Capturing Feature-Level Irregularity in Disease Progression Modeling

Kaiping Zheng
National University of Singapore
kaiping@comp.nus.edu.sg

Wei Wang
National University of Singapore
wangwei@comp.nus.edu.sg

Jinyang Gao
National University of Singapore
jinyang.gao@comp.nus.edu.sg

Kee Yuan Ngiam
National University Health System
kee_yuan_ngiam@nuhs.edu.sg

Beng Chin Ooi
National University of Singapore
ooibc@comp.nus.edu.sg

Wei Luen James Yip
National University Health System
james_yip@nuhs.edu.sg

ABSTRACT

Disease progression modeling (DPM) analyzes patients' electronic medical records (EMR) to predict the health state of patients, which facilitates accurate prognosis, early detection and treatment of chronic diseases. However, EMR are irregular because patients visit hospital irregularly based on the need of treatment. For each visit, they are typically given different diagnoses, prescribed various medications and lab tests. Consequently, EMR exhibit irregularity at the feature level. To handle this issue, we propose a model based on the Gated Recurrent Unit by decaying the effect of previous records using fine-grained feature-level time span information, and learn the decaying parameters for different features to take into account their different behaviours like decaying speeds under irregularity. Extensive experimental results in both an Alzheimer's disease dataset and a chronic kidney disease dataset demonstrate that our proposed model of capturing feature-level irregularity can effectively improve the accuracy of DPM.

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; • **Computing methodologies** → **Neural networks**; • **Applied computing** → **Health informatics**;

KEYWORDS

Healthcare; Time series; Data analytics; Gated Recurrent Unit

ACM Reference format:

Kaiping Zheng, Wei Wang, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, and Wei Luen James Yip. 2017. Capturing Feature-Level Irregularity in Disease Progression Modeling. In *Proceedings of CIKM'17, Singapore, Singapore, November 6-10, 2017*, 10 pages. <https://doi.org/10.1145/3132847.3132944>

1 INTRODUCTION

Chronic diseases are the major cause of illnesses and deaths. Among the top 10 causes of death in 2010, seven of them were resulted from chronic diseases [14]. Without timely and effective interventions,

Figure 1: A patient's several consecutive electronic medical records, describing visit-level irregularity (top) and feature-level irregularity (bottom) respectively.

chronic diseases are likely to worsen with more severe comorbidities and complications. As a consequence, patients may suffer from poor health and bear high healthcare cost. To address this issue, disease progression modeling (DPM) [22] has been proposed, which employs computational methods to model the progression of a target disease. By predicting the health state of patients accurately, DPM facilitates early detection and treatment of chronic diseases before deterioration. Moreover, DPM not only helps patients to keep good health state and save money, but also eases the allocation for limited healthcare resources.

DPM typically exploits electronic medical records (EMR) to analyze the state of patients. EMR data¹ records patients' visits to hospital with heterogeneous features, including diagnoses, lab tests, prescribed medications, procedures, doctor notes, and image data (e.g., MRI images) together with some demographic information.

One major challenge of DPM over EMR data is on handling the irregularity issue of the time series EMR data, as existing approaches mainly focus on non-time series data (e.g., [27]) or regular time series data (e.g., [19]). There are two levels of irregularities in the EMR data, namely visit-level and feature-level. Visit-level irregularity indicates that EMR data appears irregularly with time. This is due to the fact that patients visit hospital irregularly (mainly determined by their actual physical conditions). Taking Figure 1 as an

¹We shall refer EMR to EMR data in this paper for ease of reference.

This work is licensed under a Creative Commons Attribution International 4.0 License.

CIKM'17, November 6-10, 2017, Singapore, Singapore
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4918-5/17/11.
<https://doi.org/10.1145/3132847.3132944>

example, this patient visits hospital four times and the time spans between consecutive visits vary a lot, i.e., nine days, three months and six months respectively. The features appearing in the four visits include diagnoses chronic kidney disease (CKD), diabetes mellitus (DM), acute kidney failure (AKF) and lab tests glucose (GLU), haemoglobin A1c (HbA1c). Feature-level irregularity indicates that the same feature appears irregularly in EMR data with time, which is caused by two factors. First, visit-level irregularity has a direct effect on feature-level irregularity. As shown in Figure 1, for feature CKD and DM, they appear at every visit, however, the visits are irregular, which leads to irregular features. Second, patients do not receive the same set of lab tests and diagnoses at every visit. Hence, features could be missing in the records of some visits. For instance, in Figure 1, AKF only appears at the second and the last visit. Feature-level irregularity reduces to visit-level irregularity when all features appear at every visit.

Most of the existing work either ignores the visit-level irregularity by partitioning EMR data into regular time windows and aggregating the data within each window (e.g., [19]), or handles the visit-level irregularity by simply adding the time span as one feature in the feature vector [7, 9]. Pham et al. [25] propose to consider the decaying influence of the time span — the longer the time span is, the less influence of the history poses on the current state. However, they only consider the visit-level irregularity and ignore the more fine-grained feature-level irregularity.

We argue that handling the irregularity issue at the feature level rather than the visit level can further improve the accuracy of DPM together with many other EMR data analytic tasks. There are two reasons. First, more fine-grained time span information is captured at the feature level, whereas the visit-level time span is not accurate for some features. Second, different features tend to expose different decaying behaviours, where some features may decay faster than others after the same time lapse. Hence, there is a need to handle various features differently.

We have the following contributions to highlight:

- We identify the irregularity in EMR data and justify that this characteristic should be captured at the feature level for both reducing time information loss and modeling various features' decaying behaviours, such as decaying speeds.
- We propose a new model to make use of the fine-grained feature-level irregular time span information based on the standard Gated Recurrent Unit [6]. Our model differentiates various features through learning their decaying parameters. Furthermore, our proposed model focuses on processing irregular input EMR data, hence is applicable to other EMR data analytic tasks.
- We evaluate the effectiveness of our proposed Feature-Level Time Decay Model for DPM in two datasets. Through comparison with several deep learning-based methods, especially the Visit-Level Time Decay Model, and some advanced DPM methods, we demonstrate that our proposal of capturing feature-level irregularity in EMR data can achieve higher prediction accuracy.

The rest of the paper is organized as follows. In Section 2, we review existing related works. We formulate our problem in Section 3

and then describe our methodology of capturing feature-level irregularity in detail in Section 4. Section 5 contains our experimental evaluation in both a public EMR dataset and a private one from the local hospital. Finally, we conclude in Section 6.

2 RELATED WORK

There are three categories of related work for handling the issue of data irregularity in EMR data analytics. Approaches from the first category aggregate the time series data (i.e., records of all visits) of each patient into a single record and thus avoid the issue. Approaches of the second category transform the problem into a regular time series problem by partitioning the visits into regular time windows and aggregating the records within each time window. Approaches of the third category read the records of all visits directly without transformation. Some techniques are applied to handle the visit-level irregularity. However, notice that all existing methods fail to handle the feature-level irregularity.

2.1 Converting to Non-Time Series Data

Some works make use of patients' baseline features (i.e., features collected when patients visit hospital for the first time) or the average of patients' features aggregated over time. Among them, some works employ statistical regression models to discover the correlation between patients' features and patients' disease severity [11, 12, 26, 27, 29]. These statistical regression methods enjoy the advantages of simple computation and modeling.

Another line of research [23, 31, 32] is to employ multi-task learning [2]. In these works, patients' medical features collected during their first visit to hospital are utilized to predict their disease severity at multiple future time points. The prediction at each future time point is a regression task. Multi-task learning captures the relationship between multiple tasks to improve the performance of all tasks. However, some discriminative information in the historical records would be missed as they only use the record from the first visit.

One limitation in using EMR data as non-time series data might be the under-utilization of time series EMR data. For example, we cannot differentiate the importance of the features from different historical visits if the simple average operation is used to aggregate the features.

2.2 Transforming into Regular Time Series Data

After transforming the irregular time series data into regular time windows, traditional approaches for time series data can be applied. Some researchers employ dynamic Bayesian networks (DBN) or some variant graphical models [13, 28, 30]. Due to DBN's directed graph structure, it enjoys the advantages in causality and interpretability. However, in order to construct a DBN model, medical experts need to determine the causal structures in a certain medical domain in the model. This process may be time-consuming and if the exact interactions between some variables are not clear, it is difficult to determine the corresponding causal relationships. Moreover, this model construction may reduce the chance for finding some hidden variables associated with patients' disease progression

as well as some interesting, hidden causal relationships between variables.

Recently, deep learning models[18] have been applied for regular time series data. Deep learning has attracted huge interests from both academia and industry because of its dramatically better performance in many areas for feature learning, including image classification, object detection and speech recognition. Recurrent neural networks (RNN) are designed for processing sequential data. Long short-term memory (LSTM) [17] and gated recurrent unit (GRU) [6] are two specific RNN models which have drawn much attention and achieved state-of-the-art performance in many applications.

Deep learning methods have also been employed in a wide range of EMR data analytic tasks [21], including ICU mortality prediction [4, 20], diagnosis [19], and interpretable phenotype learning in clinical decision making [5]. These works partition the irregular EMR data into regular time windows (non-overlapping [5, 19, 20] or overlapping [4]). These methods may not be able to capture feature patterns within a short time (i.e., within a window) as they miss the fine-grained visit-level information.

2.3 Without Transforming Data

Choi et al. propose a couple of approaches for DPM without transforming the irregular time series data. In their “GRAM” model [8], they employ a graph-based attention model, and input EMR data of patients’ visits in chronological order without considering the intrinsic irregularity. Their focus is to incorporate domain knowledge in the form of a DAG. In their “DoctorAI” model [7], they employ a RNN model and solve the irregularity by concatenating the time span between consecutive visits in the inputs. Similarly, in their “RETAIN” model [9], they use a reverse time attention mechanism for providing precise interpretability. They utilize irregular EMR data by concatenating the visits’ timestamps in the inputs.²

In [3], Che et al. impute the missing data of patients’ records for utilization and solve the irregularity in EMR data through dynamic time warping to learn the similarity between temporal sequences. Then the proposed dynamic temporal matching technique can help the personalized prediction of Parkinson’s disease.

In [25], Pham et al. propose to use the time span as a decay term to modify the standard LSTM model for the DPM problem. Their proposed decay term is a visit-level decay term which loses feature-wise time span information and does not distinguish between various features. On the contrary, our proposed model captures feature-level irregularity with more fine-grained time information. At the same time, our method differentiates various features’ decaying behaviours over time.

3 PRELIMINARIES

In this paper, we target at predicting a patient’s future health state (or disease severity) in order to model his/her disease progression trajectory. The training data consists of a set of labeled samples, denoted as $\{< x, y, \Delta t >\}$, where x is a list of feature vectors of a patient extracted from his/her historical visits (one vector per visit), and y is the severity score after a period of time Δt since the last visit.

²This concatenation is mentioned in the appendix of [9].

The feature vector of x consists of two types of features, namely medical features and demographic features. Medical features typically include diagnoses such as DM, CKD, AKF, lab tests such as HbA1C and GLU, medications like insulin and procedures like dialysis. They appear irregularly since patients visit hospital irregularly and different lab tests or medications are given to them for each visit. For a patient, we denote the medical feature vector at time t as $(x_1^{(t)}, \dots, x_D^{(t)})^T$, where D is the total number of available medical features. The granularity of t could be a day, a week or a month. Each element could be a real number for the result of a lab test or a binary value for the existence of a diagnosis. Demographic features include height, gender, education (time) etc, which change slowly over time. Therefore, we assume the demographic features are constant in this paper. For a patient, his/her demographic feature vector is denoted as $(g^1, \dots, g^S)^T$, where S is the number of unique demographic features. We collect the features mentioned above to construct our EMR data, which is formalized as follows:

DEFINITION 1. (EMR Data $\mathcal{R}$) $\mathcal{R}$ is the union of all patients’ EMR data. For each patient p , his/her EMR data is a list of feature vectors ordered by time. By concatenating the medical features and the demographic features, we have the feature vector of a patient at time t as

$$x^{(t)} = (x_1^{(t)}, \dots, x_D^{(t)}, g^1, \dots, g^S)^T \quad (1)$$

$x_d^{(t)}$ ($d \in \{1, \dots, D\}$) denotes the value of the d -th medical feature. If the d -th medical feature is not available at time t (e.g., the lab test not conducted), then $x_d^{(t)}$ is denoted missing.

The severity score y is typically derived from the results of some lab tests. We consult doctors to select the specific lab tests for each disease. For example, the value of Mini-Mental State Examination (MMSE) test is a good indicator for the severity of an Alzheimer’s disease patient [24] and the value of Glomerular Filtration Rate (GFR) test represents the severity of a CKD patient [1]. For each patient in our training dataset, we select a cut point time to separate all visits into two parts. The early visits are used to construct x and the recent visits are used to derive y . The formal definition of the cut point is as follows.

DEFINITION 2. (CutPoint t_ψ) Suppose patient p visits hospital at a set of time points $\{t_1, t_2, \dots, t_n\}$. We define a CutPoint t_ψ such that the features at $\{t_1, \dots, t_\psi\}$ are used to construct x , and the features at $\{t_{\psi+1}, \dots, t_n\}$ are used to derive y . Consequently, multiple training samples are constructed from this patient, i.e., $\{< x, y_k, t_{\psi+k} - t_\psi >\}$ for each $k \in \{1, 2, \dots, n - \psi\}$ where the patient has the lab test (i.e., severity score) y at time point $t_{\psi+k}$.

With the description of the input $(x, \Delta t)$ and the output y above, now we define DPM formally as follows.

DEFINITION 3. (Disease Progression Modeling (DPM)) Given a set of training samples $\{< x, y, \Delta t >\}$, the objective of DPM is to obtain a mapping function Φ that minimizes the following loss function over all samples:

$$\mathcal{L}(\Phi(x, \Delta t), y)$$

The loss function $\mathcal{L}(\cdot, \cdot)$ measures the discrepancy of the predicted severity score and the ground truth score. In our experiments, the squared ℓ_2 norm is used. After training, when we deploy the

Figure 2: Disease progression modeling.

Table 1: Notations

Notation	Description
$\mathcal{R}$	EMR data
D, S	Number of unique medical features, static features
x	List of feature vectors of a patient
$x^{(t)}$	Feature vector of a patient at time t , GRU's current input at time t
t_ψ	CutPoint of a patient
$\Phi(\cdot, \cdot)$	Mapping function learned in DPM from $(x, \Delta t)$ to y
$\mathcal{L}(\cdot, \cdot)$	Loss function of DPM
$\tau^{(t)}$	Feature-level irregularity information of a patient at time t
$h^{(t)}$	GRU's hidden state at time t
$y^{(n)}$	Model's output prediction for a patient
Δt	Time between CutPoint and prediction time point
y	Labeled severity score of a patient

model for new patients, the CutPoint is always the time of his/her last visit. In other words, we use historical information to construct x , and use x to predict y . We can compute a patient's severity at different future time points by changing the input value of Δt . In this way, we model this patient's disease progression trajectory and we define this process as DPM.

We illustrate the process of DPM above in Figure 2. Notations used in this paper are summarized in Table 1.

4 METHODOLOGY

In this section, we first describe how to apply the standard GRU model for the DPM problem and then introduce our proposal for handling the issue of feature-level irregularity in EMR data.

4.1 Standard GRU for DPM Problem

GRU is a popular model that performs well in capturing long-term patterns from time series data[10]. To apply the standard GRU for

Figure 3: Neural network for DPM using standard GRU.

EMR data analytics, researchers have to either partition irregular time series into regular time windows [19] or concatenate the time span between consecutive inputs as an input feature [7, 9] as described in Section 2³. After the data transformation, the records (i.e., x) of each patient are fed into GRU as regular time series data. The detailed neural network structure is illustrated in Figure 3 and the computation is formalized as follows.

First, we compute the *reset gate* $r^{(t)}$ following Equation 2, which does affine transformation of the current input $x^{(t)}$ and the previous hidden state $h^{(t-1)}$, and then applies *sigmoid*($\cdot$) activation function.

$$r^{(t)} = \text{sigmoid}(W_r x^{(t)} + U_r h^{(t-1)}) \quad (2)$$

$$\text{sigmoid}(a) = \frac{1}{1 + e^{-a}} \quad (3)$$

Second, we compute the temporary hidden state $\tilde{h}^{(t)}$ as shown in Equation 4. We can see that the *reset gate* plays an important role in computing the new hidden $\tilde{h}^{(t)}$. If $r^{(t)}$ is near zero, then most information from the previous hidden state will be discarded, and the new hidden state is computed mainly based on the current input $x^{(t)}$. The hyperbolic tangent function $\tanh(\cdot)$ involved is defined in Equation 5. " $\odot$ " represents element-wise multiplication.

$$\tilde{h}^{(t)} = \tanh(\tilde{W} x^{(t)} + r^{(t)} \odot \tilde{U} h^{(t-1)}) \quad (4)$$

$$\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}} \quad (5)$$

Third, we compute the *update gate* $z^{(t)}$ (Equation 6) to balance the importance between the temporary hidden state $\tilde{h}^{(t)}$ and the previous hidden state $h^{(t-1)}$ for calculating the new hidden state (Equation 7).

$$z^{(t)} = \text{sigmoid}(W_z x^{(t)} + U_z h^{(t-1)}) \quad (6)$$

$$h^{(t)} = (1 - z^{(t)}) \odot \tilde{h}^{(t)} + z^{(t)} \odot h^{(t-1)} \quad (7)$$

The formulas discussed above is for one GRU at time t , where $t \in \{1, 2, \dots, n\}$ and n is the number of records for a sample. The hidden state of the final GRU is fed into the *Prediction Module* shown in Figure 3 to predict the severity score.

Some existing approaches assume the future time points are with equal time spans [23, 31, 32]. Hence, their methods cannot predict patients' severity at any future time point. Some other existing

³In [25], Pham et al. propose to revise the standard LSTM model to capture visit-level irregularity in the EMR data. We will discuss this later in Section 4.2 and compare with this method (Visit-Level Time Decay Model) in Section 5.

approaches ignore the time span between prediction and current input (predict patients' state in the next visit to hospital regardless of the time span) [8, 9]. Therefore, their methods may lose important time information for the prediction. In order to predict a patient's severity in any future time, e.g. in a time period of Δt , we include Δt as an input argument to our *Prediction Module*. As a result, our DPM model is more flexible in that we manage to predict patients' future disease progression in any future time without constraints. In particular, we combine $h^{(n)}$ and Δt through affine transformation (Equation 8) followed by a ReLU (rectified linear units) activation function (Equation 9).

$$c^{(n)} = \text{ReLU}(W_h h^{(n)} + W_\Delta \Delta t + b_h) \quad (8)$$

$$\text{ReLU}(a) = \max(0, a) \quad (9)$$

Then, we apply a linear regression in the *Prediction Module* to obtain the predicted score $y^{(n)}$ in Equation 10 where q is a vector, and “.” denotes the dot product. ε is a “noise term”, capturing other influencing factors other than $c^{(n)}$.

$$y^{(n)} = q \cdot c^{(n)} + \varepsilon \quad (10)$$

Based on the EMR data we can access, we model this prediction task as a regression problem which gives fine-grained severity prediction. Our proposed method is extendable to classification if we use a softmax function and a cross entropy loss function. Based on this, for diseases of which the severity is divided into classes, given specific medical endpoints, we can also perform classification tasks to predict changes of disease stages which are also interesting from the clinical perspective.

The discrepancy between $y^{(n)}$ and the ground truth y is evaluated by squared ℓ_2 norm as shown in Equation 11. Equation 11 is thus our training objective or loss function. Back-propagation algorithm is applied to update the model parameters, i.e., $\theta = \{W_r, U_r, \bar{W}, \bar{U}, W_z, U_z, W_h, W_\Delta, b_h, q, \varepsilon\}$.

$$\mathcal{L} = \frac{1}{|\{x, y, \Delta t\}|} \sum (y^{(n)} - y)^2 \quad (11)$$

4.2 Feature-Level Time Decay Model

The standard GRU model ignores the time span between each two consecutive time points. In other words, it assumes the time points of each sample are regularly spanned, which applies for certain applications like language modeling. However, EMR data is irregular.

Consequently, we propose a new gate in GRU which controls the decaying scale of the hidden state by considering the time span of each input feature. In this way, fine-grained feature-level irregularity is captured in the adapted GRU.

In our model, the input of each visit includes the time span of each feature. For a visit at time t , we denote each feature d 's time span from its nearest previous occurrence as $\delta_d^{(t)}$. For example, in Figure 4⁴, $\delta_4^{(6)}$ is the time from the second visit to the fourth visit, which is five weeks; $\delta_5^{(6)}$ is the time between the third visit and the fourth visit, which is three weeks. The difference between $\delta_4^{(6)}$ and $\delta_5^{(6)}$ reflects the feature-level irregularity, while the difference of the time spans between $\delta_5^{(3)}$ and $\delta_5^{(6)}$ reflects the visit-level irregularity.

⁴In Figure 4, we choose x_D as the label feature for illustration purpose. Each appearance of x_D after CutPoint gives one y for prediction.

Figure 4: Feature-level irregularity for one patient.

Figure 5: Adapted GRU model for capturing feature-level irregularity in EMR data.

For the first visit of each patient, $\delta_d^{(0)}$ is set as 0 for $\forall d \in \{1, \dots, D\}$. All time spans at time t construct a vector $\tau^{(t)} = (\delta_1^{(t)}, \dots, \delta_D^{(t)})^T$, which is the feature-level time span information and is fed into the GRU in parallel with $x^{(t)}$ as shown in Figure 5.

The time span information is incorporated in our GRU to control the decaying weight of previous hidden state. In particular, our adapted GRU (illustrated in Figure 5) computes a *decay term* q according to $\tau^{(t)}$ and multiplies q to the *update gate* $z^{(t)}$ (Equation 13) which trades off between the previous hidden state and the temporary hidden state for computing the current hidden state. The formula for q is shown in Equation 12. We employ the back-propagation algorithm for updating $\theta = \theta \cup \{W_\tau, b_\tau\}$.

$$q = 1 - \tanh(W_\tau \tau^{(t)} + b_\tau) \quad (12)$$

$$z^{(t)} = \text{sigmoid}((W_z x^{(t)} + U_z h^{(t-1)}) \odot q) \quad (13)$$

There are three principles in designing the new GRU:

- **Capacity** The model should capture the effect of time span and differentiate various features in terms of their behaviours to time span. To be specific, we propose this Feature-Level Time Decay Model based on three considerations: (i) a patient's current state can be inferred based on his/her previous state and current medical features; (ii) the influence of previous state should decay over time and the decay term should be related to

the time lapse; (iii) various medical features have different decaying behaviours/speeds over time and this is achieved through learning W_τ and b_τ .

- **Expressibility** The model is able to express different scenarios. We choose tanh activation function in our decay term ρ for its expressive power. As the value range of tanh function is $(-1, +1)$, ρ 's corresponding value range is $(0, 2)$. In this case, when $\rho = 1$, our proposed adapted GRU model turns to the standard GRU which does not capture feature-level irregularity. When $\rho < 1$, we can forget more on the history by assigning less weight to the previous hidden state. When $\rho > 1$, we give more weight to the previous hidden state to remember more of the history. We do not choose other common activation functions like $\text{sigmoid}(\cdot)$ or $\text{ReLU}(\cdot)$ due to their value ranges. For $\text{sigmoid}(\cdot)$ activation function, its range is $(0, 1)$ which corresponds to $\rho \in (0, 1)$ meaning that we always need to forget more on the history. For $\text{ReLU}(\cdot)$ activation function, its range is $[0, +\infty)$ which causes $\rho \in (-\infty, 1]$. Both cases are not suitable in our situation. Therefore, our adapted GRU with tanh function has enough expressive power to capture feature-level irregularity in EMR data.
- **Robustness** The model should be robust to feature missing. If patient p 's feature d is missing at time t , we set $\delta_d^{(t)} = 0$ and fill the feature value $x_d^{(t)}$ as d 's value in his/her nearest previous record. In other words, *forward imputation* is applied. As $\delta_d^{(t)}$ is set to 0, this imputed feature value will not influence the learning of parameters W_τ and b_τ . Consequently, missing features will not affect the learning of features' behaviours under irregularity.

5 EVALUATION

5.1 Experimental Set-Up

5.1.1 Datasets. We evaluate our proposed model using a public Alzheimer's disease dataset from Alzheimer's Disease Neuroimaging Initiative (ADNI)⁵, and a chronic kidney dataset from National University Hospital (NUH) in Singapore.

ADNI dataset is a public dataset whose EMR data includes proteomic information, lab tests, as well as some demographic features. All features of the ADNI dataset are divided into five categories as shown in Table 2. We request all patients' data from **ADNI1**, which is a six-year study launched in 2004.

We denote the time of each patient's first visit as M00 and the visit time after M00 as Mxx where xx is the number of months since the first visit. For instance, M06 indicates six months after the first visit. We use such EMR data as the input feature vector of the machine learning models. All models are evaluated based on their performance in predicting the severity scores of patients in the future. The severity is measured by MMSE test ($\in [0, 30]$), which is widely used for Alzheimer's disease patients [12, 24, 27]. The lower the MMSE is, the more severe the patient is. Using MMSE score, patients can be categorized into normal/elderly (MMSE=30), mild

Table 2: Feature Categories of the ADNI Dataset

Category	Dimension	Description & Examples
MRI	325	Extracted from MRI scans using FreeSurfer by UCSF, including Volume (Cortical Parcellation), Volume (WM Parcellation), Surface Area, Cortical Thickness Average, Cortical Thickness Standard Deviation
CSF	5	Three features: p-tau, t-tau and A β -42; two ration features: p-tau to A β -42 and t-tau to A β -42
META	146	Four types of features: cognitive scores including MMSE and ADAS-Cog; demographic features (gender, age and education time); genetic features; lab test features
PET	70	Extracted from PET scans
PROT	146	Proteomics features

impairment ($23 < \text{MMSE} < 30$) and possible Alzheimer's disease ($\text{MMSE} \leq 23$) [12].

NUH dataset is a real-world EMR dataset recording medical features including diagnoses, lab tests, medications, procedures, case notes, and demographic features for more than 100,000 patients admitted to the hospital in the year of 2012. We choose patients with Stage 3 CKD or higher as our cohort [1] and we denote it as the "**NUH-CKD**" dataset. All models are evaluated based on their performance in predicting patients' future GFR ($\in [0, 60]$), which is a lab test indicating the severity of CKD patients. The lower this GFR value is, the more severe the patient is. According to the consultation with doctors and *The Renal Association*⁶, the loss of GFR score over a year is of vital importance to CKD patient management. If a patient loses GFR more than $5\text{ml/min}/1.73\text{m}^2$ over a year, his/her health state is slowly deteriorating, and he/she needs assessment from specialists. In our experiments, we extract the first year data for each patient as input data to predict his/her future GFR scores.

We summarize the statistics of the ADNI1 dataset and the NUH-CKD dataset in Table 3. To be specific, we show the number of features (both medical and demographic ones), the number of patients and the time span for two datasets. Features within a short time period are aggregated into a single record for processing. The time period is six months for the ADNI dataset and one week for the NUH-CKD dataset. Notice that this aggregation is different from the window-based approach, which uses the records from empty windows whereas we remove empty windows. Consequently, their input data is regular and our input data is irregular. Moreover, we use multiple CutPoint settings in our experiments. The corresponding number of samples for each CutPoint setting is included in the table. The CutPoint setting is a trade-off between the sample length and number of samples. When the CutPoint is set larger (further

⁵<http://adni.loni.usc.edu/>

⁶<http://www.renal.org/information-resources/the-uk-eckd-guide/deteriorating-function>

Table 3: Dataset Statistics

Dataset	ADNI1 Dataset	NUH-CKD Dataset
# of medical features	591	603
# of demo. features	3 — age, gender, education time	2 — age, gender
# of patients	819	2740
Time span	4 years, M00 to M48, (“M” — “month”)	1 year, W00 to W52, (“W” — “week”)
# of time steps	7 (aggregated by every 6 months)	52 (aggregated by every week)
CutPoint (t_ψ) setting	M12, M18, M24	W16, W24, W32
# of samples	t_ψ =M12: 1529 t_ψ =M18: 1200 t_ψ =M24: 558	t_ψ =W16: 3601 t_ψ =W24: 2793 t_ψ =W32: 1585

from each patient’s first visit), the sample has a larger length, i.e., more records, but the number of samples is reduced.

5.1.2 Evaluation Metrics. Considering the task of patient severity prediction as a regression problem, we choose the following two performance metrics for all models, namely mean squared error (MSE), and Pearson product-moment correlation coefficient (R) value.

Suppose the predicted severity value is $y^{(n)}$ and the ground truth value is $y^{(gt)}$ for patient p , then MSE can be computed by

$$MSE = \frac{1}{|p|} \sum (y^{(n)} - y^{(gt)})^2 \quad (14)$$

We also compute R value to measure the linear relationship between predicted values and ground truth values.

$$R = \frac{\sum (y^{(n)} - \bar{y}^{(n)})(y^{(gt)} - \bar{y}^{(gt)})}{\sqrt{\sum (y^{(n)} - \bar{y}^{(n)})^2} \sqrt{\sum (y^{(gt)} - \bar{y}^{(gt)})^2}} \quad (15)$$

where $\bar{y}^{(n)}$ is the average of all predicted values and $\bar{y}^{(gt)}$ is the average of all ground truth values. This R value with the range $[-1, 1]$ measures the linear relationship between two variables, so $R > 0$ means they are positively correlated and $R < 0$ means they are negatively correlated.

To sum up, a good regression model should have a small MSE value but a large R value.

5.1.3 Baseline Methods. We denote our proposed model in Section 4.2 as **Feature-Level Time Decay Model**. It is proposed based on GRU. Hence, we first compare our model with three GRU-based baselines.

- **Window-Based Model** partitions input time series (i.e., EMR data) into regular time windows and aggregates the records within each window. This method does not consider the data irregularity. [19] employs this method for diagnosis.
- **Visit-Level Model** handles the visit-level irregularity by concatenating the time span between consecutive irregular visits as a dimension of the input feature vector [7, 9].
- **Visit-Level Time Decay Model** decays the influence of previous visits (i.e., records) on the current hidden state according to the time span between consecutive irregular visits [25].

To obtain a more thorough analysis, we compare our proposed method with several advanced disease progression modeling methods proposed in [31]. [31] employs a multi-task learning (MTL) method, which uses the EMR data from a single visit to predict patients’ severity of multiple future visits (one per task).

- **Least Convex Fused Group Lasso (cFSGL)** [31] includes three penalty terms in the objective function: (i) Lasso penalty which can select a specific set of features for each task, (ii) Group Lasso penalty which can offer a common feature selection for all tasks and (iii) Fused Lasso penalty which incorporates the temporal smoothness. According to [31], setting the parameter for the Group Lasso penalty term as 100 achieves the best results in all their settings. Hence, we also choose the parameter 100 and denote this method as **cFSGL** when reporting experimental results.
- **Least Non-Convex Fused Group Lasso (nFSGL)** [31] is a non-convex progression model proposed to solve the problem in which convex models may be sub-optimal. Zhou et al. propose two formulations of such non-convex models and we denote them as **nFSGL-1** and **nFSGL-2** in our experiments.

For all experiments, we partition all prediction samples into 80% (training data), 10% (validation data) and 10% (testing data) randomly.

5.1.4 Experimental Setting. When carrying out experiments, we need to set the CutPoint defined in Section 3. For GRU-based models, we use all input before (inclusive) CutPoint as input data. For MTL-based methods, we use the nearest historical visit before the CutPoint as the input data. Then we define the “task” in MTL according to the time span between the input and the prediction (output) visit. For instance, in the ADNI dataset, Task1 predicts the severity for the visit that is six months later from the input visit. We compute the performance by a weighted-average over all tasks, with the weight being the portion of samples of each task.⁷

We use the glorot uniform initializer [15] for initializing the parameters in all GRU-based models and choose the RMSProp optimizer [16] (with a weight decay of 0.9) which is usually good for recurrent neural networks. For training models, we choose learning rates 0.001, 0.002 and 0.004 dependent on different datasets or different CutPoint settings for the same dataset.

5.2 Experimental Results in the ADNI Dataset

In the ADNI dataset, we test the performance of all models by varying the CutPoint shown in Figure 6. The nearer CutPoint is to M00 (the first visit), the smaller the input length is for GRU-based models, while the larger the number of prediction samples are.

For MSE results, our findings are as follows. For the same CutPoint setting, from Window-Based Model to our proposed Feature-Level Time Decay Model, MSE value is mainly on the descending trend, which indicates that (i) handling the irregularity of EMR data is beneficial for DPM (Visit-Level Model achieves smaller MSE than Window-Based Model); (ii) modeling time span’s decaying effect on

⁷In MTL experiments, we delete the tasks with fewer than 10 prediction samples in the ADNI dataset and the tasks with fewer than 50 prediction samples in the NUH-CKD dataset. This is because stable performance cannot be obtained in the task with too few training samples and testing samples.

Figure 6: Experimental results in the ADNI dataset.

Figure 7: Experimental results in the NUH-CKD dataset.

time series EMR data improves the accuracy of DPM (Visit-Level Time Decay Model is more accurate than Visit-Level Model); (iii) handling irregularity at the feature level provides more benefits than at the visit level (Feature-Level Time Decay Model outperforms Visit-Level Time Decay Model).

Furthermore, when setting the same CutPoint, our proposed Feature-Level Time Decay Model can give more accurate predictions than MTL-based methods. However, not all GRU-based models can outperform MTL-based methods. When CutPoint is moved further from M00 (from M12 to M18, further to M24), MSE values of GRU-based models decrease as these methods manage to utilize more time series inputs, but the performance of MTL-based methods does not exhibit much difference. As a consequence, the advantage of GRU-based models over MTL-based methods becomes larger when CutPoint changes from M12 to M18 and then M24.

Among MTL-based methods, we observe that nFSGl-1 and nFSGl-2 obtain smaller MSE values than cFSGl, indicating that non-convex progression models can provide more accurate predictions than convex progression models. This finding agrees with the experimental results of [31].

For R value results shown in Figure 6, we can have similar observations as in MSE results. From Window-Based Model to Feature-Level Time Decay Model, the general increasing trend of R values

indicates that the performance is mainly improving. We also find that all GRU-based models tend to have higher R values than MTL-based methods. One reason is that GRU-based models exploit more time series information than MTL-based methods.

5.3 Experimental Results in the NUH-CKD Dataset

We report the experimental study in the NUH-CKD dataset using three different CutPoint settings — W16, W24, W32. The performance of all models is illustrated in Figure 7.

When looking at the MSE metric, in most cases, GRU-based models are ordered as Window-Based Model, Visit-Level Model, Visit-Level Time Decay Model and Feature-Level Time Decay Model, from larger MSE to smaller MSE. Similar to Section 5.2, such observations indicate that (i) fine-grained irregular visit-level time span benefits prediction accuracy; (ii) visit-level time span has a decaying influence in prediction; (iii) more fine-grained feature-level time span can further improve the prediction performance of DPM. For each CutPoint setting, GRU-based models outperform MTL-based methods. The reason could be the lack of sufficient samples for each task of MTL-based methods in the dataset.

When varying the CutPoint from W16 to W24, we find that GRU-based models tend to achieve larger MSE values, hence, the performance degrades. This may be caused by the decreasing number of samples from 3601 in W16 to 2793 in W24 shown in Table 3. A small sample number poses a negative effect on the prediction performance as the model is likely to overfit without enough training samples. However, when further increasing the CutPoint from W24 to W32, MSE values of GRU-based models gets smaller. The reason may be when the CutPoint is W32, GRU-based models manage to utilize more time series features despite decreasing sample number from 2793 in W24 to 1585 in W32. In a nutshell, both the sample length and sample size affect the model performance.

For R values, in the same CutPoint, we find GRU-based models exhibit similar performance as in MSE values. Furthermore, GRU-based models can achieve higher R values than MTL-based methods. The reason may be two-fold: (i) GRU-based models manage to make full use of all time series data as input; (ii) MTL-based methods are affected by the small sample number for each task.

5.4 Case Study of DPM

To give more insight of DPM, we choose three example CKD patients corresponding to three representative disease progression trajectories from the NUH-CKD dataset and plot the prediction results of our proposed model in Figure 8. To be specific, we use our proposed Feature-Level Time Decay Model trained with the CutPoint set to W32, which achieves smallest MSE value in all GRU-based models among all CutPoint settings (shown in Figure 7). We use this model to predict the three patients' future severity in terms of GFR values after CutPoint. Note that a lower GFR value corresponds to a more severe state. In Figure 8, the green triangles before the 32nd week represent patients' GFR values fed to our model as input. The orange ones denote the ground truth GFR values in the future between the 32nd and the 52nd week. As the data collected only covers one year for each patient, we have no access to patients' GFR values after the 52nd week. The blue curve is the predicted severity progression from our proposed model, which can extend to future time points after the 52nd week. These three patients correspond to three representative progression types demonstrated as follows.

Severe and deteriorating progression Patient1's GFR values first keep decreasing in the first 32 weeks and drop to around 35. Then from the 32nd week, the GFR ground truth values remain in the descending trend, and our proposed model manages to predict this trend. Furthermore, our proposed model predicts that as time further goes on, the loss of Patient1's GFR values will exceed $5\text{ml}/\text{min}/1.73\text{m}^2$ within one year. Our model would suggest Patient1 to consult specialists for expert assessment according to medical reference.

Mild yet deteriorating progression In the beginning, Patient2's state is mild as his/her GFR values indicate only moderately reduced kidney function. However, the GFR value decreases slowly over time before the 52nd week predicted by our model⁸. After the 52nd week, our model predicts that Patient2 will suffer from a minor

⁸The ground truth value fluctuates without an obvious trend for this patient.

Figure 8: Disease progression modeling illustration for three representative CKD patients in the NUH-CKD dataset (best viewed in colour).

drop in GFR value, indicating that the functioning of his/her kidney might slightly deteriorate. With this information in mind, healthcare workers may need to provide more aggressive interventions to Patient2 in advance according to his/her current condition.

Severe yet stable progression Patient3 is already in CKD Stage5 [1] when his/her data is recorded. Through the whole year, this patient progresses stably without much change in GFR value, and our model gives the prediction that this stableness will maintain for a long time.

From the analysis above, we can see that our proposed model provides accurate prediction for patients with different progression patterns. With a good DPM model, we can provide healthcare workers with the accurate disease progression trajectory of each patient. Based on the trajectory, they can give helpful advice to patients, take effective interventions beforehand and finally, have better disease management on patients. For instance, healthcare

workers may give more aggressive treatment to Patient1 and Patient2 in our previous example, and guarantee the monitoring for Patient3.

6 CONCLUSION

In this paper, we identify the irregularity characteristic residing in EMR data both at the visit level and at the feature level. We argue that capturing feature-level irregularity can benefit the performance of EMR data analytics based on: (i) the current state of a patient is inferred from his/her previous state and the current observations (i.e., lab tests, diagnoses, etc.); (ii) the influence of the previous state decays through time, and the decay term is related to the time lapse; (iii) different features have different effects on the decaying behaviours/speeds.

We propose a model to handle feature-level irregularity based on GRU. It decays the influence of previous information on patients' current state and learns decaying parameters for different features. We use DPM as a representative EMR data analytic task for evaluating our proposal. We conduct experiments in both a public ADNI dataset and a private NUH-CKD dataset for two chronic disease cohorts. In the experiments, we compare with several GRU-based models and some advanced DPM methods — MTL-based methods. Experimental results on comparison with GRU-based models demonstrate that our proposal of capturing feature-level irregularity manages to improve the prediction accuracy of DPM task effectively.

ACKNOWLEDGMENTS

The work in this paper is supported by National Research Foundation, Prime Ministers Office, Singapore under its Competitive Research Programme (CRP Award No. NRF-CRP8-2011-08). Wei Wang is supported by FY2017 SUG Grant, National University of Singapore. We would like to thank Xiangrui Cai and Associate Professor Chonho Lee for their advice on the earlier work of this paper. We would also like to thank the anonymous referees for their insightful comments on improving this work.

REFERENCES

- [1] The Renal Association. 2017. CKD stages. (2017). Retrieved Aug 30, 2017 from <http://www.renal.org/information-resources/the-uk-ckd-guide/ckd-stages>
- [2] Rich Caruana. 1998. Multitask learning. In *Learning to learn*. Springer, 95–133.
- [3] Chao Che, Cao Xiao, Jian Liang, Bo Jin, Jiayu Zhao, and Fei Wang. 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*. SIAM, 198–206.
- [4] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 507–516.
- [5] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2015. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv:1512.03542* (2015).
- [6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1724–1734.
- [7] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*. 301–318.
- [8] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2016. GRAM: Graph-based Attention Model for Healthcare Representation Learning. *arXiv preprint arXiv:1611.07012* (2016).
- [9] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 3504–3512.
- [10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- [11] Orla M Doyle, Eric Westman, Andre F Marquand, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kloszewska, Hilkka Soininen, Simon Lovestone, Steve CR Williams, et al. 2014. Predicting progression of Alzheimer's disease using ordinal regression. *PloS one* 9, 8 (2014), e105542.
- [12] Simon Duchesne, Anna Caroli, Cristina Geroldi, D Louis Collins, and Giovanni B Frisoni. 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *NeuroImage* 47, 4 (2009), 1363–1370.
- [13] Konstantinos P Exarchos, Themis P Exarchos, Christos V Bourantas, Michail I Papafaklis, Katerina K Naka, Lampros K Michalis, Oberdan Parodi, and Dimitrios I Fotiadis. 2013. Prediction of coronary atherosclerosis progression using dynamic Bayesian networks. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE, 3889–3892.
- [14] National Center for Chronic Disease Prevention and Health Promotion. 2017. Chronic Disease Overview. (2017). Retrieved Aug 30, 2017 from <http://www.cdc.gov/chronicdisease/overview/>
- [15] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, Vol. 9. 249–256.
- [16] Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. In *COURSERA: Neural Networks for Machine Learning*.
- [17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
- [19] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2016. Learning to diagnose with LSTM recurrent neural networks. *International Conference on Learning Representations* (2016).
- [20] Zachary C Lipton, David C Kale, and Randall Wetzell. 2016. Modeling Missing Data in Clinical Time Series with RNNs. In *Machine Learning for Healthcare*.
- [21] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2017. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* (2017), bbx044.
- [22] DR Mould. 2012. Models for disease progression: new approaches and uses. *Clinical Pharmacology & Therapeutics* 92, 1 (2012), 125–131.
- [23] Liqiang Nie, Luming Zhang, Yi Yang, Meng Wang, Richang Hong, and Tat-Seng Chua. 2015. Beyond doctors: future health prediction from multimedia and multimodal observations. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 591–600.
- [24] Jeffrey R Petrella, R Edward Coleman, and P Murali Doraiswamy. 2003. Neuroimaging and Early Diagnosis of Alzheimer Disease: A Look to the Future 1. *Radiology* 226, 2 (2003), 315–336.
- [25] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 30–41.
- [26] Matthias B Schulze, Kurt Hoffmann, Heiner Boeing, Jakob Linseisen, Sabine Rohrmann, Matthias Möhlig, Andreas FH Pfeiffer, Joachim Spranger, Claus Thamer, Hans-Ulrich Häring, et al. 2007. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes care* 30, 3 (2007), 510–515.
- [27] Cynthia M Stonnington, Carlton Chu, Stefan Klöppel, Clifford R Jack, John Ashburner, Richard SJ Frackowiak, Alzheimer Disease Neuroimaging Initiative, et al. 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage* 51, 4 (2010), 1405–1413.
- [28] Marcel AJ Van Gerven, Babs G Taal, and Peter JF Lucas. 2008. Dynamic Bayesian networks as prognostic models for clinical patient management. *Journal of biomedical informatics* 41, 4 (2008), 515–529.
- [29] P Vemuri, HJ Wiste, SD Weigand, LM Shaw, JQ Trojanowski, MW Weiner, David S Knopman, Ronald Carl Petersen, CR Jack, Alzheimer's Disease Neuroimaging Initiative, et al. 2009. MRI and CSF biomarkers in normal, MCI, and AD subjects predicting future clinical change. *Neurology* 73, 4 (2009), 294–301.
- [30] Xiang Wang, David Sontag, and Fei Wang. 2014. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 85–94.
- [31] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, and Jieping Ye. 2012. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1095–1103.
- [32] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. 2011. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 814–822.