An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease

Chao Che*† Cao Xiao*‡ Jian Liang§ Bo Jin¶ Jiayu Zhou‖ Fei Wang**

Abstract

Parkinson's disease (PD) is a chronic disease that develops over years and varies dramatically in its clinical manifestations. A preferred strategy to resolve this heterogeneity and thus enable better prognosis and targeted therapies is to segment out more homogeneous patient sub-populations. However, it is challenging to evaluate the clinical similarities among patients because of the longitudinality and temporality of their records. To address this issue, we propose a deep model that directly learns patient similarity from longitudinal and multi-modal patient records with an Recurrent Neural Network (RNN) architecture, which learns the similarity between two longitudinal patient record sequences through dynamically matching temporal patterns in patient sequences. Evaluations on real world patient records demonstrate the promising utility and efficacy of the proposed architecture in personalized predictions.

1 Introduction

Parkinson's disease (PD) is a neurodegenerative disorder encompassing both motor and non-motor symptoms. It progresses over years and varies dramatically in its clinical manifestations and overall prognosis. The PD individuals may have different ages of onset (AO), and may be affected with motor impairment, motor severity, cognitive status, sleep disorders, and cardiac autonomic dysfunction - to varying extents and at varying speed [10, 26, 20]. For any single domain, e.g. motor severity, some patients may have late AO and show rapid progression throughout the course of PD, while some others may have early AO but a slow and stabilized progression [26]. The heterogeneity reveals different underlying disease mechanism such as the biological mechanism (e.g. different dopamingeric dysfunction levels might trigger distinct PD pathology and thus lead to various PD trajectories), however, it may contribute to inaccurate assessment of PD status and, thereby, affects treatment decisions, including the selection of patients for targeted therapies.

A preferred strategy to resolve this heterogeneity and thus enable better prognosis and targeted therapies is to segment out more homogeneous patient subpopulations with similar clinical characteristics. With those similar patient cohorts we can perform targeted prognosis and design customized therapies. Previous comparative studies also provide evidence that personalized models could provide improved predictive performance over global and local models across a range of different healthcare informatics tasks [17, 22, 28].

To identify similar patient cohort, there is no universally accepted consensus on the grouping criteria, also a "top-down" polling based on prior assumptions would be difficult since the approach relies on accurate clinical observations to recognize patterns from all available wide breath of clinical features [19]. Therefore, there is an urgent need for a data-driven approach, in which similarity of the underlying disease mechanism arise from complex patient data. Among existing methods, traditional vector based approaches aggregate patient record sequences to obtain vector based patient representation, and calculate similarity on top of those patient vectors. For example, [28] proposes to use a local spline regression (LSR) based method to embed patient events into an intrinsic space, and then measure the patient similarity by the Euclidean distance in the embedded space. However, this vector based representation neglects all temporal information in patient records.

More recently, deep learning techniques have been adopted in patient representation learning. In [21], the authors develop a deep neural network composed of a stack of denoising autoencoders to process electronic health records (EHR) in an unsupervised manner that captured stable structures and regular patterns in the data and generate a patient representation. Patient similarity is then calculated based on such representation. In [6], the authors adopt "Word2Vec" technique to train a two-layer neural network from a record cor-

*equal contributions
†Key Lab of ADIC, Dalian University
‡Univ. of Washington and IBM T.J. Watson Research Center
§Dept. of Automation, Tsinghua University
¶Computer Science, Dalian University of Technology
‖Computer Science and Eng., Michigan State University
**Weill Cornell Medical School, Cornell University

pus to map each event into a vector space encoding the event contextual correlations. The similarities (e.g. cosine distance) evaluated in such embedded vector space reflect the contextual associations (e.g., event A and B with high similarity suggests they tend to appear in the same context). However, these techniques do not explicitly model dynamic temporal information or tackle the challenges from heterogeneous data sources.

To summarize, despite of their initial successes, the existing methods still have the following limitations:

- They often learn patient similarity via some intermediate representations, rather than directly from sequences. The complexity of relationship between different events could be over-simplified and some critical information could be overlooked in the independently learned representations.
- Temporal dynamics are either not accounted in the learned representation or not accounted in computing patient similarities. The temporal mismatch between patient event sequences (e.g. due to wide range of event lengths and interval lengths) may impact the similarity measure depending on the nature of events, disease mechanism and other factors.
- The experiments in previous works often only involve one type of patient data, e.g. diagnosis in patient Electronic Health Records (EHR). There is a lack of work that learns patient similarity based on the data obtained from multiple modalities.

The proposed work in this paper, which is inspired by the ideas in [27] and [12], aims at addressing those challenges. In [12], the authors use multi-directional RNN to access contextual information and successfully apply the model on the digit recognition task with warped test set. We propose a method to directly learn patient similarity from longitudinal and multi-modal patient data with an RNN architecture that can encode the similarity of two sequences and dynamically match temporal pattern in data. In particular, we treat the inter-relations between events from two sequences as a two-dimensional domain, with prefixes as contextual information. Gated Recurrent Unit (GRU) [5] is used to overcome the vanishing gradient problem and control the information flow to adapt to different time scales. The learned pair-wise similarity representation is robust to event warping and fed into a linear function to learn the final similarity score in a similar fashion as the learning-to-rank approach. Our main contributions are:

- We design a deep learning model to directly compute pairwise patient similarities which capture potentially complex relationships between heterogeneous and longitudinal patient records.
- We borrow the idea of Dynamic Time Warping (DTW) combining with a 2D-RNN architecture using a ranking loss function to learn the similarity between two temporal sequences varying in speed of evolving records. An optimal similarity representation that is robust to warping will be learned.
- We perform extensive experiments on a large scale real world patient dataset obtained from a longitudinal cohort study. Results show that performance of the proposed method significantly outperforms baselines in both similarity learning and personalized predictions tasks.

The rest of the paper is organized as follows. In section 2, we introduce the building blocks of our models and discuss related work in Section 3. Next we introduce the proposed architecture in Section 4, and evaluate it with real world data in Section 6. We also discuss results in Section 7. Last, we conclude our work and highlight future directions in Section 8.

2 Background

2.1 Recurrent Neural Networks (RNN) The recurrent neural network (RNN) is a feed-forward neural net that computes a fixed sequence of learned nonlinear transformations to convert an input pattern into an output pattern. The structure of RNN enables the networks to capture the temporal dynamics and/or perform sequential prediction. The hidden states works as the memory of network such that the current state of the hidden layer depends on previous time. This enables the RNNs to handle variable-length sequence input. Two prominent variants with sophisticated gating mechanisms are widely used: the long short-term memory (LSTM) unit [14], and the gated recurrent unit (GRU) [5]. They are designed to overcome the vanishing gradient problem as well as capture the effect of long-term dependencies. Empirical studies show that GRU and LSTM have comparable prediction performance[7, 16]. However, GRU with a simpler architecture have fewer parameters, which could reduce calculating time, especially in a complex architecture. Thus in this work, we choose to use GRU to overcome vanishing gradient problem and control the information flow to adapt to different time scales.

2.2 Dynamic Time Warping (DTW) The dynamic time warping (DTW) is an approximate pattern detection algorithm that measures similarity between two temporal sequences which may vary in speed. It uses a dynamic programming approach to minimize a predefined distance measure (e.g. Euclidean distance) so that two time series are optimally aligned through a

warping path. DTW has been successfully applied in the speech recognition field to tolerate nonlinear rate variations [25], and also shows great performance in online streaming monitoring [23], DNA sequence mining [1] and entertainment [31]. It is considered the best measure for time series pattern matching across a wide range of application domains[24].

In this work, to handle the temporal dynamics in patient data, we use 2D-RNN to mimic the dynamic programming formula in DTW with the learned gate parameters. By such design, our model would adopt the benefit of DTW and have better alignment for pairs of sequences of events with significant temporal dynamics (e.g. varying inter-event interval lengths and event duration).

3 Related Work

3.1 Patient Similarity Patient similarity is one major research topic in the healthcare informatics domain. Much effort has been done in this area in the past years. For example, [3] proposed a patient similarity algorithm that weights similarity measures using Support Vector Machine. [28] proposed to use a Local Spline Regression based method to embed patient events into an intrinsic space, then measure the patient similarity by the Euclidean distance in the embedded space. These methods do not take the temporal information into consideration when evaluating patient similarities. [29] presented an One-Sided Convolutional Matrix Factorization for detection of temporal patterns. However, none of these methods directly learns patient similarity from heterogeneous multi-modal data with considering temporal timestamp information.

Deep learning models become a recent trend of patient similarity learning. In [4], the authors proposed an adjustable temporal fusion scheme using CNN-extracted features. In [21], the authors develop a deep neural network composed of a stack of denoising autoencoders to process electronic health records (EHR) in an unsupervised manner and then compute patient similarity based on such representation. While in [6], the authors adopt "Word2Vec" technique to train a two-layer neural network from a record corpus to map each event into a vector space encoding the event contextual correlations. The similarities (e.g. cosine distance) evaluated in such embedded vector space reflect the contextual associations (e.g., event A and B with high similarity suggests they tend to appear in the same context). However, these techniques do not explicitly model dynamic temporal information or tackle the challenges of learning from heterogeneous data sources.

3.2 Personalized Prediction in Healthcare Personalized models in the context of healthcare applications have recently been investigated. In [17], the authors performed a series of comparative studies and found that across a range of different bioinformatics classification tasks, personalized models can provide improved accuracy over global and local models. In [22], the authors performed personalized prediction by matching clinical similar patients with a locally supervised metric learning measure. [18] proposed an integrated method to provide personalized treatment and drug design. In [30], the authors proposed a multi-task learning approach to provide personalized prediction model, by building one different predictive model for each patient.

This work extends the patient similarity learning and personalized prediction along a number of important dimensions, including: 1) a deep architecture to directly compute similarity with considering of the temporal dynamics across data obtained from multiple modalities, and 2) personalized prediction of target measures over a few time stamps based on clusters of similar patients.

4 The RNN Architecture with Dynamic Temporal Matching

To enable the challenging personalized prediction task, we introduce an RNN architecture to compute the similarity between sequences of patient data with a DTW-like structure that brings better alignment for sequences with significant temporal dynamics. As shown in Figure 1, the matching structure comprises of three steps: 1) obtain the distance between two patient data sequences, 2) apply 2D-RNN to compute the global distance between two patient data tensors, and 3) apply a linear scoring function to obtain the final distance.

4.1 Similarity between Patient Data Sequence For each patient, we normalize the data from multiple modalities into a unified vector as follows. Given two patient vectors $u(w_i)$ and $u(v_j)$ where $u(w_i) = (w_{i1}, w_{i2}, \ldots, w_{in})$ and $u(v_j) = (v_{j1}, v_{j2}, \ldots, v_{jn})$, a typical way to compute their similarity is to calculate their distance with Euclidean distance measure:

$$\boldsymbol{d}_{ij} = \sqrt{\sum_{k=1}^{n}(w_{ik} - v_{jk})^2}.$$

However, such direct computation is not enough due to its limitation in modeling temporal dynamics. Particularly, although similar patients are more likely to show similarity in overall trends of illness, the temporal events in their data may vary in speed and local arrangement.

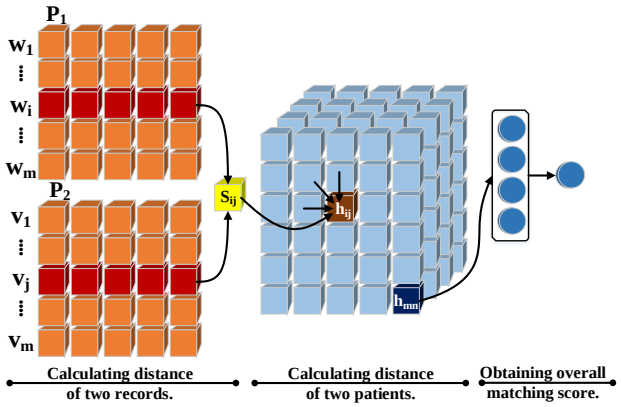


Figure 1: The proposed RNN architecture with dynamic temporal matching

For example, two similar patients go to clinics and get the same diagnosis, but the time stamps of diagnosis can be different, which could introduce noise in learning patient similarity.

4.2 Dynamic Temporal Matching To lift the aforementioned limitations, we get inspired by DTW and design a deep matching structure that warps sequences non-linearly in the time dimension to reduce noise. The DTW distance $dtw(i,j)$ is computed using Eq. 4.1 in a recursive way:

$$dtw(i,j) = d(w_i, v_j) + \min\{dtw(i-1,j), dtw(i,j-1), dtw(i-1,j-1)\}, \tag{4.1}$$

where $d(w_i, v_j)$ refers to the distance of two observations w_i and v_j.

Figure 2 provides a simplified illustration of matching process between patients P_1 and P_2 using the DTW-inspired design. In the illustration, we only consider binary features for simplicity. Given patients $P_1 = (0,0,1,1,1,0,0,0,1,1)$ and $P_2 = (0,0,1,1,0,0,0,1)$, we denote the i-th feature value of P_1 as w_i and the j-th feature value of P_2 as v_j. P_1 and P_2 are represented using waveform in Figure 2. The waveform of P_1 and P_2 look similar overall, however, their similarity cannot be measured using Euclidean distance due to different dimensions. Using DTW, we measure their similarity by aligning P_1 and P_2 in time dimension. For example, when comparing at time stamp 5, we "warp" the time axis by computing the distance of (w_5, v_4) instead of (w_5, v_5) since the distance of (w_5, v_4) is shorter. Similar warping is also needed at (w_8, v_7) and (w_{10}, v_8).

4.3 2D-GRU for Dynamic Temporal Matching In this work, we borrow the idea of DTW and combine

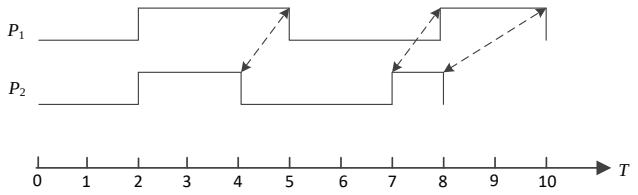


Figure 2: The matching process of DTW

a 2D-RNN architecture with a ranking loss function to learn the similarity between two temporal sequences. The design is described as follows. Denote the records for a patient as $P_1 = \{w_1, ..., w_i, ..., w_m\}$, where w_i is the i-th event. Also denote the prefixes $P[1:i] = \{w_1, ..., w_i\}$ as the sequence of events from the first event to i-th event. Given two patients' sequences of events $P_1 = \{w_1, ..., w_m\}$ and $P_2 = \{v_1, ..., v_n\}$, the distance between prefixes $P[1:i]$ and $P[1:j]$ is determined by distance of sub-prefixes $\vec{h}_{i-1,j}$, $\vec{h}_{i,j-1}$,$\vec{h}_{i-1,j-1}$ and the distance of the current record as shown in Formula 4.2.

$$\vec{h}_{i,j} = f(\vec{h}_{i-1,j}, \vec{h}_{i,j-1}, \vec{h}_{i-1,j-1}, \vec{d}(w_i, v_j)), \tag{4.2}$$

where $\vec{h}_{i-1,j}$ is the distance between prefixes $P_1[1:i-1]$ and $P_2[1:j]$, $\vec{d}(w_i, v_j)$ indicates the distance between the i-th record of P_1 and the j-th record of P_2. Here the f function we choose is a 2D-GRU(gated recurrent unit) as illustrated in Figure 3.

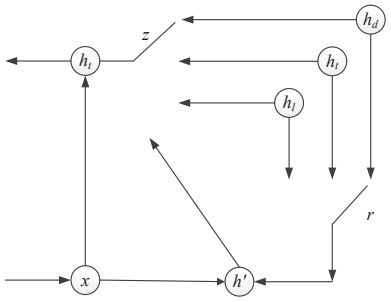


Figure 3: Illustration of 2-D Gated Recurrent Units. The update gate $\vec{z}$ and reset gate $\vec{r}$ control previous unit input of three directions. x directly input its value to output h_{ij}

GRU utilizes two gates to tackle the gradient vanishing and exploding problems of RNN. In this paper, we extend the traditional GRU to 2D-GRU. In 2D-GRU, we use a reset gate $\vec{r}$ to decide how much information of the previous hidden state is discarded and an update gate $\vec{z}$ to decide how much information stored in the hidden state is updated. For a given position (i,j), the input of previous units comes from three directions $(i-1,j)$, $(i,j-1)$ and $(i-1,j-1)$, denoted as l, t, d

respectively. Therefore, the update gate $\vec{z}$ and the reset gate $\vec{r}$ actually control the input information of three directions. Besides the input of previous units, the unit also needs the distance $\boldsymbol{d}_{ij}$ as the input. Thus, input vector $\vec{q}$ is constructed by concatenating four vectors $\vec{h}^T_{i-1,j}$, $\vec{h}^T_{i,j-1}$, $\vec{h}^T_{i-1,j-1}$, and d_{ij}:

$$\vec{q}^T = [\vec{h}^T_{i-1,j}, \vec{h}^T_{i,j-1}, \vec{h}^T_{i-1,j-1}, d_{ij}].$$

Given a input vector $\vec{q}$, $\vec{r}$ and $\vec{z}$ are computed by :

$$\vec{r} = \sigma(W^{(r)}\vec{q} + \vec{b}^{(r)})$$

$$\vec{z} = \sigma(W^{(z)}\vec{q} + \vec{b}^{(z)}),$$

where $W^{(r)}$ and $W^{(z)}$ indicate weight coefficient of reset gate and update gate, respectively, and $\vec{b}^{(r)}$ and $\vec{b}^{(z)}$ are thresholds of reset gate and update gate, respectively. The global matching score $\vec{h}_{i,j}$ is then computed as following:

$$\begin{aligned} \vec{h}'_{ij} &= \phi(\vec{w}d_{ij} + U(\vec{r} \odot [\vec{h}^T_{i-1,j}, \vec{h}^T_{i,j-1}, \vec{h}^T_{i-1,j-1}]^T) + \vec{b})) \\ \vec{h}_{ij} &= W^{(m)}(\vec{z} \odot [\vec{h}^T_{i-1,j}, \vec{h}^T_{i,j-1}, \vec{h}^T_{i-1,j-1}]^T) \\ &\quad + U^{(m)}(1 - \vec{z}) \odot \vec{h}'_{ij} + \vec{w}d_{ij}, \end{aligned} \tag{4.3}$$

where $\vec{h}'_{ij}$ denotes the hidden state, $W^{(m)}$ and $U^{(m)}$ are the parameters of the reset gate, $\vec{w}$ is weight coefficient of distance between two patients. In addition, we use *tanh* for ϕ, and $\odot$ for Hadamard product, of which each element multiply the element in the same position.

4.4 Linear Scoring Function The 2D-GRU scans the input recursively from position $(0, 0)$ to position (m, n), therefore the last output of the model is $\vec{h}_{mn}$, which reflects the global distance between two patients. The overall matching score can be obtained using Formula 4.4:

$$M(P_1, P_2) = W^{(s)}\vec{h}_{mn} + \vec{b}^{(s)}, \tag{4.4}$$

where $W^{(s)}$ and $\vec{b}^{(s)}$ are the parameters of the linear function.

4.5 Optimization We choose pairwise ranking loss as the loss function. Given a triplet (P_1, P_2^+, P_2^-) where the matching distance of (P_1, P_2^+) is shorter than that of (P_1, P_2^-), the loss function is defined as the one in Formula 4.5:

$$\begin{aligned} &L(P_1, P_2^+, P_2^-) = \\ &\max(0, 1 + M(P_1, P_2^+) - M(P_1, P_2^-)) + \lambda\|\Theta\|_2^2 \end{aligned} \tag{4.5}$$

where $M(P_1, P_2^+)$ and $M(P_1, P_2^-)$ are the matching scores of two patients. And the parameters of RNN based matching structure are $\Theta = \{W^{(r)}, \vec{b}^{(r)}, W^{(z)}, \vec{b}^{(z)}, \vec{w}, U, \vec{b}, W^{(m)}, U^{(m)}, W^{(s)}, \vec{b}^{(s)}\}$. Among them, $W^{(r)}$ and $\vec{b}^{(r)}$ are parameters of the reset gate, $W^{(z)}$ and $\vec{b}^{(z)}$ are parameters of the update gate, $\vec{w}$, U and $\vec{b}$ are the parameters of the memory cell, $W^{(m)}$ and $U^{(m)}$ are parameters of dimension transformation, $W^{(s)}$ and $\vec{b}^{(s)}$ are the parameters of the linear function. To minimize the loss, all these parameters are trained using back-propagation and mini-batch Stochastic Gradient Descent with AdaGrad [9].

5 Personalized Prediction

The learned patient similarity could be used in personalized prediction. For a queried patient we firstly retrieve the top N similar patients to form a sub-population. Then we train predictive models based on these similar patients to make personalized prediction for the queried patient. In this paper, we use such approach to predict a set of ordered scores that would reflect the cognitive stages of patients with Parkinson's disease. Note that since we will predict ordered scores for the next several time periods, the prediction task could be considered as an ordinal classification problem. As for the classifiers, the proposed personalized prediction step is flexible to admit any classifier for multi-class classification. Here in this work we employ two classifiers: multiclass logistic regression and multiclass support vector machine (SVM) [8]. For multiclass SVM, we take the one-versus-rest strategy. In addition, since logistic regression takes vector sequences as inputs while a patient often has several records at different time stamps in our data, we consider the latest record next to the prediction window as the input for logistic regression.

6 Experiments

6.1 Datasets To evaluate the performance of the proposed architecture in the task of personalized prediction, we use the data from the Parkinson's Progression Markers Initiative (PPMI) challenge dataset[1]. The PPMI is a landmark observational clinical study that comprehensively evaluates PD cohorts using heterogeneous sources of data including advanced imaging, biologic sampling and clinical and behavioral assessments to identify the conditions of PD progression for patients. The data is sparse, irregular, and longitudinal with a great deal of temporal information embedded in the medical events underpinning the long period progression path of PD, adding more difficulty in learning.

To preprocess the data, we extract all features and select the features that are observed in at least 400 patients' records. Then we leave out the prodromal cohorts and only keep 683 patients whose primary

[1] http://www.ppmi-info.org/

diagnosis are either Idiopathic PD (case) or "No PD nor other neurological disorder" (control). As a result, we have 466 cases and 217 controls.

Then we impute most missing values as well as handle data anomaly. For most of the missing values we use the last occurrence carry forward strategy [11]. If the first record of a patient is missing, we impute it with the first ever observed record of the patient. If all entries of one feature of a patient are missing, we used the mean value of all observed values of this feature across the entire population to impute. For features with integer values, we round up such mean values. For categorical features, we transform and normalize them into one-hot form [13]. In addition, we also handle data anomalies by encoding abnormal entries as 1 and 0 otherwise. Furthermore, we remove those patients with less than 3 sequences. After the above processing, we have data records of 617 patients with 15636 record sequences in all, i.e., each patient has 25 data sequences on average. Table 1 summarizes the data used in evaluation. As the loss function of our model is pairwise, we need to construct a triplet (P_1, P_2^+, P_2^-) by finding a positive patient P_2^+ and a negative patient P_2^- for each patient P_1 . After ranking all the patients according to their similarities with P_1, we select four most similar patients as positive patients and randomly select four patients from those ranking after 200 as negative patients. Consequently, sixteen triplets are built for each patient and the size of the training set increases to 16 times of the number of patients, which helps tackle the over-fitting issue to some extent.

Table 1: Summary of PPMI dataset

Item	Count
patient number	617
record number	15636
feature dimension	319

6.2 Features and Targets For features, we followed [26] and chose 319 raw features in the following 7 categories: 1) motor symptoms/complications (MCs) (SPES/SCOPA sections Motor/Complications), 2) cognitive functioning (SCOPACOG), 3) autonomic symptoms (SCOPA-AUT), 4) psychotic symptoms (SCOPA-PC, items 1–5), 5) nighttime sleep problems and excessive daytime sleepiness (SCOPASLEEP), 6) depressive symptoms [PROPARK: Beck Depression Inventory (BDI); 7) ELEP: Hospital Anxiety and Depression Scale (HADS)8.

In addition, we used Hoehn and Yahr (NHY) scale scores [15] from the last 3 months as the prediction targets. NHY scale is a commonly used system describing how the motor functions of PD patients deteriorate. The scores are ordered and discrete, ranging from 0 to 5. Score 1.0 means that the PD is limited to one side of the body. Other motor conditions such as tremor, rigidity, reduced arm swing, and slowness are present only on one side. Score 2.0 refers to problems affecting both sides. The higher the score, the more severe the condition is.

6.3 Evaluation Method In this section we discuss the evaluation methods for 1) patient similarity learning, and 2) personalized prediction. To evaluate patient similarity learning, we use the Precision@K patients retrieved as the performance measure [2]. Precision@K is calculated as $\frac{\sum_{n=1}^{N} P@K_n}{N}$, where N is the number of patients in the test data set and $P@K_n$ indicates the precision for the n-th test data with a set of top K similar patients by the RNN based matching structure. As there is no ground truth for patients similarity learning, we employed the Euclidean distances of patients as the ground truth for similarity learning. The Euclidean distances are measured by the average values of all the 82 targets from the last 3 month period.

To evaluate the performance of personalized prediction in terms of an ordinal classification task, we choose Root-Mean-Square Error (RMSE) as the measure, which is calculated as $RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}$, where $\hat{y}_t$ indicates the predicted NHY score for patient t and y_t is actual NHY score for patient t. RMSE is a performance measure for ordinal classification. It could measure how close a wrong prediction is to the true value. In addition to RMSE, we also use confusion matrix and micro-average precision, recall, F1 score to demonstrate the experiment results of personalized prediction.

For both tasks, we divide the data of 617 patients into training, validation, and testing sets with a ratio of 8:1:1 and report the performance on test set.

6.4 Experiment Results In the experiments, we set the batch size of Stochastic Gradient Descent (SGD) to be 10. It's a small value since there are only 617 patients in our dataset. All the parameters are initialized randomly following a Uniform distribution. The dimension of the RNN based matching structure is set as 5 after validating values of $d = 2, 5, 10$.

In this paper, we compare with several state-of-the-art baselines to evaluate the performance of our similarity learning and personalized prediction approaches. For similarity learning, we consider Euclidean distance measure as the baselines, which measures the distance

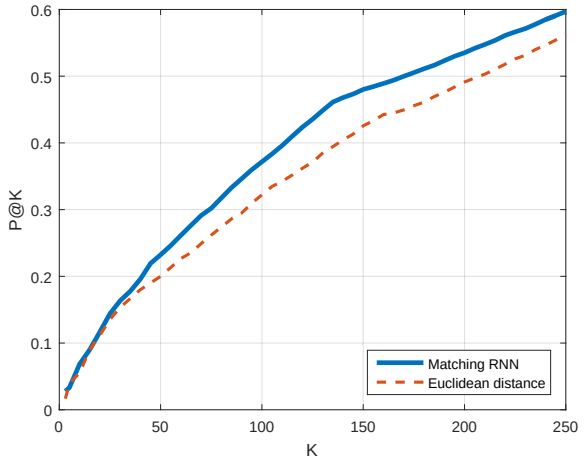


Figure 4: P@K of matching RNN and Euclidean distance.

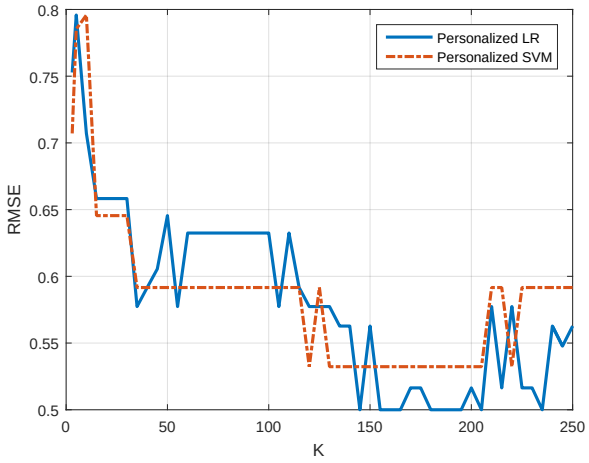


Figure 5: RMSE of personalized LR and personalized SVM varying with K similar patients.

by calculating Euclidean distance of the average feature vectors of two patients. For personalized prediction, we employ Multiclass LR, K-Nearest Neighbors(KNN), multiclass SVM and Long Short-Term Memory (LSTM) as the baselines. To distinguish the traditional methods, the LR and SVM trained on similar patients are called personalized LR and personalized SVM, respectively.

Figure 4 shows the similarity learning results of our method. The evaluation terms (precision @K patients retrieved) of different patients were averaged to obtain the experimental results. The proposed method performed better than the baseline.

To optimize the values of the parameter K on the validation data, we compare the results for K varying from 5 to 250 (Figure 5) and observe that both personalized LR and personalized SVM could achieve the best RMSE at more than one values of K. Since a smaller value of K can increase computational efficiency, we choose the smallest K that generates the best performance. For personalized LR, we set $K = 145$, while for personalized SVM, $K = 120$.

Table 2: Performance of different prediction models, where *Pers.* is the abbreviation for Personalized, and *Mult.* stands for Multiclass

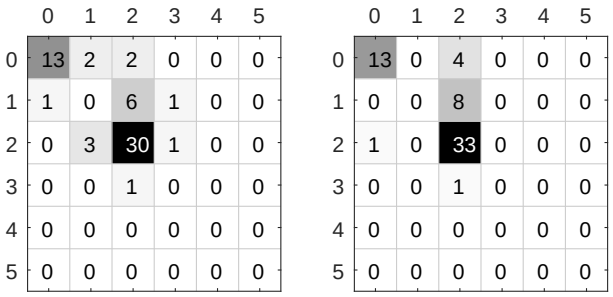
Model	RMSE	micro-P	micro-R	micro-F1
Pers. LR	**0.6583**	**0.7167**	**0.7167**	**0.7167**
Pers. SVM	**0.6952**	**0.7667**	**0.7667**	**0.7667**
Mult. LR	0.7188	0.6833	0.6833	0.6833
Mult. SVM	0.7416	0.7500	0.7500	0.7500
KNN	0.9574	0.5667	0.5667	0.5667
LSTM	0.7853	0.3833	0.3833	0.3833

Table 2 compares the performance of our methods against four baselines including Multiclass LR, Multiclass SVM, KNN, and Long Short-Term Memory (LSTM) [14] in the term RSME, micro-average precision(micro-P), micro-average recall(micro-P) and micro-average F1 score(micro-F1). In addition, we also make performance comparison using the confusion matrix in Figure 6, where columns represent the predicted patient classes and rows represent their true classes. From the results we can observe that the proposed models can gain the largest numbers in cell diagonals of the matrix, i.e. the most correct predicting results. Moreover, compared with Multiclass LR and Multiclass SVM, the proposed Personalized LR and Personalized SVM can gain less wrong predicting results, for example at the cell of line 1 and column 3 in the matrix in Figure 5(a) to 5(d).

7 Discussion

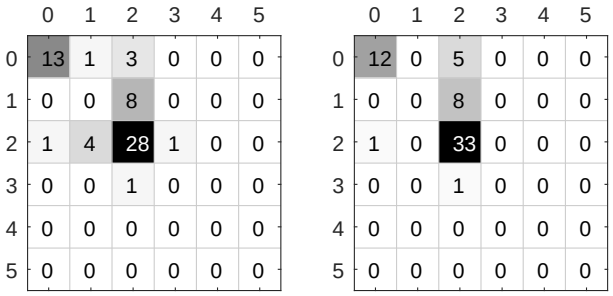
From the experiment, we have the following findings.

- Performance of personalized prediction models is closely related to the size of similar patient subgroup K. From Figure 5, the RSME initially declines as K increases and then becomes stable or increases after K reaches 220.
- The personalized prediction model outperforms baselines including LSTM, one most popular RNN architecture, with a large performance gain due to the personalized prediction setting, as well as the effectiveness of the learned cohorts with similar patients.
- The personalized prediction adopts similar concept



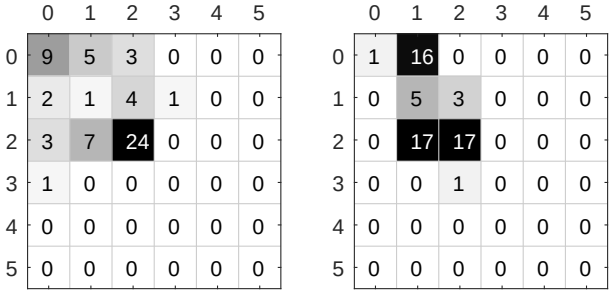
(a): Personalized LR

(b): Personalized SVM



(c): Multiclass LR

(d): Multiclass SVM



(e): KNN

(f): LSTM

Figure 6: Confusion Matrix of six prediction models

as KNN in terms of prediction based on K most similar patients, however, significantly outperforms KNN. The results indicate that the similar patients obtained by our model can enhance the performance of prediction model more effectively than those acquired by Euclidean distance of mean feature vectors since the matching structure might catch the underlying interaction between different patients from the point view of prediction targets.

- In the experiment, the HNY scores are extremely unbalanced. Figure 6 shows that most patients are in stage 0 and 2, while there are few or even none patients in stage 3, 4 and 5. Thus, the stages 1 and 3 next to stage 0 and 2 are prone to misclassification.

8 Conclusion and Future Work

In this work, we have developed a novel deep model for patient similarity learning. The proposed approach directly learns patient similarity from longitudinal and multi-modal patient data with an RNN architecture that can encode the similarity of two sequences and dynamically match temporal patterns in patient data. Based on the learned similarity, we further develop a personalized prediction framework that is flexible to admit various classifiers in the prediction step. We further apply our proposed model on real-world patient data obtained from a longitudinal study of Parkinson's disease, which demonstrates promising utility and efficacy of our method. Potentially, this study could be extended along both the directions of similarity learning and personalized prediction. For example, we could explore better loss functions for more accurate similarity learning. It is worth designing a more effective framework for multi-target predicting.

Acknowledgement

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (http://www.ppmi-info.org/data). For up-to-date information on the study, visit http://www.ppmi-info.org. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including abbvie, Avid, Biogen, Bristol-Mayers Squibb, Covance, GE, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merk, Meso Scale Discovery, Pfizer, Piramal, Roche, Sanofi, Servier, TEVA, UCB and Golub Capital. The work of Chao Che is supported by NSFC No. 91546123. The work of Fei Wang is partially supported by NSF IIS-1650723. The work of Jiayu Zhou is supported in part by ONR N00014-14-1-0631, NSF IIS-1565596 and IIS-1615597.

References

[1] J. Aach and GM. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17:495–508, 2001.

[2] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. SIGIR '04, 2004.

[3] L. Chan, T. Chan, L. Cheng, and W. Mak. Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy. In *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 2010.

[4] Y. Cheng, F. Wang, P. Zhang, and J. Hu. Risk

prediction with electronic health records: A deep learning approach. 2016.

[5] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014.

[6] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. KDD '16, 2016.

[7] Junyoung Chung, ***c***Caglar Gül***c***cehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[9] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[10] SM. Fereshtehnejad, SR. Romenets, JB. Anang, V. Latreille, JF. Gagnon, and RB. Postuma. New clinical subtypes of parkinson disease and their longitudinal progression: A prospective cohort comparison with other phenotypes. *JAMA Neurol.*, 72(8):863–73, 2015.

[11] Andrew. Gelman and Jennifer Hill. *Data analysis using regression and multilevel and hierarchical models*. Cambridge University Press, 2007.

[12] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Multi-dimensional recurrent neural networks. In *Proceedings of the 17th International Conference on Artificial Neural Networks*, pages 549–558, 2007.

[13] David Harris and Sarah Harris. *Digital Design and Computer Architecture, Second Edition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2012.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[15] MM. Hoehn and MD. Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17:427–42, 1967.

[16] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning*.

[17] N. Kasabov. Global, local and personalised modeling and pattern discovery in bioinformatics: An integrated approach. *Pattern Recogn Lett.*, 28(6):673–85, 2007.

[18] N. Kasabov and Y. Hu. Integrated optimisation method for personalised modelling and case studies for medical decision support. In *International Journal of Functional Informatics and Personalised Medicine*, 2010.

[19] M. Lawton, F. Baig, C. Rolinski, C. Ruffman, K. Nithi, M. May, Y. Ben-Sholomo, and M. Hu. Parkinson's disease subtypes in the oxford parkinson disease center discovery cohort. *Journal of Parkinson's Disease*, 5:269–79, 2015.

[20] C. Marras and KR. Chaudhuri. Nonmotor features of parkinson's disease subtypes. *Mov Disord.*, 31(8):1095–102, 2016.

[21] R. Miotto, L. Li, BA. Kidd, and JT. Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6, 2016.

[22] K. Ng, J. Sun, J. Hu, and F. Wang. Personalized predictive modeling and risk factor identification using patient similarity. In *AMIA Summits on Translational Science Proceedings*, 2015.

[23] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Trans. Knowl. Discov. Data*, 7(3), 2013.

[24] Chotirat Ann Ratanamahatana and Eamonn Keogh. Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*, 2004.

[25] Hiroaki Sakoe and Seibi Chiba. Readings in speech recognition. chapter Dynamic Programming Algorithm Optimization for Spoken Word Recognition, pages 159–165. 1990.

[26] SM. Van Rooden, F. Colas, P. Martínez-Martín, M. Visser, D. Verbaan, J. Marinus, JN. Chaudhuri RK, and JJ. van Hilten. Clinical subtypes of parkinson's disease. *Mov Disord.*, 26(1):51–8, 2011.

[27] Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. Match-SRNN: Modeling the recursive matching structure with spatial RNN. 2016.

[28] F. Wang, J. Hu, and J. Sun. Medical prognosis based on patient similarity and expert feedback. In *ICPR '12*, page 1799–1802, 2012.

[29] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *KDD '12*, page 453–461, 2012.

[30] Jianpeng Xu, Jiayu Zhou, and Pang-Ning Tan. Formula: Factorized multi-task learning for task discovery in personalized medical models. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 496–504. SIAM, 2015.

[31] Y. Zhu and D. Shasha. Warping indexes with envelope transforms. *Bioinformatics*, 17:181–192, 2003.