate data availability. However, five subjects were found to have greater than 40% abnormal samples and were filtered out. Figure 13 shows the bar-plot of the abnormal proportions for all 25 subjects, showing the 5 subjects exceeding the 40% threshold for inclusion.

To select features for the HMM-FLDA sleep/wake segmentation algorithm over the full time course of 0-270 hours, we designated two one hour periods of the day, 3:00-4:00 and 21:00-22:00, respectively, as sleep and wake (resting). As contrasted
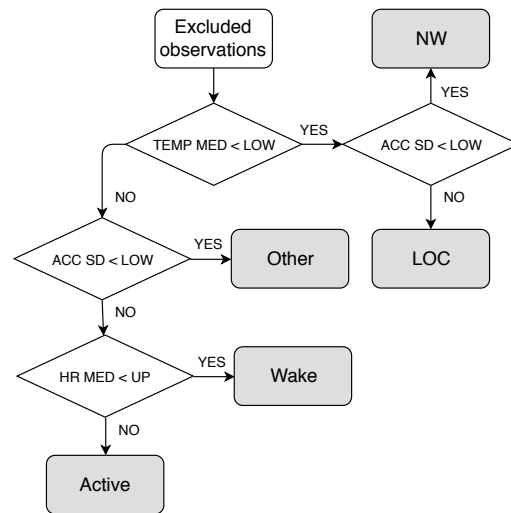
Fig. 12: Decision tree for abnormality classification module for the HVC data. Non-normal samples (excluded observations) identified by the abnormality filter in first stage of the pipeline of Fig. 1b are excluded from the training set used by the second stage transfer learning algorithm. The abnormality classifier re-evaluates these samples for possible re-insertion and labeling of physiologically meaningful abnormalities. The classifier classifies the samples into final categories of: device not worn (NW), loss of contact (LOC), Wake, Active and Other. Samples that are classified as "Active" or "Wake" are re-inserted into the corresponding session.
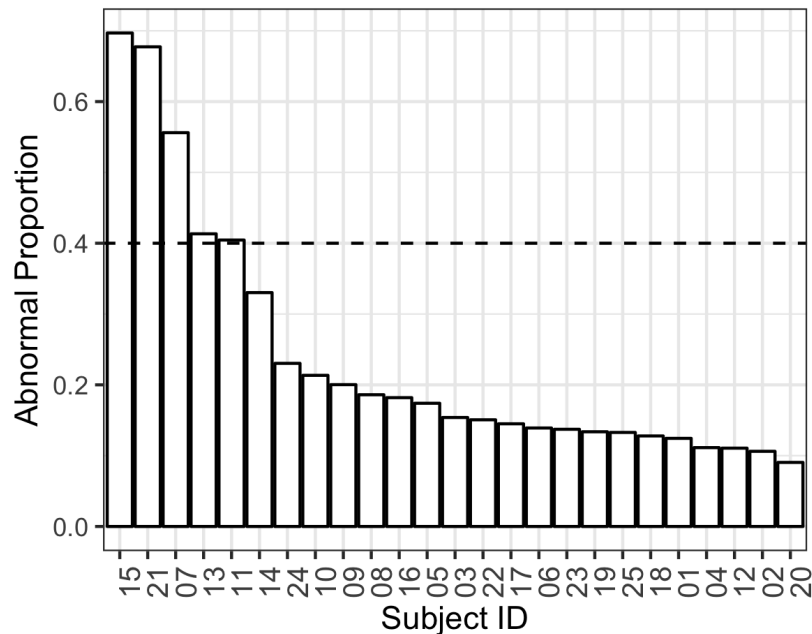

Fig. 13: The bar-plot of abnormal data proportions over all time points (0-270) in HVC study data. Bar-plot of availability (analog to Fig. 9) is not shown since no subjects had more than 40% missing data over this 270 hour time period.

to the pre-infection segmentation in which we used 3 hour periods in the 60 hours (2 nights) of available data, here we could take advantage of the higher specificity of one hour periods since 11 nights are available over the 270 hours. The sleep/wake separability indices (SI), defined in (6), of each of the 12 temporally localized features were computed over all time points in order to evaluate their discrimination power. Figure 14 shows box-plots of the SWSI values for the 12 local features.

Four local features had SWSI greater than 0.7 for over over 75% of the subjects: HR MEAN, HR MED, HR SD, ACC SD. Since HR MEAN and HR MED are highly correlated, HR MEAN was excluded from the top 4 SI features as its 25% quantile is lower than that of HR MED. The remaining features were then used as feature variables in the second stage of the pipeline to identify sleep/wake sessions for each subject. Like in the 0-60 hour segmentation, we again note that the EDA features have lower sleep/wake separability indices than the other E4 variables.