Data Mining and Knowledge Discovery (2021) 35:1760–1784
https://doi.org/10.1007/s10618-020-00724-6

Feature extraction from unequal length heterogeneous EHR time series via dynamic time warping and tensor decomposition

Chi Zhang[1] · **Hadi Fanaee-T**[2] · **Magne Thoresen**[1]

Received: 6 January 2020 / Accepted: 9 November 2020 / Published online: 4 January 2021

Abstract
Electronic Health Records (EHR) data is routinely generated patient data that can provide useful information for analytical tasks such as disease detection and clinical event prediction. However, temporal EHR data such as physiological vital signs and lab test results are particularly challenging. Temporal EHR features typically have different sampling frequencies; such examples include heart rate (measured almost continuously) and blood test results (a few times during a patient's entire stay). Different patients also have different length of stays. Existing approaches for temporal EHR sequence extraction either ignore the temporal pattern within features, or use a predefined window to select a section of the sequences without taking into account all the information. We propose a novel approach to tackle the issue of irregularly sampled, unequal length EHR time series using dynamic time warping and tensor decomposition. We use DTW to learn the pairwise distances for each temporal feature among the patient cohort and stack the distance matrices into a tensor. We then decompose the tensor to learn the latent structure, which is consequently used for patient representation. Finally, we use the patient representation for in-hospital mortality prediction. We illustrate our method on two cohorts from the MIMIC-III database: the sepsis and the acute kidney failure cohorts. We show that our method produces outstanding classification performance in terms of AUROC, AUPRC and accuracy compared with the baseline methods: LSTM and DTW-KNN. In the end we provide a detailed analysis on the feature importance for the interpretability of our method.

Keywords Electronic health records · Dynamic time warping · Tensor decomposition · Patient similarity

Responsible editor: Panagiotis Papapetrou.

Extended author information available on the last page of the article

1 Introduction

Electronic health record, or EHR data is patient data routinely generated from health institutions, including demographics, diagnoses, vital measurements, clinical notes, laboratory test results and medical images. EHR data can provide valuable information for analytical tasks including but not limited to disease detection and classification, medical concept embedding and data augmentation (Xiao et al. 2018). However, EHR data can be challenging due to multi-modality of data types, lack of outcome labels and missingness, temporality and irregularity (Ghassemi et al. 2018; Kruse et al. 2016). Temporal EHR sequences are difficult to model due to two sources of variability in length of sequence: feature-wise and subject-wise. Different features (also known as variables or parameters) can vary greatly in terms of measurement frequency, from measured nearly continuously (blood pressure) to daily or whenever necessary (laboratory test). Different patients (or subjects) can have varying periods of stays in the hospital or intensive care units. It is crucial to represent the data in meaningful ways to proceed with further analytical tasks such as clinical event predictions, hence the heterogeneity in sequence length poses challenges.

There have been several approaches to represent temporal EHR data. One simple approach is to compute sample statistics (minimum, maximum, mean, standard deviation, number of measurements, first measurement) for features at predefined intervals, for instance the first 48 h of the hospital stay (Harutyunyan et al. 2018; Johnson et al. 2017; Guo et al. 2020a). Classification tasks are then completed using classifiers such as logistic regression or gradient boosting machines. This approach produces human readable and interpretable features and can adapt to both feature and subject sequence length variability, however loses the temporal dependency which is valuable for modeling pathophysiologic evolution and disease progression (Luo et al. 2016; Alaa and van der Schaar 2018). Recent developments in deep learning, especially Recurrent Neural Networks (RNN) are able to capture the temporal pattern in multiple features. Long Short Term Memory (LSTM) networks are a type of state-of-the-art RNN, and have shown many successful applications in temporal healthcare data representation and classification. Such examples include sepsis prediction (Scherpf et al. 2019), unplanned intensive care unit readmission (Lin et al. 2019), mortality risk monitoring (Kaji et al. 2019; Purushotham et al. 2018) and other clinical event detection and diagnosis (Lipton et al. 2016). Bidirectional LSTM (BiLSTM) is a variation of LSTM which takes both forward and backward sequence dependency into account, and has been successful in disease inference and predictions (Yu et al. 2020; Guo et al. 2020b).

LSTMs are typically trained on a specified window (first 24 or 48 h of patient records) therefore ignore the irregular sequence lengths (Suresh et al. 2018; Purushotham et al. 2018; Song et al. 2018; Lei et al. 2018). When there are missing values or variables, it is necessary to impute: either with mean, zero or with more complex approaches such as Gaussian Processes (Lipton et al. 2016; Moor et al. 2019). Despite of their outstanding performance in classification tasks, most deep learning methods require a large amount of data to train and remain complex with tens of thousands of hyperparameters that are hard to interpret (Lipton 2016). Recent works to improve interpretability in deep learning using 'attention' mechanisms still require complex architecture (Alaa and van der Schaar 2018; Song et al. 2018). We need therefore

a more transparent tool that can account for temporal sequences sampled at various frequencies for irregular length of periods from different patients.

Modeling unequal length temporal EHR features directly in the raw form requires either extracting the summary statistics at a snapshot or segmenting the sequences into a regular window, as outlined above. On the other hand, if we model the relations between sequences such as similarity instead of the raw sequence itself, the segmentation could be avoided. It is natural to study the similarity or distance (we use these two terms interchangeably) as patients with similar conditions might display similar patterns of physiological trends (Luo et al. 2016). This forms clusters of patients that can be used for personalized predictions and treatments (Che et al. 2017; Ruffini et al. 2017) and to help understand the underlying patient characteristics, also known as phenotyping (Ho et al. 2014a, b; Perros et al. 2017). Powerful data mining tools such as dynamic time warping (Keogh and Pazzani 1999) can align and compare two time series of unequal lengths, and has proven effective in EHR temporal sequence learning (Che et al. 2017; Moor et al. 2019). The distances computed for different features then need to be integrated in some way for further classification tasks. Luo et al. (2016) used frequent subgraph mining to group patients with similar temporal trends, then used subgraph groups to predict 30-day mortality. Moor et al. (2019) proposed to use a hybrid of dynamic time warping (or DTW in short) and the K-nearest neighbor ensemble algorithm to classify each feature, then ensemble the predictive score together to predict sepsis onset. Outside healthcare related applications, nearest neighbor type classifiers with some distance metric remains one of the most powerful time series classification methods (Tan et al. 2019; Bagnall et al. 2016).

Instead of classifying each feature individually and then integrate, an alternative to collect all features together is to put the DTW distance matrices into a multidimensional array: a tensor. In this way data from more than two dimensions can be captured conveniently. This tensor contains information about all features that were originally irregular at feature level and subject level, and its decomposition can provide useful insights on the characteristics of the features and the cohort. We therefore propose a novel method to represent irregular length temporal EHR data via dynamic time warping and tensor decomposition. Instead of using a fixed window of data, we use the full patient sequences from various features that typically differ for each patient. We learn the patient-pairwise feature distance for each feature using dynamic time warping. Based on these distance matrices we construct a third order tensor, then decompose the tensor using CANDECOMP/PARAFAC decomposition (Kiers 2000). Our approach is referred to as DTW-CP. The learned latent feature matrix contains information that can further produce patient representation for supervised learning tasks.

We test DTW-CP on two different cohorts from the open MIMIC-III critical care database with an in-hospital mortality prediction experiment, and compare with baseline results produced by LSTM. With sufficient number of latent components, DTW-CP has consistently better classification performance on both cohorts in three metrics. We provide a detailed analysis of the features and learned latent components to provide insight on which features contain more information for the classification performance.

The rest of the paper is organized as follow. Section 2 provides background information for dynamic time warping and CP decomposition, and describes our methodology of patient representation in detail. Section 3 outlines the experimental evaluation and implementation details. Section 4 provides results for the experiments and analysis of feature importance. Section 5 discusses the strength, limitation and future works and conclude the paper.

2 Methodology

We give a brief review of dynamic time warping and tensor decomposition in Sect. 2.1, then describe our method for patient time series representation in Sect. 2.2.

2.1 Background

We first introduce the notations used in the paper (consistent with Kolda and Bader 2009). A tensor is a multidimensional array, the number of dimensions is called order, modes or ways. In this work we focus on third order tensors. A slice is a two dimensional section of a tensor with two fixed modes. For example $\boldsymbol{X}_{1::}$ is a horizontal slice, which is the first layer or top matrix of a tensor (Table 1).

2.1.1 Tensor decomposition

Tensor decomposition has wide applications in signal processing and data mining (Sidiropoulos et al. 2017; Acar et al. 2017), and has been applied successfully in helathcare informatics (Ho et al. 2014a; Henderson et al. 2017, 2018). In this paper we focus on CANDECOMP/PARAFAC or CP decomposition for short. For a third order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, a CP decomposition for a chosen number of components

Table 1 List of notations

Symbol	Definition
$\boldsymbol{X}, \boldsymbol{D}, \boldsymbol{M}$	Matrix
$\boldsymbol{X}^T$	Matrix transpose
$\boldsymbol{x}_r$	r-th column of $\mathbf{X}$
$\mathcal{X}, \mathcal{D}$	Tensor
x_{ij}, x_{ijk}	Elements of a matrix and a tensor
$\boldsymbol{X}_{i::}, \boldsymbol{X}_{:j:}$	Horizontal, lateral slice of tensor
$\boldsymbol{X}_{::k}$ or simply $\boldsymbol{X}_k$	Frontal slice of tensor
$\boldsymbol{x}, \boldsymbol{y}$	Vector
$\circ$	Outer product

$r = 1, \ldots, R$ can be formalized in the following way:

$$\min_{\hat{\mathcal{X}}} ||\mathcal{X} - \hat{\mathcal{X}}|| \quad \text{where} \quad \hat{\mathcal{X}} = \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \tag{1}$$

Here $\boldsymbol{a}_r, \boldsymbol{b}_r, \boldsymbol{c}_r$ are column vectors of size I, J, K. The vectors can be reorganised into factor matrices $[\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]$ where $\boldsymbol{A} \in \mathbb{R}^{I \times R}$, $\boldsymbol{B} \in \mathbb{R}^{J \times R}$, $\boldsymbol{C} \in \mathbb{R}^{K \times R}$, $\boldsymbol{A} = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \ldots \quad \boldsymbol{a}_R]$. If the columns of $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are normalized to unit length, then the weights are absorbed into $\boldsymbol{\lambda} \in \mathbb{R}^R$,

$$\hat{\mathcal{X}} = \sum_{r=1}^{R} \lambda_r \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r. \tag{2}$$

More details on tensors and the CP decomposition can be seen in (Rabanser et al. 2017; Kolda and Bader 2009) and references therein.

2.1.2 Dynamic time warping

Dynamic time warping (DTW) is a technique to find the optimal alignment between two time dependent sequences, specifically with time deformation and different speed (Keogh and Pazzani 1999; Muller 2007). Given two time series $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ and $\boldsymbol{y} = (y_1, y_2, \ldots y_M)$, construct a cost matrix $\boldsymbol{C} \in \mathbb{R}^{N \times M}$ with elements $c_{n,m} = d(x_n, y_m)$. Here d is a distance measure. With squared Euclidean distance, $d(x_n, y_m) = (x_n - y_m)^2$.

A warping path $W = (w_1, \ldots, w_Q)$ is a set of matrix indices that defines a mapping between $\boldsymbol{x}$ and $\boldsymbol{y}$ where Q is the length of the warping path. Let $w_1 = (1, 1)$, $w_Q = (N, M)$, indicating that the warping path starts and ends in the opposite corner cells of the matrix (boundary conditions). W also need to satisfy additional continuity and monotonicity conditions (Keogh and Pazzani 1999). Let the total cost of a warping path W between $\boldsymbol{x}, \boldsymbol{y}$ be

$$TC_W(\boldsymbol{x}, \boldsymbol{y}) = \sum_{q=1}^{Q} c_{w_q}, \tag{3}$$

The optimal warping path W^* is the one that minimizes the total cost among all possible paths, and the DTW distance is the total cost associated with W^*,

$$\begin{aligned} DTW(\boldsymbol{x}, \boldsymbol{y}) &= TC_{W*}(\boldsymbol{x}, \boldsymbol{y}) \\ &= min\{TC_W(\boldsymbol{x}, \boldsymbol{y})\}. \end{aligned}$$

It is time consuming to find the optimal warping path. By restricting the difference between possible alignment indices between time series pairs, the search window is narrowed around the diagonal of the warping cost matrix. Two well known global

constraints are the Sakoe–Chiba band (Sakoe and Chiba 1978) and Itakura parallelogram (Itakura 1975). A comparison between these two constraints has been made by Geler et al. (2019). More recent works have investigated learning constraints from the data for faster computation and better accuracy (Ratanamahatana and Keogh 2004; Niennattrakul and Ratanamahatana 2009; Salvador and Chan 2007; Dau et al. 2017). It is worth mentioning that constraints work well when the time series lengths do not differ much, otherwise the warping path might not exist (Giorgino 2009).

2.2 Representation of EHR time series

In this section we describe the workflow of representing patient time series of unequal length and sampling frequency. Each unique variable of such physiological time series such as temperature or white blood cell count is referred to as a *feature*. We use the term *distance* and *similarity* interchangeably. Denote the patient index $i, i = 1, \ldots N$ and feature index $k, k = 1, \ldots K$. The length of stay for different patients varies, leading to patient-specific time index denoted by $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{iT})$. The temporal sequence of feature k associated to patient i is recorded as

$$\boldsymbol{p}_{ik} = (p_{ik,t_{i1}}, \ldots, p_{ik,t_{iT}}). \tag{4}$$

2.2.1 Learning latent feature structure

Due to the irregularity in lengths of feature sequences across patients and features, we transform the problem from modeling the individual feature itself for all patients to modeling the similarity of feature between pairs of patients. As dynamic time warping (DTW) can align and compute the distances between pairs of univariate sequences with varying lengths, for each feature k, we compute the distance between each pair of patients (i, j) denoted by

$$d_{ijk} = DTW(\boldsymbol{p}_{ik}, \boldsymbol{p}_{jk}). \tag{5}$$

The procedure is illustrated in Fig. 1a. This forms a third order pairwise distance tensor $\mathcal{D} \in \mathbb{R}^{N \times N \times K}$ where the three modes correspond to patient, patient, feature respectively (Fig. 1b). Each frontal slice $\boldsymbol{D}_{::k}$ represents the pairwise distance matrix for feature k. Elements in the same slice $\boldsymbol{D}_{::k}$ have 0 as diagonal elements, $d_{iik} = 0$ for $i = 1, \ldots N$.

We then proceed by decomposing the tensor $\mathcal{D}$ via CP decomposition with chosen number of components R (Fig. 1c). The motivation for this step is twofold. On the one hand, using CP allows us to learn the latent variables from a complex set of data of multiple unequal-length time series across different patients; on the other hand we reduce the dimensionality of the data and make it possible to carry out predictive tasks. CP produces three factor matrices $M_1 \in \mathbb{R}^{N \times R}$, $M_2 \in \mathbb{R}^{N \times R}$, $M_3 \in \mathbb{R}^{K \times R}$ that are the combinations of the vectors from the rank-one components. They represent patient, patient and feature modes. We refer to M_3 as the *feature factor* matrix where each element $M_{k,r}$ is the loading or weight for feature k on component r.

2.2.2 Patient representation for prediction

We further examine the distance tensor $\mathcal{D}$ from another perspective: its lateral slices $\boldsymbol{D}_{:i:}$. Our approach is similarity based, hence it is necessary to have a common key or *pivot* patient to compare with. A pivot patient is defined as the patient I in the cohort whose features of other patients $i = 1, \ldots N, i \neq I$ are compared to. For instance, the first slice on the left $\boldsymbol{D}_{:1:} \in \mathbb{R}^{N \times K}$ contains DTW distances for all features comparing patient $I = 1$ with all other patients (Fig. 1d). Such a matrix is referred to as a *pivot distance matrix*. Each pivot distance matrix is partial as it only contains distances compared with one key patient. In order to complete a predictive task such as mortality classification, directly using the distance matrix as input creates problems because there is no rule as for which lateral slice (i.e. which pivot patient) to choose. Each component of the feature factor matrix M_3, however, contains feature information (loadings) collected from all patient pairs that can be used for prediction. We produce patient representation $P_I \in \mathbb{R}^{N \times R}$ by projecting the pivot distance matrix onto the feature factor matrix (Fanaee-T et al. 2013) as shown in Fig. 1e,

$$P_I = \boldsymbol{D}_{:I:} M_3. \tag{6}$$

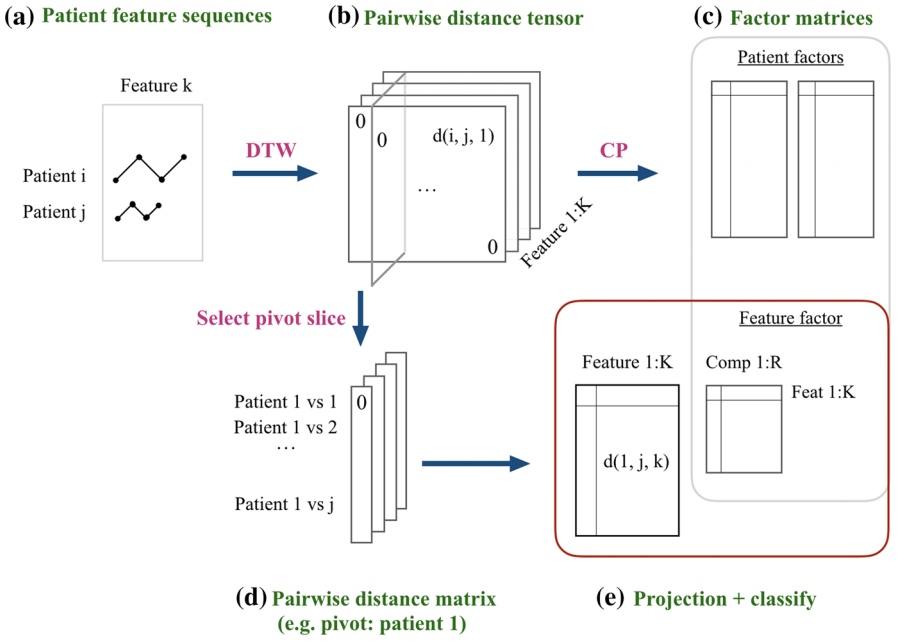

Fig. 1 **a**, **b**, **c**: Procedure of learning latent feature factor, where $d(i, j, 1)$ is the DTW distance between patients (i, j) for feature 1. **d**, **e**: Learning patient representation for prediction. Similarly, $d(1, j, k)$ is the DTW distance between patients $(1, j)$ for feature k

2.2.3 Training and testing procedure

Now we describe the workflow of the training and testing. First the data set is randomly split into training and test sets with 70/30 proportion and class stratification. In the training set, DTW distances are computed for all features. The distance tensor is constructed and then decomposed, producing the feature factor matrix for a pre-chosen number of latent components. To create the projection matrix for classification, we choose an arbitrary lateral slice (of pivot patient I) from the training tensor, and project it onto the feature factor matrix. In the illustration of Fig. 1, $I = 1$ but it can be different. This projection is used for training the classifier. For the test set, it is necessary to compute the DTW distance for feature sequences between the test subjects and the pivot patient I, then make the projection onto the feature factor matrix. This is because we need to make the distance representation consistent: both training and test distances for projection need to be compared with the same pivot.

3 Experimental evaluation

We carry out experiments using a publicly available database, the Medical Information Mart for Intensive Care (MIMIC III) database (Johnson et al. 2016). This is a single center database that contains information about patients admitted to critical care units at Beth Israel Deaconess Medical Center, Boston, USA. The data types include, but are not limited to structured data such as temporal physiological signs and laboratory test results, static demographic information such as age and gender, as well as unstructured data such as free text clinical notes. In the current work we will focus on the structured temporal data. Recent works on reproducible studies using MIMIC-III data make it possible to extract consistent patient cohorts and features. We select two cohorts for our experiments, and our selection criteria is in line with (Johnson et al. 2017).

3.1 Cohort and feature selection

3.1.1 Sepsis cohort

The first cohort we examine is a subset of the sepsis cohort originally studied in (Ribas Ripoll et al. 2014) then reproduced by Johnson et al. (2017). We choose patients who have a sepsis diagnosis (ICD-9 code 995.92 or 785.52) and Simplified Acute Physiology Scores (SAPS) (Le Gall et al. 1993). We only keep patients who have been in the ICU for no more than seven days (168 h), making a cohort of 1425 ICU stays in total. Of these patients, 38.9% are associated with a mortality outcome. Our study period is longer than other works using DTW or similarity-based methods that used only 12 or up to 48 h (Luo et al. 2016; Moor et al. 2019). It is of interest to see whether DTW still works well for longer sequences. We design an incremental inclusion criterion: group 1 contains all subjects with below 24 h (1 day) records, group 2 contains subjects with below 48 h (2 days) records and so on, until 7 days. This suggests that patients within groups with shorter stays are also included in those

Table 2 Information for the sepsis and the acute kidney injury (AKI) cohorts

Cohort index	Length of stay (h)	N patient (Case, Control)
Sepsis 1	[5, 24]	225 (136, 89)
Sepsis 2	[5, 48]	546 (240, 306)
Sepsis 3	[5, 72]	833 (338, 495)
Sepsis 4	[5, 96]	1048 (410, 638)
Sepsis 5	[5, 120]	1202 (468, 734)
Sepsis 5	[5, 144]	1329 (521, 808)
Sepsis 5	[5, 168]	1425 (554, 871)
AKI 1	[5, 24]	652 (189, 463)
AKI 2	[5, 48]	1676 (370, 1306)
AKI 3	[5, 72]	2448 (515, 1933)
AKI 4	[5, 96]	2959 (611, 2348)
AKI 5	[5, 120]	3284 (683, 2601)
AKI 6	[5, 144]	3521 (757, 2764)
AKI 7	[5, 168]	3705 (801, 2904)

Note that for the AKI experiment we use 50 fixed size of stratified random samples (500 subjects in total with 150 cases, 350 controls) for each subgroup 1–7

with longer stays. In this way we can observe DTW's performance on data with smaller and larger sequence length variability.

3.1.2 Acute kidney injury cohort

The second cohort is the acute kidney injury (referred to as AKI in the rest of the paper) cohort based on Johnson et al. (2017). We select patients who have ICD-9 diagnosis of acute kidney injury (code 584.9) who have no more than seven days stay, similar to the previous section. We end up with a cohort of 3705 patients (21.6% mortality). Similar to the previous cohort, we segment the cohort into seven incremental groups: below 24, 48, 72, 96, 120, 144, 168 h corresponding to 1 to 7 days. We modify the experiment slightly to assess the stability of our method in a more controlled scenario. We fix two aspects of the cohorts: sample size and class distribution. We perform experiments on 50 *random samples* of fixed size 500 subjects from the five subgroups corresponding to length of stay, shown in Table 2. The class distribution within each sample is set to 30% case (dead) and 70% control (alive). This produced 350 random samples in total.

3.1.3 Feature selection

In the temporal EHR prediction literature there have been some frequently used physiological and laboratory test variables (Johnson et al. 2017; Moor et al. 2019; Luo et al. 2016; Suresh et al. 2018). The majority of these features are the same, such as heart rate, oxygen saturation and others. Nonetheless, there are some study-specific

Table 3 Extracted features and abbreviations for our experiment

Heart rate	HR	Mean blood pressure	MBP
Systolic blood pressure	SBP	Diastolic blood pressure	DBP
Respiratory rate	RR	Temperature	Temp
Oxygen saturation	SpO2	Glasgow coma scale total	GCS
GCS motor, verbal, eyes	GCS_m, v, e	Urine output	UO
Endotracheal flag	EndoFlag		
Anion gap	AG	Albumin	ALB
Immature Band forms	Band	Base excess	BE_bg
Bilirubin	BIL	Blood Urea Nitrogen	BUN
Bicarbonate	HCO3*	Carboxyhemoglobin	CoHB_bg
Calcium	Ca_bg	Chloride	CL*
Creatinine	CR	Glucose	Glu*
Glucose chart	Glu_c	Hematocrit	HCT*
Hemoglobin	HGB*	Lactate	LAC*
Methemoglobin	MetHb_bg	International Normalized Ratio	INR
Partial pressure (Oxygen)	PO2_bg	Partial pressure (CO2)	PCO2_bg
pH	pH	Platelets	PLT
Prothrombin time	PT	Partial thromboplastin time	PTT
Potassium	K*	Sodium	Na*
Total CO2 concentration	totalCO2_bg	White blood cell count	WBC

The top panel consists of vital signs as well as urine output, Glasgow coma scales and endotracheal flag. The bottom panel contains laboratory test variables. *bg*: arterial blood gas measurement
* Indicate that this feature has more than one measurement source, the other being blood gas

features included in each paper, for instance, Luo et al. (2016) uses volumes of gas exchanged per minute which is not included in other studies. For consistency, we extract a reproducible set of features from Johnson et al. (2017), listed in Table 3. Repeated feature names such as glucose is due to multiple sources of data produced in different test procedures, as explained by the authors (finger-stick glucose or arterial blood gas glucose). The final number of features is 52.

3.2 Implementation details

For features other than lab test variables, we use the period starting from patient admission into ICU until their discharge (from '0h' to end of stay illustrated in Fig. 2). For lab test features, we include an extended period of 24 h before admission (from '-24 h' to the end of stay). These features are typically measured less frequently, and an additional period may contain useful information (Johnson et al. 2017). Each feature is standardized by substracting the mean and dividing by the standard deviation of its own cohort.

We evaluate our DTW-CP method on a binary classification task: in-hospital mortality. DTW is carried out on 52 standardized features for the training set, producing

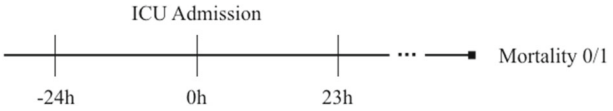

Fig. 2 Time intervals for feature extraction from an individual patient's ICU stay

52 pairwise distance matrices. We use three options for the warping path of DTW: without any constraint, Itakura parallelogram, and Sakoe–Chiba constraint with bandwidth that is half of the maximum length of the two series of interest. In the case when a constrained warping path does not exist, we use the unconstrained warping path to compute the distance. The distance matrices are then stacked into a third order tensor as described in Fig. 1 for CP decomposition. The selected number of components to decompose into is from 2 to 30. The rest of the procedure is as described in Sect. 2.2. We use logistic regression, support vector machine with linear and radial basis function kernel as our classifiers. The tuning parameters for SVM are chosen via 5-fold cross validation. We use three options of pivot patient in our experiments: (1) a random pivot such as the first patient (Sect. 4.1.1, part 1); (2) all patients as pivots (Sect. 4.1.1, part 2); (3) we choose 10 random pivots, split the training set into training and validation data and fit the models with each pivot. The one that has the best validation AUC is picked as the final pivot (Sect. 4.1.2). The metrics to evaluate the classification performance on the test data are Area under Receiver Operating Characteristic curve (AUROC, or AUC in the rest of the paper), Area Under Precision Recall curve (AUPRC) and accuracy (defined by the proportion of correct classifications). The use of AUPRC is to provide a better metric when the class distribution is imbalanced. We report the average of the above three metrics over the random splits from the test sets from each experiment.

We consider two types of comparison methods: K-nearest neighbor combined with dynamic time warping (DTW-KNN), and Long Short Term Memory (LSTM) neural networks. For DTW-KNN, we use the DTW distance computed in the previous task. For all features, we sum up the pairwise DTW distances matching the patient index: the resulting matrix is the multivariate DTW distance matrix with elements $dm_{i,j} = \sum_{k=1}^{52} d_{i,j,k}$ for patients i, j. This is equivalent to the independent multivariate DTW distance (Shokoohi-Yekta et al. 2017). We experiment KNN classifiers with $k = 1, 3, 5$.

There are numerous variations of LSTM architectures (Harutyunyan et al. 2018; Song et al. 2018). A typical LSTM application of temporal EHR data requires each patient record to have at least 24 h of records, then only take the first 24 h for modeling, indicated as the interval between '0h' to '23h' in Fig. 2. While producing good classification results with huge amount of training data, this inclusion criterion ignores patients with shorter records. We adjust this approach to make patient inclusion more flexible. For cohorts with shorter than 24 h records (day1), we make predictions on data periods of both 12 and 18 h for subjects who have at least 12 and 18 h records, respectively. For cohorts with longer records we use 12, 18, 24 h. We use the average performance over these windows as our final metric for that cohort.

We fit LSTM type models of three different architectures for the hidden layers: (1) one LSTM hidden layer; (2) two LSTM hidden layers and (3) one bidirectional LSTM hidden layer (BiLSTM). We use the rectified linear unit (ReLu) activation for hidden layers, and the sigmoid activation for the dense output layer to complete the binary classification. We test two different numbers of units, 64 and 128, for the LSTM layers. We use RMSProp as our optimizer. The batch size is fixed at 32. We train each model with 20 epochs and we use an early stopping if the validation loss stops decreasing for 5 epochs. It is uncommon to have more than two LSTM layers in practice as the number of parameters to estimate explodes. An optimal set of hyperparameters for LSTM does not exist in the literature and the impact of number of units or architecture can be insignificant (Reimers and Gurevych 2017). Our choice of configuration should be representative for this type of methods. We compute the average AUC, AUPRC and accuracy over the random splits from the test set for each window.

Software for implementation: `R` (version 3.6.1) has been used for data preparation, DTW (with `dtw` package created by Giorgino 2009) and classification. `MATLAB Tensorlab` (Vervliet et al. 2016) has been used for CP decomposition. `Keras` (Chollet 2015) with TensorFlow backend has been used for LSTM models.

4 Results

In this section, we report the classification performance tested on both the sepsis and the acute kidney injury (AKI) cohorts, followed by the analysis of features using data from the sepsis cohort. We answer the following questions: (1) how does our method perform in data sets with different combinations of feature sequence heterogeneity compared to the baseline methods? (2) are we able to identify features that are important for the patient representation and the classification performance?

4.1 Classification performance

4.1.1 DTW-CP performance analysis

Figure 3 compares the classification performance (measured by AUC) using DTW distances computed with three warping path options (unconstrained, Itakura parallelogram, and Sakoe–Chiba band) as features, and logistic regression(LR) and linear SVM as classifiers for the sepsis cohort for each group (sepsis 1–7, Table 2) over 10 random splits of the training and test sets. The pivot is fixed at the first patient. We use 'group' and 'day' interchangeably in the rest of the paper. On the x-axis is the number of components or latent features as predictors. Overall, different constraints do not give very different classification performance. Computation times for the constraints can be found in Sect. 4.3. When the number of components grows, the AUC increases and stabilizes for all groups. This upward-then-stable pattern is common for tensor-based phenotyping and prediction applications (Ho et al. 2014a, b). The exception is day1 (dark green) with the LR classifier: after component 20 to 25 the performance start to deteriorate yet still stays above 0.8. This was not the case for

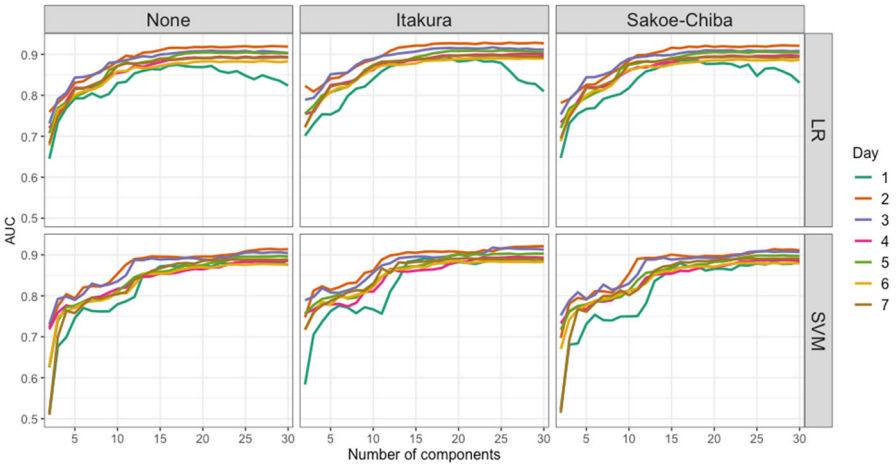

Fig. 3 AUC of DTW-CP with LR and linear SVM classifier on 1–7 groups for the sepsis cohort, fixed pivot at the first patient. DTW distances are computed with no constraint ('None'), Itakura parallelogram, and Sakoe–Chiba band. Lines represent the mean AUC over 10 randomly split test sets for each number of components from 2 to 30

the SVM classifier. This indicates that the optimal component r for day1 with LR is smaller than 30; alternatively a classifier with penalization could be applied to mitigate the problem.

Similarly, we show the mean AUC for the acute kidney injury (AKI) cohort for group 1–7 over 50 *randomly sampled* test sets under different DTW constraints and classifiers (Fig. 4). Similar to the sepsis cohort results, the AUC displays an upward-then-stable trend as the number of components increases. As the day grows (hence the variation among the sequences within the cohort) the performance slightly deteriorates. The patterns in the AKI cohort is more consistent than the sepsis cohort and less variable.

We then test the performance of DTW-CP over different choices of pivots: we carry out classification tasks using all the pivot patients in the training tensor for each day from the sepsis cohort (with only one random split) with component $R = 30$ and no DTW constraint, and report the mean and standard error of the metrics in Table 4. The results can be compared with Figs. 3 and 6. Apart from day1 where the classification performance is slightly worse and with higher LR standard errors, the other metrics fluctuate with an SE around 0.02. It is not straight-forward to identify the potential outliers in the cohort because there are multiple features, and all distances are relative to which pivot to compare with. Instead of iterating over all possible pivots, one way to choose the pivot is to randomly choose a few (for instance, 10) and pick one that produces the best validation AUC.

4.1.2 Comparison with KNN and LSTM

In this section we compare DTW-CP with KNN and LSTM methods. We make use of the procedure described in the previous section: we randomly choose 10 training pivot patients and take the pivot with best validation AUC as the optimal pivot. The

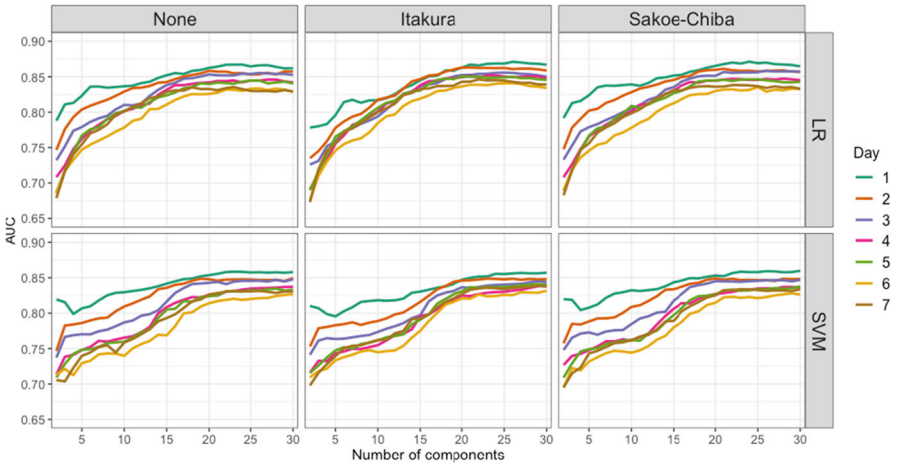

Fig. 4 AUC of DTW-CP with LR and linear SVM classifier on 1–7 groups for the AKI cohort, fixed pivot at the first patient. DTW distances are computed with no constraint ('None'), Itakura parallelogram, and Sakoe–Chiba band. Lines represent the mean AUC over 50 randomly split test sets for each number of components from 2 to 30

Table 4 Performance (mean, SE) on the sepsis cohort with different pivot patients

Metric/Classifier	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
AUC LR							
Mean	0.836	0.876	0.900	0.881	0.852	0.874	0.884
SE	0.044	0.020	0.018	0.021	0.020	0.018	0.022
AUC SVM							
Mean	0.873	0.871	0.889	0.864	0.832	0.863	0.872
SE	0.030	0.023	0.020	0.023	0.024	0.020	0.024
Accuracy LR							
Mean	0.773	0.809	0.850	0.819	0.798	0.830	0.841
SE	0.042	0.021	0.019	0.020	0.019	0.019	0.016
Accuracy SVM							
Mean	0.806	0.805	0.831	0.800	0.788	0.822	0.830
SE	0.040	0.021	0.020	0.020	0.020	0.020	0.019
AUPRC LR							
Mean	0.853	0.883	0.890	0.874	0.801	0.820	0.850
SE	0.050	0.017	0.018	0.022	0.026	0.030	0.024
AUPRC SVM							
Mean	0.897	0.876	0.879	0.848	0.778	0.802	0.827
SE	0.041	0.021	0.022	0.026	0.031	0.029	0.030

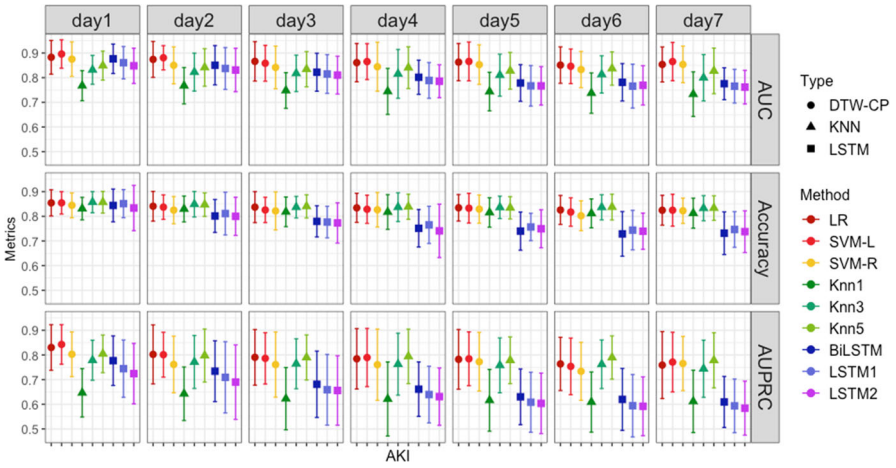

Fig. 5 Three metrics (mean, 95% CI) for the AKI cohort, comparing DTW-CP with LSTM, KNN over 7 groups. DTW-CP (LR = logistic regression, SVM-L = SVM with linear kernel, SVM-R = SVM with radial basis function kernel) performance are extracted at component 30 with unconstrained DTW

results are averaged over 10 random splits from the sepsis cohort, and 50 random splits from the AKI cohort. We focus on the case with 30 components. In Fig. 5 we illustrate the three metrics from the AKI cohort. It can be seen that for DTW-CP, LR and SVM-L classifiers produce similar results, while SVM-R is slightly worse. The impact of adding sequence length variation (from day 1 to day 7) is not as obvious as in Fig. 4, and the AUC is higher after selecting the optimal pivot. Compared with the baseline methods, DTW-CP with LR or SVM-L produce the best AUC. When it comes to the accuracy and AUPRC, DTW-CP is constantly better than LSTM methods, and has better or similar performance as the best KNN method from day 1 to day 5. As the prediction horizon increases, the deterioration of the LSTM methods is more obvious. This is not surprising, as there is not enough information to predict in a long term by using only the first 24 h without huge amounts of training data. BiLSTM seems to have the best performance among the LSTM methods.

In Fig. 6 we show the performance comparison on the sepsis cohort. DTW-CP outperforms LSTMs and KNN (k = 1) in all metrics on all groups except day1. It also produces better or equal performance in all metrics as the best KNN in day 3, 4 and 5. In day 2, 6 and 7, DTW-CP has comparable or marginally lower performance than the best KNN (k = 5) in one of the three metrics. We also observe that the performance of DTW-CP in day 1 is worse than the other groups in terms of AUC and accuracy, although the metric values are still decent. This is consistent with Fig. 3 and Table 4.

We make the following comments on the performance differences between the sepsis and the AKI cohort. With DTW-CP, when we use any random pivot (such as the first patient, Figs. 3, 4), the overall performance in terms of AUC is better in the sepsis than the AKI cohort, with an exception day1. The worse performance in sepsis-day1 compared to other days in the sepsis data is probably due to much fewer samples (only 225, see Table 2); when the sample size is larger (AKI) , day1 has better performance than all other days. In all the other days, sepsis has larger sample size than AKI, which

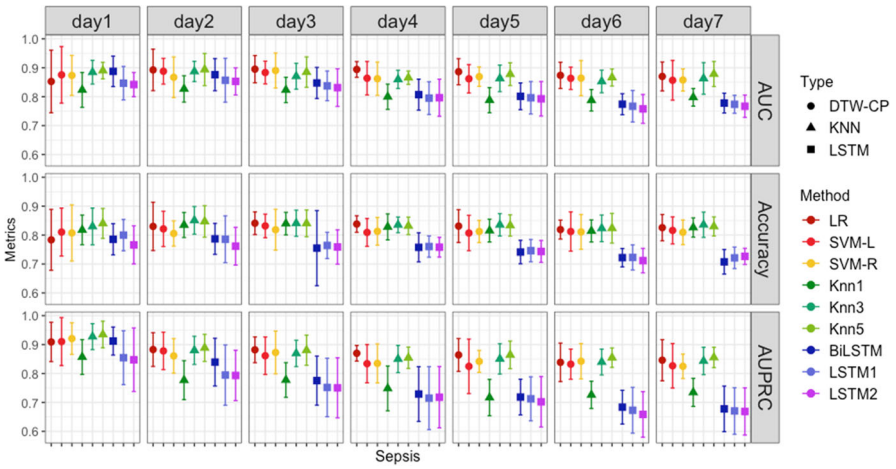

Fig. 6 Three metrics (mean, 95% CI) for the sepsis cohort, comparing DTW-CP with LSTM, KNN over 7 groups. DTW-CP performance are extracted at component 30 with unconstrained DTW

could explain why performance is still competitive when sequence length variation gets bigger. With controlled sample size in all its subgroups, the AKI cohort displays rather constant deterioration as length variation grows (day1, 2 has better AUC than day6, 7). We summarize that DTW-CP could perform better under two conditions: when there is more data, and when the sequences are shorter. If we select the pivot that produces that best validation AUC among a few randomly chosen ones, then the sequence length variation has less impact on the performance.

4.2 Analysis of feature importance

Following the good classification performance, we further investigate the interpretability using data from the sepsis cohort. The aim is to understand which features play an important role in the patient representation. We look at three aspects, namely the measurement frequency of the features, the distance matrix for one pivot patient and the learned latent feature matrix from CP decomposition. The feature names and abbreviations are consistent with Table 3.

4.2.1 Measurement frequency

The features we use vary greatly in terms of measurement frequency, and consequently, in terms of total number of measurements and length of sequence. Figure 7 illustrates the average number of measurements for patients in the sepsis cohort. The time stamp of feature recording is rounded to the nearest hour; if more than one measurement per hour is made, an average is taken. The total number of hours of patient stay in hospital or intensive care unit (length of stay, LOS) is therefore the maximum number of measurements for this patient. The cohort mean (median) length of stay is 56.65 (52) h. Vital features such as heart rate, blood pressures and oxygen saturation are

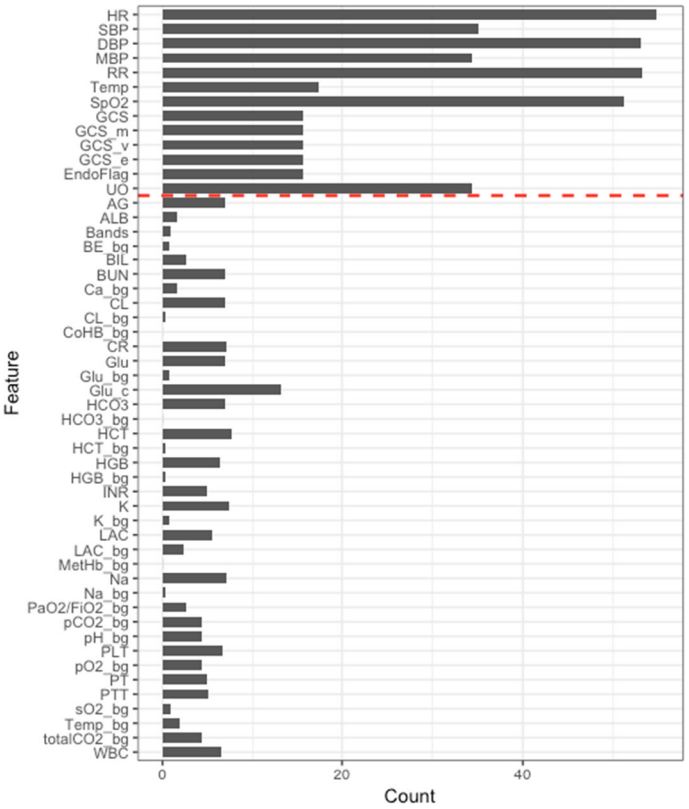

Fig. 7 Average number of total measurements for features in sepsis cohort. Cohort mean (median) length of stay is 56.65 (52) h. The red dashed line distinguishes the non-lab and lab test features

measured almost hourly while laboratory tests are taken only a few times during a patient's entire hospital stay. At the same time, even within the same feature (i.e. heart rate), the number of measurements can vary across patients given different LOS.

4.2.2 Distance matrix

To deal with the heterogeneity of time series outlined in the previous section, we work with the similarity (distance) between patient pairs computed via DTW. Figure 8 presents a heatmap for a pivot distance matrix for an arbitrary patient, as described in Fig. 1. It is important to point out that this matrix varies for different pivot patients. The X-axis represents the subject index of the cohort. Each colored element represents the DTW distance for each individual patient compared to the pivot for the corresponding feature, plotted on the y-axis. The features are ordered in the same way as Fig. 7. The top rows represent very frequently measured features (vitals and procedures) having close to zero distances with low variability, colored in deep red. Most blood gas test results (end with _bg) are measured very infrequently and display the same pattern as the vitals. This effect could be interpreted as follows: frequently measured features are

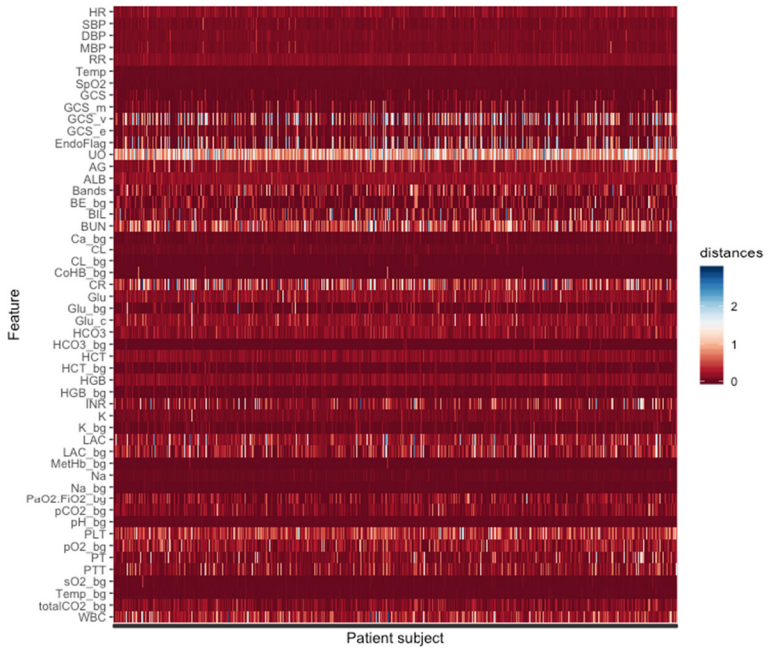

Fig. 8 Distance matrix for one patient from the sepsis cohort

vital signs that are inherently similar, hence little distance; while features with very few measurements simply contain too little information.

Noticeably, for this pivot patient the urine output, GCS measurements, blood urea nitrogen, creatinine, lactate, PaO2/FiO2 ratio, platelet counts and white blood cell count display higher distance variability, colored in white and blue. We assume features with high variance provide more information for classification.

4.2.3 Latent feature matrix

The pivot distance matrix only contains DTW distances of one particular patient compared to others in the cohort, therefore it is patient-specific. Tensor decomposition (CP) provides a useful tool to summarize information from the whole cohort. The latent feature matrix of the sepsis cohort is 52 rows (feature) by $R = 2, \ldots 30$ columns (component). By examining each component, we can identify which feature was important or unimportant by examining the magnitude of its loadings. In contrast to to Principal Component Analysis (PCA), the first component from CP does not necessarily correspond to the direction explaining the largest variance: there is no ordering among the components. We illustrate with an example of three arbitrary components out of 30 from the CP decomposition in Fig. 9, as it is infeasible to visualize more than three dimensions. We normalize loadings of each component to unit length. In this particular factorization, it can be observed that most features have low factor weights or loadings, as they are concentrated around 0, and some are more spread out.

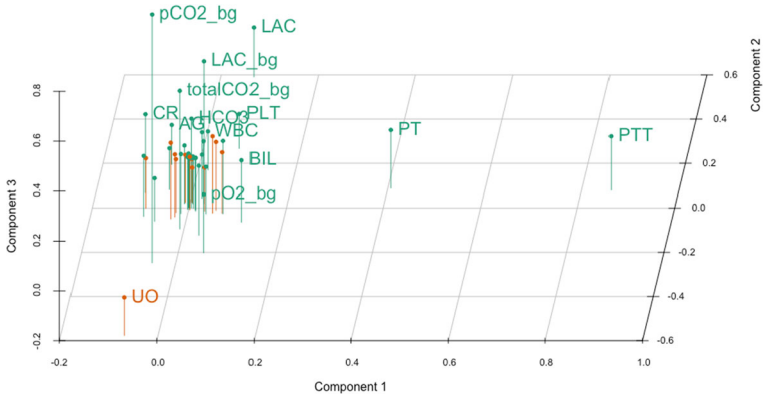

Fig. 9 Normalized factor loadings for three arbitrary components out of 30 from the CP decomposition, corresponding to the sepsis day 1 cohort. Green and Red color indicate feature categories Lab and Vital. Only loadings of magnitude greater than 0.1 in any direction is labeled for readability

To further investigate the importance of each particular feature, we calculate the average factor loading in the following way. For the decomposed feature factor matrix $M \in \mathbb{R}^{52\times R}$ where each row corresponds to the k_{th} feature's R_{th} component loading, and we define the average loading of feature k as the average magnitude of all its R loadings. We illustrate the average loadings for the sepsis tensor decomposition over 7 days with fixed number of components, $R = 30$ in Fig. 10. Comparing with Fig. 7 it can be observed that the factor loading does not correspond with the measurement frequency: temperature and SpO2 are measured rather frequently but have low loadings across all 7 days constantly; creatinine, lactate, PaO2/FiO2 fraction are measured fewer times but have greater loadings. Regarding trend corresponding to one to seven day data, features display various patterns: increasing (GCS verbal), constant (heart rate) and decaying (lactate). This examination also reveals which features play very little role (close to zero loading for all 7 days) in the patient tensor structure.

From the factor loadings we can try to link to the physical meanings of feature importance. Urine output is measured frequently and is a marker for acute kidney injury that is associated with high hospital mortality (Legrand and Payen 2011; Zhang et al. 2014). Lactate (serum and blood gas) has both shown up as important features, and lactate level elevation is associated with increased risk of death (Sanderson et al. 2018; Filho et al. 2016; Trzeciak et al. 2007). The other features such as PaO2/FiO2 ratio (Allardet-Servent et al. 2009), glucose (Park et al. 2013), creatinine, bilirubin, platelet counts, INR (Murali et al. 2014; Li et al. 2018) are indicators for functionality in different organs, and GCS scores (Ting et al. 2010) provides information for the mobility of a patient. Our method could be one step forward to understanding which features are most indicative for classification for similar datasets, in contrast to including all features available and utilizing models with complex architecture. It is crucial to point out that physiological patterns are extremely complex especially for critically ill patients, and all interpretations are data and context dependent. Therefore any use of machine learning models need to be carefully verified by clinicians.

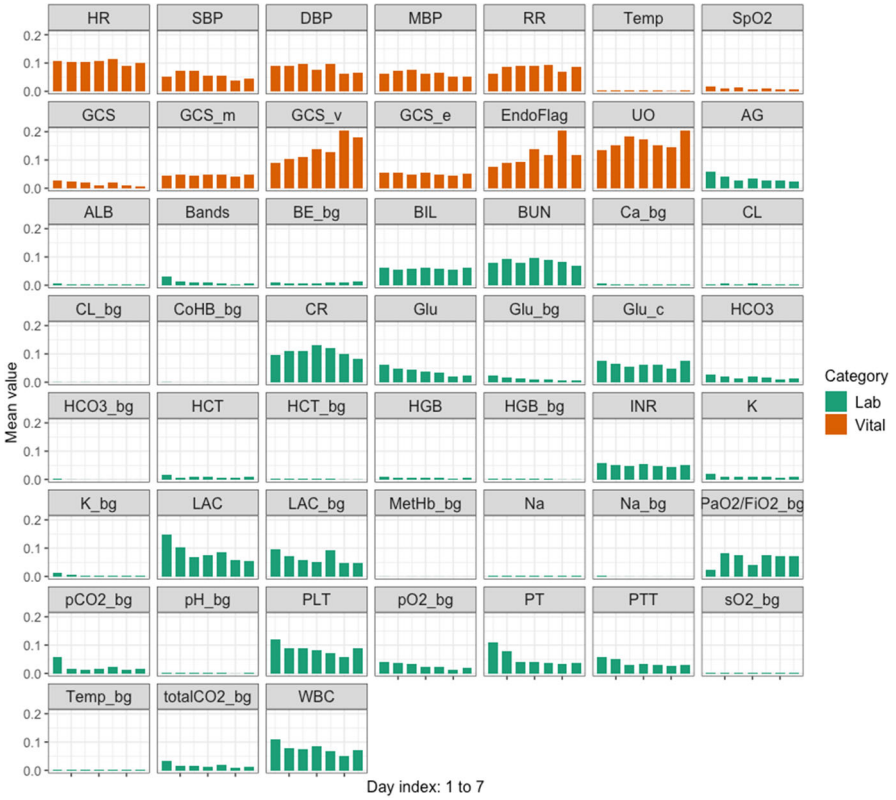

Fig. 10 Average loading for 52 features over all $R = 30$ components for the CP decomposition, sepsis data, day 1 to 7

4.3 Scalability of DTW and CP

We provide the execution time for Dynamic time warping and CP decomposition for the sepsis dataset. Computations are performed on a High Performance Computer running Red Hat Enterprise Linux 7. The hardware includes Intel ®Xeon ®Platinum 8160 (2.10 GHz) CPU and 1TB of RAM.

The average time for DTW computations in hours (mean, standard deviation) for all features is reported in Table 5. Itakura parallelogram and Sakoe–Chiba constraint (of bandwidth half of the maximum sequence length) improve the DTW speed compared to unconstrained DTW. The higher standard deviation in the unconstrained DTW is due to longer time required for features with longer sequences, such as heart rate (Table 6).

We also provide the time required for CP decomposition with varying size of tensors and number of components to decompose. The computation is carried out using MATLAB tensorlab toolbox. We report the execution time in seconds for the sepsis data set, day 1 to 7 subgroups (averaged over 10 random splits) where the dimension of target tensor grows from $158 \times 158 \times 52$ to $998 \times 998 \times 52$.

Table 5 Average DTW execution time (h) for 52 features

Constraint	Mean	SD
None	1.047	0.351
Itakura	0.980	0.178
Sakoe–Chiba	0.904	0.096

Table 6 CP decomposition execution time (seconds) into 10, 20, 30 components

Data index Size of patient mode	Day 1 158	Day 2 383	Day 3 584	Day 4 734	Day 5 842	Day 6 931	Day 7 998
10	13.01	36.37	56.62	80.66	90.49	98.74	114.09
20	18.81	111.99	134.69	156.64	190.05	229.46	323.64
30	143.30	169.91	211.02	275.45	283.10	302.19	356.80

5 Discussions and conclusion

We have proposed a novel approach, a hybrid of dynamic time warping with tensor decomposition (DTW-CP) to tackle a prevalent but challenging issue with temporal EHR sequences: varying sampling frequency among features for different lengths of patient stays. Our approach utilizes DTW to learn information about feature similarities for patients in the cohort, and consequently uses tensor decomposition to learn the latent feature structures. In addition, we have done a detailed analysis of the temporal features used in many clinical prediction applications using the MIMIC-III database. We illustrated that the importance of a feature (i.e. high factor loading from decomposition) is not directly related to how often it is measured, and linked the 'important' features to their clinical interpretations.

Among all the works using DTW or tensor decomposition in healthcare, we are the first to combine these two. Moreover, we have extended the DTW time period to up to seven days, and illustrated how classification performance changes with different variation in sequence length. We carried out careful experiments using (1) distance matrices computed by different DTW constraints (Itakura parallelogram, Sakoe–Chiba band versus unconstrained DTW); (2) different pivot options; (3) different classifiers (logistic regression, linear and radial basis function kernel SVM). By comparing with two baseline methods: LSTM with three architectures, and DTW-KNN methods, we have shown that our method is able to outperform them in three different metrics. We also give interpretations of the classification performance with different data sets and different settings.

DTW-CP is a similarity (distance) based approach, this has two implications. Firstly it is necessary to compute the distance between all pairs of patients in the cohort for each feature. This step can be time consuming when the sequences are long and when the cohort is large, as pointed out in Moor et al. (2019) (who did not use any constraint, but used fastDTW in their implementation). Although DTW computation time can be reduced with constraints, it can only be used when the sequence length do not differ much; also it is unclear which constraint is the best (Geler et al. 2019). Secondly, the

interpretation of features is based on patient similarity instead of the feature value themselves. This means there is always a need for pivot patient to compare the rest of the cohort with, to make the interpretation meaningful. We have chosen to optimize the choice of pivot based on maximizing AUC. This choice should of course be guided by which metric is most important for any given application.

Our choice of decomposition algorithm (CP) does not have non-negative constraint, hence the interpretation of latent feature matrix distinguishes itself from Ho et al. (2014a, b); Afshar et al. (2018) and others where each component is a combination of positive phenotype memberships. There is no standard way to choose the number of components to decompose into, hence we suggest that in practice this should be where the classification performance stabilizes. Lastly, we have only utilized temporal EHR sequences. Most works on patient clustering and clinical event predictions include static demographic data in addition to the dynamic data (Suresh et al. 2018; Purushotham et al. 2018), thus combining static data with temporal sequences is a direction we could investigate further. Possible solutions include coupled matrix and tensor factorization (CMTF).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Acar E, Levin-Schwartz Y, Calhoun VD, Adali T (2017) Tensor-based fusion of EEG and FMRI to understand neurological changes in schizophrenia. In: Proceedings—IEEE international symposium on circuits and systems, pp 1–4. https://doi.org/10.1109/ISCAS.2017.8050303

Afshar A, Perros I, Papalexakis EE, Searles E, Ho J, Sun J (2018) COPA: constrained PARAFAC2 for sparse and large datasets. In: The 27th ACM international conference on information and knowledge management (CIKM '18). https://doi.org/10.1145/3269206.3271775

Alaa AM, van der Schaar M (2018) Forecasting individualized disease trajectories using interpretable deep learning. arXiv:1810.10489

Allardet-Servent J, Forel JM, Roch A, Guervilly C, Chiche L, Castanier M, Embriaco N, Gainnier M, Papazian L (2009) FiO2 and acute respiratory distress syndrome definition during lung protective ventilation. Crit Care Med 37(1):202–207. https://doi.org/10.1097/CCM.0b013e31819261db

Bagnall A, Bostrom A, Large J, Lines J (2016) The great time series classification bake off: an experimental evaluation of recently proposed algorithms. Extended version arXiv:1602.01711

Che C, Xiao C, Liang J, Jin B, Zho J, Wang F (2017) An RNN architecture with dynamic temporal matching for personalized predictions of Parkinson's disease. In: Proceedings of the 2017 SIAM international conference on data mining, pp 198–206. https://doi.org/10.1137/1.9781611974973.23

Chollet F (2015) Keras. https://keras.io

Dau HA, Silva DF, Petitjean F, Forestier G, Bagnall A, Keogh E (2017) Judicious setting of Dynamic Time Warping's window width allows more accurate classification of time series. In: Proceedings—2017 IEEE international conference on big data, big data 2017. https://doi.org/10.1109/BigData.2017.8258009

Fanaee-T H, Oliveira M, Gama J, Malinowski S, Morla R (2013) Event and anomaly detection using tucker3 decomposition. In: Proceedings of 20th European conference on artificial intelligence (ECAI'2013)-ubiquitous data mining workshop, vol 1, pp 8–12. arXiv:1406.3266v1

Filho RR, Rocha LL, Correa TD, Pessoa CMS, Colombo G, Assuncao MSC (2016) Blood lactatte levels cutoff and mortality prediction in sepsis—time for a reappraisal? A retrospective cohort study. Shock 46(5):480–485. https://doi.org/10.1097/SHK.0000000000000667

Geler Z, Kurbalija V, Ivanovic M, Radovanovic M, Dai W (2019) Dynamic time warping: Itakura vs Sakoe–Chiba. In: IEEE international symposium on innovations in intelligent systems and applications, INISTA 2019—Proceedings. https://doi.org/10.1109/INISTA.2019.8778300

Ghassemi M, Naumann T, Schulam P, Beam AL, Ranganath R (2018) Opportunities in machine learning for healthcare. arXiv:1806.00388

Giorgino T (2009) Computing and visualizing dynamic time warping alignments in R: the dtw package. J Stat Softw 31(7):1–24. https://doi.org/10.18637/jss.v031.i07

Guo C, Lu M, Chen J (2020a) An evaluation of time series summary statistics as features for clinical prediction tasks. BMC Med Inform Decis Mak 20(1):1–20. https://doi.org/10.1186/s12911-020-1063-x

Guo D, Duan G, Yu Y, Li Y, Wu FX (2020b) A disease inference method based on symptom extraction and bidirectional Long Short Term Memory networks. Methods 173(April 2019):75–82. https://doi.org/10.1016/j.ymeth.2019.07.009

Harutyunyan H, Khachatrian H, Kale DC, Steeg GV, Galstyan A (2018) Multitask learning and benchmarking with clinical time series data. arXiv:1703.07771

Henderson J, Ho JC, Kho AN, Denny JC, Malin BA, Sun J, Ghosh J (2017) Granite: diversified. Sparse tensor factorization for electronic health record-based phenotyping. In: IEEE international conference on healthcare informatics (ICHI). https://doi.org/10.1109/ICHI.2017.61

Henderson J, Malin BA, Ho JC (2018) PIVETed-granite: computational phenotypes through constrained tensor factorization. arXiv:1808.02602v1

Ho J, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, Sun J (2014a) Limestone: high-throughput candidate phenotype generation via tensor factorization. J Biomed Inform 52:199–211. https://doi.org/10.1016/j.jbi.2014.07.001

Ho J, Ghosh J, Sun J (2014b) Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 115–124. https://doi.org/10.1145/2623330.2623658

Itakura F (1975) Minimum prediction residual principle applied to speech recognition. IEEE Trans Acoust Speech Signal Process 23(1):67–72. https://doi.org/10.1109/TASSP.1975.1162641

Johnson AE, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi AL, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:160035. https://doi.org/10.1038/sdata.2016.35. https://www.nature.com/articles/sdata201635

Johnson AEW, Pollard TJ, Mark RG (2017) Reproducibility in critical care: a mortality prediction case study. In: 2nd machine learning for healthcare conference, PMLR, vol 68. http://proceedings.mlr.press/v68/johnson17a.html

Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, Oermann EK (2019) An attention based deep learning model of clinical events in the intensive care unit. PLoS ONE 14(2):1–17. https://doi.org/10.1371/journal.pone.0211057

Keogh EJ, Pazzani MJ (1999) Scaling up dynamic time warping to massive datasets. Princ Data Min Knowl Discov 1704(Derriere):1–11. https://doi.org/10.1007/978-3-540-48247-5_1

Kiers HAL (2000) Towards a standardized notation and terminology in multiway analysis. J Chemom 14:105–122

Kolda TG, Bader BW (2009) Tensor decompositions and applications. SIAM Rev 51(3):455–500. https://doi.org/10.1137/07070111X

Kruse CS, Goswamy R, Raval Y, Marawi S (2016) Challenges and opportunities of big data in health care: a systematic review. JMIR Med Inform 4(4):e38. https://doi.org/10.2196/medinform.5359

Le Gall JR, Lemeshow S, Saulnier F (1993) Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA 270(24):2957–2963

Legrand M, Payen D (2011) Understanding urine output in critically ill patients. Ann Intensive Care 1(1):13. https://doi.org/10.1186/2110-5820-1-13. http://www.annalsofintensivecare.com/content/1/1/13

Lei L, Zhou Y, Zhai J, Zhang L, Fang Z, He P, Gao J (2018) An effective patient representation learning for time-series prediction tasks based on EHRs. In: Proceedings—2018 IEEE international conference on bioinformatics and biomedicine, BIBM 2018. https://doi.org/10.1109/BIBM.2018.8621542

Li Y, Chaiteerakij R, Kwon JH, Jang JW, Lee HL, Cha S, Ding XW, Thongprayoon C, Ha FS, Nie CY, Zhang Q, Yang Z, Giama NH, Roberts LR, Han T (2018) A model predicting short-term mortality in patients with advanced liver cirrhosis and concomitant infection. Medicine 97(41):e12758

Lin YW, Zhou Y, Faghri F, Shaw M, Campbell R (2019) Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLoS ONE 14(7):e0218942. https://doi.org/10.1371/journal.pone.0218942

Lipton ZC (2016) The mythos of model interpretability. arXiv:1606.03490

Lipton ZC, Kale DC, Elkan C, Wetzel R (2016) Learning to diagnose with LSTM recurrent neural networks. In: 4th international conference on learning representations, ICLR 2016—conference track proceedings, pp 1–18. arXiv:1511.03677

Luo Y, Xin Y, Joshi R, Celi L, Szolovits P (2016) Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In: 30th AAAI conference on artificial intelligence, AAAI 2016, pp 42–50

Moor M, Horn M, Rieck B, Roqueiro D, Borgwardt K (2019) Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping. arXiv:1902.01659

Muller M (2007) Dynamic time warping. In: Information retrieval for music and motion, Springer, Berlin, Heidelberg, chap 4, pp 69–84

Murali AR, Devarbhavi H, Venkatachala PR, Singh R, Sheth KA (2014) Factors that predict 1-month mortality in patients with pregnancy-specific liver disease. Clin Gastroenterol Hepatol 12(1):109–113. https://doi.org/10.1016/j.cgh.2013.06.018

Niennattrakul V, Ratanamahatana CA (2009) Learning DTW global constraint for time series classification. arXiv:0903.0041

Park BS, Yoon JS, Moon JS, Won KC, Lee HW (2013) Predicting mortality of critically ill patients by blood glucose levels. Diabetes Metab J 37:385–390

Perros I, Papalexakis EE, Wang F, Vuduc R, Searles E, Thompson M, Sun J (2017) SPARTan: scalable PARAFAC2 for large and sparse data. In: KDD. https://doi.org/10.1145/3097983.3098014

Purushotham S, Meng C, Che Z, Liu Y (2018) Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 83:112–134. https://doi.org/10.1016/j.jbi.2018.04.007

Rabanser S, Shchur O, Günnemann S (2017) Introduction to tensor decompositions and their applications in machine learning, pp 1–13. arXiv:1711.10781

Ratanamahatana CA, Keogh E (2004) Making time-series classification more accurate using learned constraints. In: SIAM proceedings series, pp 11–22. https://doi.org/10.1137/1.9781611972740.2

Reimers N, Gurevych I (2017) Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. arXiv:1707.06799

Ribas Ripoll VJ, Vellido A, Romero E, Ruiz-Rodríguez JC (2014) Sepsis mortality prediction with the quotient basis kernel. Artif Intell Med 61(1):45–52. https://doi.org/10.1016/j.artmed.2014.03.004

Ruffini M, Gavaldà R, Limón E (2017) Clustering patients with tensor decomposition 68. https://doi.org/10.1002/dei. arXiv:1708.08994

Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process ASSP 26(1):43–49

Salvador S, Chan P (2007) FastDTW: toward accurate dynamic time warping in linear time and space. Intell Data Anal 11(5):561–580. https://doi.org/10.3233/ida-2007-11508

Sanderson M, Chikhani M, Blyth E, Wood S, Moppett IK, Mckeever T, Simmonds MJR (2018) Predicting 30-day mortality in patients with sepsis: an exploratory analysis of process of care and patient characteristics. J Intensive Care Soc 19(4):299–304. https://doi.org/10.1177/1751143718758975

Scherpf M, Gräßer F, Malberg H, Zaunseder S (2019) Predicting sepsis with a recurrent neural network using the MIMIC III database. Comput Biol Med 113(June):103395. https://doi.org/10.1016/j.compbiomed.2019.103395

Shokoohi-Yekta M, Hu B, Jin H, Wang J, Keogh E (2017) Generalizing DTW to the multi-dimensional case requires an adaptive approach. Data Min Knowl Disc 31(1):1–31. https://doi.org/10.1007/s10618-016-0455-0

Sidiropoulos ND, De Lathauwer L, Fu X, Huang K, Papalexakis EE, Faloutsos C (2017) Tensor decomposition for signal processing and machine learning. IEEE Trans Signal Process 65(13):3551–3582. https://doi.org/10.1109/TSP.2017.2690524. arXiv:1607.01668

Song H, Rajan D, Thiagarajan JJ, Spanias A (2018) Attend and diagnose: clinical time series analysis using attention models. In: 32nd AAAI conference on artificial intelligence, AAAI 2018, pp 4091–4098. arXiv:1711.03905

Suresh H, Gong JJ, Guttag J (2018) Learning tasks for multitask learning: heterogenous patient populations in the ICU. In: KDD. https://doi.org/10.1145/3219819.3219930. arXiv:1806.02878

Tan CW, Petitjean F, Webb GI (2019) FastEE: fast ensembles of elastic distances for time series classification. Data Min Knowl Discovy. https://doi.org/10.1007/s10618-019-00663-x

Ting H, Chen M, Hsieh Y, Chan C (2010) Good mortality prediction by Glasgow Coma scale for neurosurgical patients. J Chin Med Assoc 73(3):139–143. https://doi.org/10.1016/S1726-4901(10)70028-9

Trzeciak S, Dellinger RP, Chansky ME, Arnold RC, Schorr C, Milcarek B, Hollenberg SM, Parrillo JE (2007) Serum lactate as a predictor of mortality in patients with infection. Intensive Care Med 33:970–977. https://doi.org/10.1007/s00134-007-0563-9

Vervliet N, Debals O, Sorber L, Van Barel M, De Lathauwer L (2016) Tensorlab 3.0

Xiao C, Choi E, Sun J (2018) Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 25(10):1419–1428. https://doi.org/10.1093/jamia/ocy068

Yu K, Zhang M, Cui T, Hauskrecht M (2020) Monitoring ICU mortality risk with a long short-term memory recurrent neural network. Pac Symp Biocomput 25:103–114. https://doi.org/10.1142/9789811215636_0010

Zhang Z, Xu X, Ni H, Deng H (2014) Urine output on ICU entry is associated with hospital mortality in unselected critically ill patients. J Nephrol 27:65–71. https://doi.org/10.1007/s40620-013-0024-1

Affiliations

Chi Zhang[1] · Hadi Fanaee-T[2] · Magne Thoresen[1]

✉ Chi Zhang
chi.zhang@medisin.uio.no

Hadi Fanaee-T
hadi.fanaee@hh.se

Magne Thoresen
magne.thoresen@medisin.uio.no

[1] Department of Biostatistics, Domus Medica, University of Oslo, Oslo, Norway

[2] Center for Applied Intelligent Systems Research, Halmstad University, Halmstad, Sweden