Temporal Clustering with External Memory Network for Disease Progression Modeling

Zicong Zhang[1], Changchang Yin[1,2], Ping Zhang[1,2]
[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, USA
[2]Department of Biomedical Informatics, The Ohio State University, Columbus, USA
Email: {zhang.5157, yin.731, zhang.10631}@osu.edu

***Abstract*—Disease progression modeling (DPM) involves using mathematical frameworks to quantitatively measure the severity of how certain disease progresses. DPM is useful in many ways such as predicting health state, categorizing disease stages, and assessing patients' disease trajectory, etc. Recently, with the wider availability of electronic health records (EHR) and the broad application of data-driven machine learning methods, DPM has attracted much attention yet remains two major challenges: (i) Due to the existence of irregularity, heterogeneity, and long-term dependency in EHRs, most existing DPM methods might not be able to provide comprehensive patient representations. (ii) Lots of records in EHRs might be irrelevant to the target disease. Most existing models learn to automatically focus on the relevant information instead of explicitly capture the target-relevant events, which might make the learned model suboptimal. To address these two issues, we propose Temporal Clustering with External Memory Network (TC-EMNet) for DPM that groups patients with similar trajectories to form disease clusters/stages. TC-EMNet uses a variational autoencoder (VAE) to capture internal complexity from the input data and utilizes an external memory work to capture long-term distance information, both of which are helpful for producing comprehensive patient health states. Last but not least, the k-means algorithm is adopted to cluster the extracted comprehensive patient representation to capture disease progression. Experiments on two real-world datasets show that our model demonstrates competitive clustering performance against state-of-the-art methods and is able to identify clinically meaningful clusters. The visualization of the patient representations shows that the proposed model can generate better patient health states than the baselines.**

***Index Terms*—disease progression modeling, deep learning, temporal clustering**

I. Introduction

With the recent development of deep learning and the accumulation of electronic health records (EHR), also known as time-series data, there has been an increasing effort in clustering EHR data in order to discover meaningful patterns throughout longitudinal health information. Moreover, chronic diseases, such as Parkinson's disease (PD) and Alzheimer's disease (AD), can have various outcomes even with a limited number of patients. Such diseases are heterogeneous in nature and often evolve at unique patterns that trigger distinct responses to therapeutic interventions based upon different conditions [1]. Thus, it has become crucial to develop a disease progression modeling (DPM) system to capture certain progression patterns, provide early detection to critical situations, and yield clinically helpful information to improve the quality of care.

Traditionally, DPM or disease clustering/staging is developed by domain experts with extensive clinical experience, in which disease stages are defined separately and based solely on the values of one or a few biomarkers [2], [3]. Nevertheless, developing a DPM system requires long-term observation and human labor, and the result is often based on known biomarkers and acknowledged covariants, which makes it difficult to develop a DPM system for disease with limited knowledge on biomarkers that have not been well-studied. In recent years, the rapid growth of data-driven machine learning methods has motivated a great effort in developing DPM models. There are two main approaches when it comes to DPM: 1) The problem is formed as a risk prediction task with label information based on patient representation that is extracted from the last layer of the model. [4]–[8]. 2) The problem is formed as a traditional unsupervised, patient clustering/subtyping problem where the model is trained to separate the patient into multiple groups [9]–[11]. Leveraging disease outcomes during the training process can prevent the model from forming heterogeneous clusters. However, for certain diseases, diagnosis labels are often unavailable at each patient visit due to limited knowledge of the disease. Moreover, deep learning models that are designed for supervised tasks may not perform well when training in an unsupervised fashion. Therefore, there is a need for developing a DPM framework that can handle both situations with respect to the availability of training labels. However, most developed deep learning models for disease progression modeling suffers from the following limitations:

- **Irregularity and heterogeneity**: Many diseases are heterogeneous in nature and EHR data often has high internal complexity. Due to the complexity of effectively encoding various health conditions into patient representation, accurate DPM still remains a challenging problem.
- **Long-term Dependency**: RNNs are long known to suffer from modeling long-term dependency since it tends to forget earlier information when the input sequence is long. Disease progression modeling, especially for chronic disease, requires long-term observation of the patient in order to provide a comprehensive view for decision making.
- **Target Awareness**: Most rnn-based methods derive patient representations directly from the hidden states of

the model. Such an approach neglect the contribution of target-relevant information. In fact, real-world clinical decisions made by doctors are often based upon past diagnoses as well.

To address these challenges, we propose Temporal Clustering with External Memory Network (TC-EMNet) for disease progression modeling via both supervised and unsupervised settings. TC-EMNet leverages a variation autoencoder framework and a memory network to deal with data irregularity and long-term dependency problems of RNNs respectively. At each time step, TC-EMNet takes EHR medical records as input and encodes the input feature using a recurrent neural network to get hidden representations. Then TC-EMNet samples from the hidden state to form a latent representation. Meanwhile, the hidden state is stored in a global-level memory network, which in turn outputs a memory representation based on current memory cells. The memory representation is then concatenated with the current latent representation to form the patient representation at the current time step. When the training label is available, the model also employs a patient-level memory network to process label information up to most recent visit and outputs target-aware memory representation. We combine memory representations from global-level and patient-level memory networks using a calibration process. TC-EMNet is trained with reconstruction objective under unsupervised setting and prediction objective under supervised setting.

In this paper, our contributions are four fold:

- We propose a novel deep learning framework, namely TC-EMNet for disease progression modeling under both supervised and unsupervised settings.
- TC-EMNet uses a combined recurrent neural network and variational auto-encoder (VAE) architecture to capture the irregularity in data and heterogeneity nature of the disease.
- Under superviesd setting, TC-EMNet employs dual memory network architecture to leverage both hidden representations from the input data and clinical diagnosis to produce accurate patient representations.
- Experiments on two world datasets show that TC-EMNet yields competitive clustering performance over state-of-the-art methods and is able to find clinically interpretable disease clusters/stages.

The remainder of the paper is organized as follows. Section II briefly reviews existing works related to DPM, temporal clustering, and VAE. Section III describes the technical details of the proposed model (TC-EMNet). Section IV and V present experimental results and discussions. Finally, Section VI concludes the paper.

II. Related Work

A. Disease Progression Modeling

Disease progression modeling (DPM) plays a very important role in the healthcare domain, especially for chronic diseases such as Parkinson's Disease (PD) and Alzheimer's Disease (AD). A well-performed disease progression modeling system can not only provide early detection or diagnosis but also discover clinically meaningful patterns for certain groups of trajectories. Most probabilistic models for DPM are based on the hidden markov model (HMM). For example, [12] derived a deep probabilistic model based on sequence-to-sequence architecture to model progression dynamic on UK Cystic Fibrosis Registry. [9] introduced a continuous-time Markov process to learn a discrete representation of each progression state. Moreover, deep learning methods have also been developed for disease progression modeling. [13] proposed a CNN-based model to jointly learn features from MR images combined with demographic information to predict Alzheimer's Disease progression patterns. [14] designed a prediction framework using generative models to forecast the distribution of patients' outcomes. DPM can be regarded as a classification problem, where diagnosis labels are leveraged in favor of model training. On the other hand, DPM can also be seen from an unsupervised perspective where the goal is to discover potential disease states or patient subtypes throughout patients' medical history [15]. However, DPM still remains a challenge due to the high complexity of data introduced by irregular progression patterns for certain chronic diseases.

B. Temporal clustering

Temporal clustering, widely known as time-series clustering, is a data-driven method to cluster patients into subgroups based on time-series observation. Temporal clustering can be considered a challenging task often because of the high dimensionality of the dataset and multiple time steps for each data sample. Recent advances have been focused on leveraging the latent representation learned by recurrent neural network (RNN) for temporal clustering, which was motivated by the success of RNN modeling time-series data. Moreover, due to the emerging availability of electronic health records (EHR) that introduced large-scale and normalized context for individual patients, the deep learning approach become capable of learning more comprehensive patterns and achieving better performance on several critical tasks. [16] introduces a time-aware mechanism to long short term memory cells to capture progression patterns with irregular time-interval. [4] proposed an actor-critic algorithm for predictive clustering where, instead of defining a similarity measure for clustering, a cluster embedding is trained to represent each disease stage. [17] proposed an auto-encoder to reconstruct relevant features for sepsis with attention and showed that the proposed model can identify interpretable patient subtypes. Nevertheless, there is only limited literature that focuses on DPM using temporal clustering techniques.

C. Variational Autoencoder

Variational autoencoder (VAE) is a type of generative model that can handle complicate distributions. VAEs are effective against modeling complex data structures and are widely adopted to solve many real-world problems range from image generation to anomaly detection [18]–[20]. It has also several successful applications with healthcare data

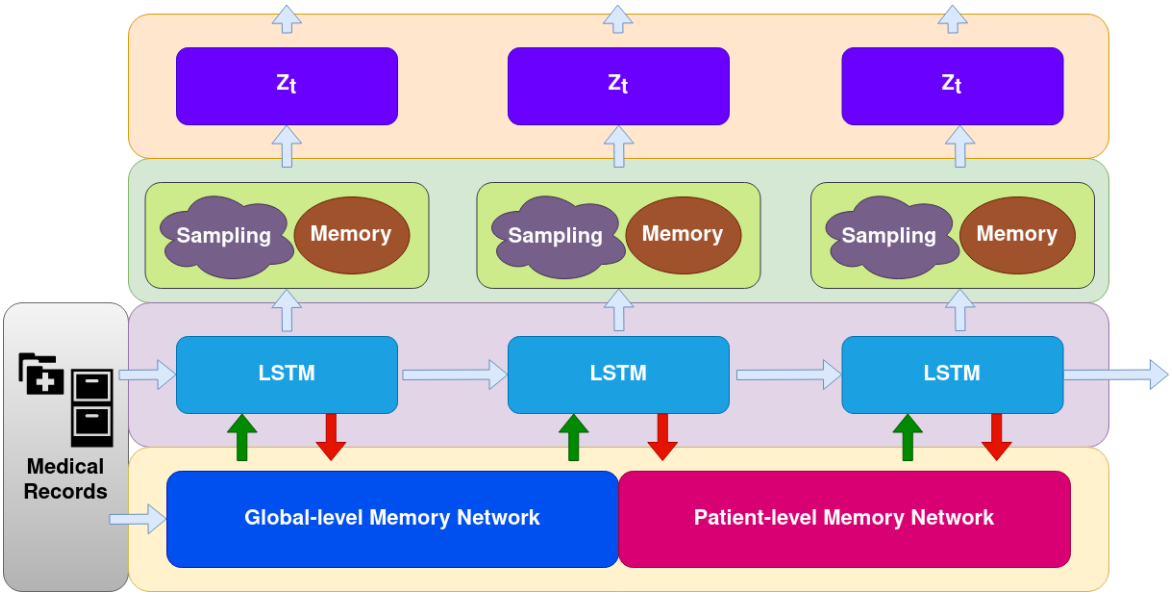


Fig. 1: Overview of the proposed framework. At each timestamp, hidden representation from the encoder network is updated with the memory state to produce disease clusters/stages based on current and previous observations.

[21]. [22] proposed to use VAE to impute missing values for electronic health data with uncertainty-aware attention. Experiments on real-world datasets show that VAE can capture the complexity of EHR distribution. [14] leveraged the VAE framework to forecast disease states for Parkinson's Disease (PD) and Alzheimer's Disease (AD). Nonetheless, the latent representation learned from VAE can be drawn from unrealistic distribution if trained without any constraints.

III. Methodology

A. Problem Definition

Let $x \subseteq \boldsymbol{X}$ and $y \subseteq \boldsymbol{Y}$ be the random variables for input feature space and label space accordingly. Here we focus on a clustering problem, where we are given a population of time-series data $D = \{(x_t^n, y_t^n)_{t=1}^T\}_{n=1}^N$ consisted of paired sequences of observations (x, y) for N patients. $t \subseteq 1, ..., T$ denotes the time stamps for each patients at which the observations are made.

We aim to identify K clusters for time-series data, each corresponding to a disease stage. Each cluster consists of homogeneous data samples, represented by the centroids based on certain similarity measures.

B. Method

This section presents our proposed framework. Here we discuss disease progression modeling under both supervised and unsupervised settings, where our proposed question requires estimating the underlying distribution of all possible disease stages. Such a DPM framework can help the doctors identify meaningful characteristics in both times when a disease has certain diagnosis labels but possible underlying disease stages and when a disease has no well-defined diagnosis labels.

The framework consists of three components: the encoder, the memory network, and the clustering network. For each patient, a recurrent neural network is deployed to encode the patient's information. The memory network controls the overall long-term information at each timestamp. Specifically, when a hidden representation h_i is generated based on current and previous observation $X_{<i}$ at timestamp i, the hidden state is read by the memory network and updates the memory storage. Next, a latent variable z_i is drawn from the prior distribution $p_\theta(z_i|X_{<i})$ conditioned on the hidden state that is generated from the memory network. Then, we either yield prediction outcomes or reconstruct the current observation X_i accordingly. We take the hidden presentation from the last layer of the model for clustering.

1) Encoder Network: The encoder network takes the current observation and the hidden state from the previous timestamp and yields the hidden representation that can interact with the external network. Specifically, a LSTM cell is adopted to generate and update the hidden state:

$$h_t = LSTM(X_t, h_{t-1}), \tag{1}$$

where X_t is the current observation at timestamp t and h_{t-1} is the hidden state from previous step. At each timestamp, the encoder network maps a sequence of time-series input $x_{1:t}$ to a hidden representation $z_t \subseteq Z$, where Z is the subspace of latent representation. The hidden representation will be interacting with the external memory network to form an accurate representation.

2) Memory Network: Long-term information plays an important role in disease progression modeling, since, in the context of chronic disease, the health conditions from the past will affect the current disease stages of the patient. In addition,

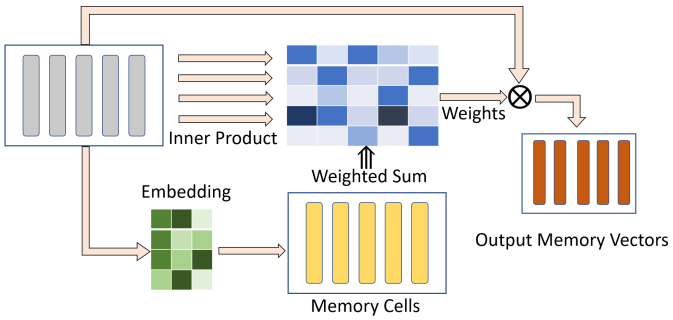


Fig. 2: Overview of the proposed memory network. Hidden states are first written into the memory cells and read by the clustering network to produce a comprehensive representation.

historical information should be stored in an efficient way such that it can provide useful guidance towards the patient's current health state at different timestamps. To this end, we propose an external memory network to capture long-term information throughout the progression modeling process. Our proposed memory network is closely related to [23], which has several successful applications in the field of natural language processing. Similarly, we define memory slots to represent historical information that can be extracted and retrieved at any given timestamp. At each timestamp, the hidden state from the encoder network is recorded and written into the memory cells. By pushing through a series of observations, the memory network will process continuous representations for each individual visit so that a more comprehensive review of the patient can be utilized during the clustering/staging process.

a) Memory Reading: We denote a clinical sequences record $r_t, t = 1, ..., T$, where t stands for index or timestamps of the given record. In memory network, after receiving a hidden representation z_t from the encoder network, the network will produce an external representation e_t based on reading weight $w_t^l, l = 1, ..., T$ of the memory slots. Specifically, e_t can be expressed as:

$$\begin{aligned} e_t &= \sum_l w_t^l m_t, \\ w_t^l &= \textbf{softmax}(\alpha_t, m_t, h_t) \\ &= \frac{\exp(\alpha_t^l) C(h_t, \mathbf{M}_t)}{\sum_j \exp(\alpha_t^j) C(h_t, \mathbf{M}_t)} \end{aligned} \tag{2}$$

where $l = 1, ...L$ denotes the number of memory slots, $m_t \subseteq \mathbb{R}^{1\times D}$ is the memory representation with hidden size D. α is the strength vector that can be learned through the reading operation and $C(\dots)$ is the cosine similarity measure. Memory reading operation is built upon the idea that not all records in the sequence contribute equally to the current health state of the patient. Hence, the weights are computed using the softmax function based on the cosine similarity of the current hidden states and all the previous memories.

b) Memory Writing: Memory writing stores latent representation into memory slots. We use a fixed number of slots to denote the overall memory size. The dimension of the continuous space for each memory slot is d and we use D to denote the dimension of hidden representation z_t. The hidden state is non-linearly projected into the memory space using a $d\times D$ matrix $\mathbf{A}$, $m_t = Az_t$, where m_t is the new input memory representation. Memory writing aims to filter out non-related information and stores only personalized information based on the current hidden state. Mathematically, memory writing can be expressed as:

$$\mathbf{M}_t = r_t \mathbf{M}_{t-1} + v_t h_t, \tag{3}$$

where r_t and v_t are gated vectors that control the information flow between the previous and current memory vector.

3) Clustering Network: After obtaining the representation of the observation h_t through the encoding network, i.e the prior network, and updating the memory cell m_t at current timestamp t, we follow the traditional framework of variational autoencoder (VAE) [24] to compute the mean and standard deviation vectors through the posterior network. We assume that the output is a Gaussian distribution. The computation process can be expressed as:

$$\begin{aligned} \mu_{z_t} &= f_{posterior}([h_t, x_i]) \\ \sigma_{z_t} &= f_{posterior([h_t, x_i])} \end{aligned} \tag{4}$$

where h_t is the hidden state and x_i is the observation at timestamp t. $f_{posterior}$ is posterior functions described by feed-forward neural networks. We then draw samples from the posterior Gaussian distribution using the reparameterization trick:

$$z_t = \mu_{z_t} + \sigma_{z_t} \odot \epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, 1)$, and z_t is the latent representation. $\odot$ indicates element-wise multiplication. The reparameterization trick allows the gradient to backpropagate through the sampling process. Lastly, depends on the availability of diagnosis labels, the clustering network will be trained on two different objectives. When diagnosis label is used, the clustering network is directly trained to predict the label information:

$$\hat{Y}_t = f_{pred}(f_{con}([z_t, m_t])), \tag{6}$$

where f_{pred} is a feed-forward network that outputs probabilities of each label. When diagnosis label is not available, we trained the framework to reconstruct the observation x_i from the latent variable z_t conditioned on the memory state m_t, denote as:

$$\hat{X}_t = f_{recon}(f_{con}([z_t, m_t])), \tag{7}$$

where $\hat{X}_t$ is the reconstructed input, $frecon$ is a feed forward network and f_{con} is the concatenation. During cluster phase, we use euclidean distance-based k-means algorithm on the latent variable z_t.

4) Dual Memory Network Architecture: Under a clinical setting, doctors often provide diagnosis labels based on patients' current and past medical events. Such information can be target health conditions or a diagnosis. Under supervised setting when the label is available during training, we further

Algorithm 1 TC-EMNet

Initialize encoder and decoder network parameters θ, σ;
Initialize memory embedding and memory slots;
for (every time stamp t) **do**
    Compute patient hidden encoding through Encoder network via Eq. (1);
    Read from global-level memory network to extract recent memory representations via Eq. (2);
    if diagnosis is available **then**
        Read from patient-level memory network to extract recent memory representations via Eq. (2);
        Compute memory representation via Eq. (8);
    end if
    Compute loss via Eq. (4) - (7);
    Write to corresponding memory slots via Eq. (3);
end for
Update parameters by optimizing Eq. (12), (13) accordingly;

utilize a patient-level memory network to capture diagnosis information during each visit. Compared to global-level memory network, patient-level memory network at current memory slot M_t can only access diagnosis up to previous timestamp, namely, $\{Y_i | i = 1, 2, ..., t-1\}$. patient-level memory network only reads and writes diagnosis information which later is combined with a global-memory network for clustering. We propose a calibration process to integrate representations from two memory networks, as follows:

$$\begin{aligned} h_t^{global} &= M_{global}(X; \alpha_t^{global}, m_t^{global}, h_t^{global}), \\ h_t^{patient} &= M_{patient}(Y_{<t}; \alpha_t^{patient}, m_t^{patient}, h_t^{patient}), \\ h_t^{final} &= h_t^{global} \odot \sigma(f_{embed}(h_t^{patient})), \end{aligned} \tag{8}$$

where M_{global} and $M_{patient}$ is the global-level and patient-level memory network respectively. This memory calibration process can be regarded as a point-wise attention mechanism.

C. Objective Function and Optimization

Here, we present our training objectives and optimization process. As mentioned in previous sections, the entire network can be trained from end to end using maximum likelihood estimation (MLE). To solve the intractable marginalization for the latent variable z_i, we use the variational lower bound parameterized by q_ϕ to approximate the true distribution, which we assume to be Gaussian. After the memory work reading and writing, We use the latent variable z_i at timestamp i to identify the disease stages. Here we restrict the latent variable to be a multivariate Gaussian distribution, which enforces the same for the posterior. We learn the generative parameter θ using maximum likelihood estimation (MLE):

$$\theta = \underset{\theta}{\operatorname{argmin}} \sum_i \log \int [p_\theta(x_i|z_i, X_{<i}) \times p_\theta(z_i|X_i)] \, dz_i \tag{9}$$

However, the marginalization of z_i is intractable for complicated functions (for instance neural networks). Thus, we need to derive a variational lower bound (i.e. variational Bayesian method) to approximate the logarithm of the marginal probability of the observation, which is as follows:

$$\begin{aligned} &\log p_\theta(x_i|z_i, X_{<i}) \\ &= \log(\mathbf{E}_{q\phi}[\frac{p_\theta(x_i, z_i|X_{<i})}{q_\phi(z_i|x_i, X_{<i})}]) \\ &\geq \mathbf{E}_{q\phi}[\log p_\theta(x_i, z_i|X_{<i})] \\ &- \mathbf{E}_{q\phi}[\log q_\phi(z_i|x_i, X_{<i})], \end{aligned} \tag{10}$$

where the inequality can be obtained using Jensen's inequality and the variational lower bound involves the probability q_ϕ that are parameterized by ϕ, which eventually approximate the intractable true posterior distribution $p_\theta(z_i|X_i)$. Since health-related data is often associated with high-dimensional and general more complicated distribution, we introduce the latent variable z_t to capture the internal stochasticity from the data. We can train the entire clustering network end-to-end using stochastic optimization techniques. After obtaining the variational lower bound, the optimization follows the KL divergence that is the difference of log-likelihood and the variational lower bound:

$$\begin{aligned} &\mathcal{L}_{variation}(\theta, \phi) \\ &= -KL[q_\phi(z_t|h_t, X_{<i})||p_\theta(z_t|h_t, X_{<i})] \\ &= \mathbf{E}_{\log q_\phi(z_t|z_t, X_{<i})} - \mathbf{E}_{\log p_\theta(z_i, x_i|X_{<i})} \end{aligned} \tag{11}$$

where ϕ and θ represents the model parameter and proxy posterior accordingly. The equation holds if the distribution of q_ϕ is equal to the true distribution. When diagnosis label is used during training, we use the cross-entropy loss to directly predict the outcome from the combined latent representation denoted as:

$$\begin{aligned} &\mathcal{L}_{objective}(\theta, \phi, x_t) = \\ &\alpha \mathcal{L}_{variation}(\theta, \phi) + \mathcal{L}_{CE}(Y, \hat{Y}; \theta, \phi, x_t), \end{aligned} \tag{12}$$

When the model is trained in a unsupervised manner, the overall objective function combined with the reconstruction loss becomes:

$$\begin{aligned} &\mathcal{L}_{objective}(\theta, \phi, x_t) = \\ &\alpha \mathcal{L}_{variation}(\theta, \phi) + \mathcal{L}_{recon}(X, \hat{X}; \theta, \phi, x_t), \end{aligned} \tag{13}$$

where we use the mean square error (MSE) for reconstruction loss and α is a hyperparameter to prevent VAE from KL vanishing problem. We adopt a linear annealing schedule for α based on training steps denoted as:

$$\alpha = \min(1, \frac{\textit{training step}}{x}), \tag{14}$$

where x is a threshold value. Last but not least, we use the k-means algorithm [25] on the patient representation to perform clustering.

TABLE I: Results of proposed methods and other methods on ADNI datasets. ↓ indicates that the smaller the better (0=best, and 1=worst). ↑ indicates that the greater the better (0=worst, and 1=best).

	w/o label			with label		
Model	Purity ↑	NMI ↑	RI ↑	Purity ↑	NMI ↑	RI ↑
RNN	0.6799±0.00	0.1415±0.01	0.1406±0.02	0.8532±0.00	0.4020±0.01	0.3805±0.01
Bi-LSTM	0.6810±0.02	0.1540±0.02	0.1559±0.02	0.8674±0.00	0.4092±0.01	0.4042±0.02
RETAIN	0.6903±0.02	0.1787±0.01	0.1671±0.01	0.7144±0.02	0.2572±0.01	0.1838±0.03
Dipole	0.6839±0.00	0.1707±0.01	0.1452±0.00	0.8904±0.01	0.4674±0.01	0.4776±0.02
StageNet	0.6943±0.01	**0.2002±0.01**	0.1791±0.01	0.8513±0.01	0.4045±0.03	0.3744±0.01
AC-TPC	-	-	-	0.8214±0.03	0.3362±0.07	0.3827±0.09
VAE	0.6651±0.02	0.1023±0.02	0.1117±0.02	0.6495±0.04	0.1718±0.05	0.1042±0.04
Memory Network	0.6887±0.02	0.1392±0.01	0.1584±0.02	0.8262±0.01	0.3603±0.01	0.3538±0.02
TC-EMNet^{-u}	**0.7040±0.01**	0.1967±0.02	**0.1891±0.02**	0.8904±0.00	0.4679±0.01	0.4889±0.01
TC-EMNet^{-s}	-	-	-	**0.9126±0.01**	**0.4789±0.01**	**0.4923±0.02**

TABLE II: Results of proposed methods and other methods on PPMI datasets. ↓ indicates that the smaller the better (0=best, and 1=worst). ↑ indicates that the greater the better (0=worst, and 1=best).

	w/o label			with label		
Model	Purity ↑	NMI ↑	RI ↑	Purity ↑	NMI ↑	RI ↑
RNN	0.7221±0.00	0.3089±0.01	0.3120±0.01	0.7640±0.02	0.4222±0.04	0.3663±0.03
Bi-LSTM	0.7264±0.00	0.3170±0.00	0.2976±0.01	0.7674±0.03	0.4456±0.05	0.3575±0.05
RETAIN	0.5241±0.02	0.1188±0.01	0.0619±0.01	0.7510±0.01	0.4072±0.03	0.3361±0.01
Dipole	0.7233±0.00	0.3200±0.00	0.3153±0.00	0.8033±0.01	0.4947±0.01	0.4476±0.02
StageNet	0.7252±0.01	0.3305±0.00	0.3234±0.01	0.7839±0.01	0.4700±0.03	0.3840±0.01
AC-TPC	-	-	-	0.8151±0.01	0.4984±0.03	**0.5129±0.01**
VAE	0.7161±0.00	0.3576±0.01	0.3153±0.00	0.7942±0.01	0.4452±0.00	0.3782±0.01
Memory Network	0.6996±0.01	0.2809±0.01	0.2581 ±0.02	0.7689±0.01	0.4482±0.01	0.4597±0.01
TC-EMNet^{-u}	**0.7452±0.00**	**0.3773±0.00**	**0.3742±0.01**	0.8256±0.00	**0.5053±0.00**	0.4823±0.01
TC-EMNet^{-s}	-	-	-	**0.8339±0.00**	0.5035±0.00	0.4993±0.01

IV. Experiments

We evaluated our proposed model on two real-world datasets, Alzheimer's Disease Neuroimaging Initiative (ADNI) and Parkinson's Progression Markers Initiative (PPMI) dataset. All dataset can be accessed on IDA website[1]. The code can be found on GitHub[2].

A. Datasets

1) ADNI Dataset: Alzheimer's disease (AD) is a chronic neurodegenerative disease that is often related to behavior and cognitive impairment. ADNI is a longitudinal study that aims to explore early detection and tracking of AD based on imaging, biomarkers, and genetic data collected throughout the process [26]. The dataset consists of a total of 11651 visits over 1346 patients with 6 months intervals. For each patient, 21 variables are collected and processed, including 16 time-varying features (brain function, cognitive tests) and 5 static features (background, demographics). 3 diagnose labels are assigned by doctors at each visit for the patient, including control normal (CN), Mild Cognitive Impairment (MCI), and AD, which indicates the severity of how AD symptoms have progressed on each patient.

2) PPMI Dataset: Parkinson's Progression Markers Initiative (PPMI) is a longitudinal study aiming to evaluate patients' progression on Parkinson's disease (PD) based on biomarkers [27]. The dataset consists of a total of 13685 visits over 2145 patients with irregular time intervals. For each patient, 79 features based on motor and non-motor symptoms are collected, including cognitive assessment, lab tests, demographic information, and biospecimens. Since the dataset does not provide a diagnosis label per visit for each patient, we use Hoehn and Yahr (HY) scores as labels for our evaluation. HY scores, ranges from 0 to 5, indicate the severity of patients' symptoms of Parkinson's disease. We use the mean and last occurrence carried forward method to impute missing values.

[1]https://ida.loni.usc.edu/

[2]https://github.com/Ericzhang1/TC-EMNet.git

B. Baselines

We compare our proposed model to several state-of-the-art methods, ranged from vanilla RNNs to multi-layer attention models. Since here we consider disease progression modeling under both supervised and unsupervised settings, we adjusted the architecture of the baseline models to fit the objective accordingly. For baselines that cannot be modified interchangeably, we did not collect the result under the corresponding setting. For all experiments, we use k-means clustering on the hidden representations from the last layer to report the clustering performance.

- **RNN** [28]: A single RNN cell with an additional layer of feed-forward neural network. The model is trained with cross-entropy loss and reconstruction objective accordingly.
- **Bi-LSTM** [29]: Similar to RNN model, a Bi-directional LSTM is used with a reconstruction objective, the model takes both directions of the sequence data into account and is showed to capture richer information compare to single direction.

- **RETAIN** [30]: An interpretable deep learning model that is based on recurrent neural network and reverse time attention mechanism. The RETAIN model learns the importance of hospital records through attention weights. We modify the last layer of RETAIN and train the model based on the prediction and reconstruction objective.
- **Dipole** [31]: A interpretable bidirectional recurrent neural network that employs attention mechanism to leverage both past and future visits. We use concatenation-based attention mechanism for testing and, similar to RETAIN, we adjust the last layer of the model accordingly.
- **StageNet** [5]: A recent risk prediction model that learned to extract disease progression patterns during training and leveraged modified LSTM cell with an attention mechanism. The progression pattern at each timestamp is re-calibrated accordingly using a convolution network.
- **AC-TPC** [4]: A recent deep predictive clustering network that consists of an encoder network, selector, and a predictor. The model is first initialized using a prediction objective and then optimized to train a cluster embedding using the actor-critic algorithm. This method cannot be trained without label information.
- **VAE** [32]: A vanilla variational autoencoder model using a LSTM cell as encoder and trained with prediction and variation objective respectively. Note that this baseline method can be served as an ablation example against our proposed method.
- **Memory Network**: A vanilla global-level memory network with reading and writing mechanism described previously. The network reads and writes the EHR sequence directly and the k-means algorithm is applied directly to the hidden memory representation.
- **TC-EMNet**$^{-u}$: Unsupervised version of TC-EMNet. When the training label is not available, only a global-level memory network is used to produce memory representation. We also train the model for the prediction task and set it as an ablation example against supervised version of TC-EMNet.
- **TC-EMNet**$^{-s}$: Supervised version of TC-EMNet. When the training label is available, a patient-level memory network is used to combine with a global-level memory network to produce target-aware memory representations.

TABLE III: Hyperparameter Searching Space

Hyperparameter	Range
hidden size	$[32, 64, 128, 256]$
latent variable size	$[32, 64, 128, 256]$
x	$[500, 700, 900]$
learning rate	$[1e-2, 1e-3, 1e-4, 1e-5]$
batch size	$[64, 128]$

C. Model Training and Implementation Details

As mentioned previously, our proposed network is continuous and differentiable. We can train the network using stochastic optimization techniques. All neural networks in the proposed network are feed-forward networks. We implemented our solution using Pytorch [33] and trained the model on a single Nvidia Volta V100 GPU with 16GB memory. We adopt gradient accumulation when dealing with out-of-memory problems. We select hyperparameters through random search as shown in table III. For our model, we set both hidden size and latent variable size to be 128. We adopt Adam optimizer with a learning rate of $1e-3$. The model is trained with batch size 32 for 70 epochs. x is set to 700. We split the dataset into training, validation, and testing set with a ratio of $3/1/1$ and report the performance of 5 fold cross-validation for both datasets. A detailed description of the optimization process of our proposed framework can be found in Algorithm 1. The average running time of our proposed framework on both datasets is about 2 hours for cross-validation. For the implementation of other baseline methods, we implement RNN and Bi-lstm methods with Pytorch. We adopt implementations from Pyhealth [34] for RETAIN, Dipole, and StageNet. And we adopt implementation from [4] for AC-TPC. All baseline methods share the same hyperparameter searching space.

D. Evaluation Metrics

To evaluate the clustering performance of our model, we use purity score (purity), normalized mutual information (NMI) [35], and adjusted rand index (ARI) [36]. Purity score is ranged between 0 to 1, indicating the extent to which a cluster is consist of single class. NMI (0 to 1) represents the mutual information between each clusters with 1 being perfect clustering. ARI derives from the Rand index and measures the percentage of the correct cluster assignment. Mathematically, the metrics can be expressed as follows:

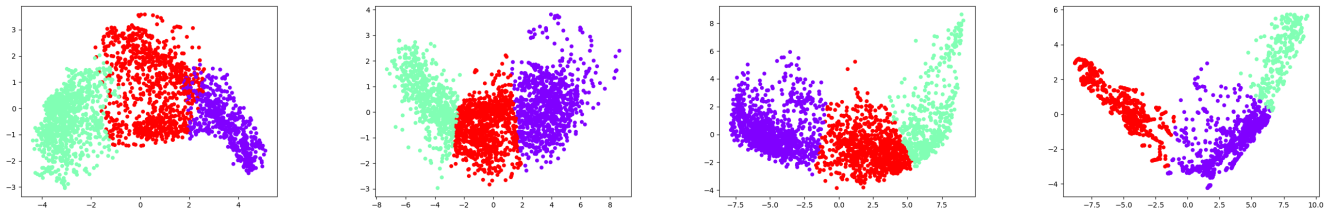
$$\begin{aligned} \textbf{Purity} &= \frac{1}{N}\sum^{j}\max|c_j \cap l_j|, \\ \textbf{NMI} &= \frac{2\cdot \mathbf{I}(c_j, l_j)}{[H(C)+H(L)]}, \\ \textbf{ARI} &= \frac{RI - E(RI)}{\max(RI) - E(RI)}, \end{aligned} \tag{15}$$

where N is the total number of samples, c_j and l_j denotes the cluster assignment and true label respectively, $I(\cdot)$ is the mutual information function and $H(\cdot)$ is the entropy, $E(\cdot)$ and *RI* are the expectation value and Rand index accordingly.

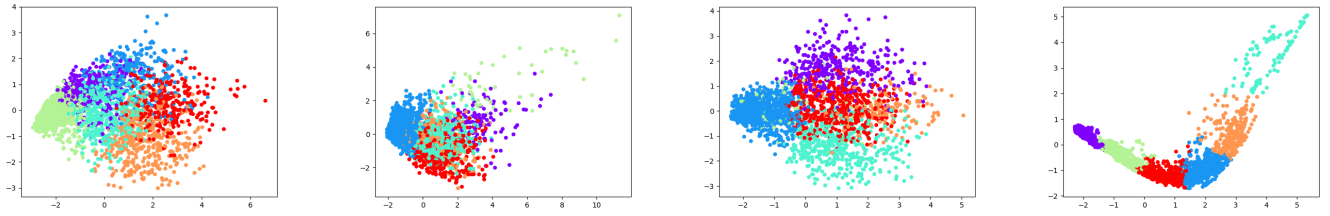
V. Results

A. Clustering Performance

A quantitative comparison of the clustering performance on ADNI and PPMI dataset is shown in table I and table II respectively. We set the cluster assignments to the number of class/diagnosis for each dataset, i.e. 3 for ADNI (diagnosis label) and 6 (NHY score) for PPMI. We want the model to identify the individual disease stages both when there is only limited knowledge known to a certain disease, i.e class/diagnosis is not available and when diagnosis label is available, and thus provide insightful and interpretable information to help discover corresponding treatment to individual treatment. We compare our proposed method with the



(a) Bilstm (b) StageNet (c) Dipole (d) Ours



(e) Bilstm (f) StageNet (g) Dipole (h) Ours

Fig. 3: Visualization of the clusters for ADNI (first row) and PPMI (second row) using PCA: Bilstm (1st column), StageNet (2nd column), Dipole (3rd column), Ours (4th) column).

TABLE IV: Most significant features in each cluster measured by first order gradient for ADNI and PPMI dataset.

ADNI Dataset				
	Features			
Cluster I	RAVLT_learning	Ventricles	WholeBrain	ICV
	RAVLT_perc_forgetting	RAVLT_forgetting	ADAS13	RAVLT_immediate
Cluster II	ICV	RAVLT_perc_forgetting	ADAS13	Ventricles
	serial	RAVLT_immediate	CDRSB	
Cluster III	RAVLT_perc_forgetting	serial	ICV	RAVLT_learning
	Entorhinal	Hippocampus	Ventricles	WholeBrain
PPMI Dataset				
	Features			
Cluster I	Global Spontaneity of Movement	Speech	Anxious Mood	Arising from Chair
	Right leg	Getting Out of Bed	Pronation-Supination (left)	
Cluster II	Posture	Rest tremor amplitude	Dopamine	Rigidity
	Saliva + Drooling	Anxious Mood	Global Spontaneity of Movement	
Cluster III	Postural Stability	Cognitive Impairment	Rest Tremor Amplitude	Pronation-Supination (left)
	Dopamine	Standing	Rigidity	
Cluster IV	Pronation-Supination (left)	Standing	Postural Stability	Chewing
	Cognitive Impairment	Dopamine	Right Hand	
Cluster V	Dopamine	Cognitive Impairment	Hallucinations	Chewing
	Dressing	Pronation-Supination (left)	Arising from Chair	
Cluster VI	Rigidity	Serial	Rigidity	Standing
	Apathy	Constipation Problems	Cognitive Impairment	Dopamine

TABLE V: Complexity comparison between models

Model	# of trainable parameters
Dipole	279k
StageNet	283k
AC-TPC	143k
TC-EMNet^{-u}	163k
TC-EMNet^{-s}	174k

aforementioned baselines in terms of clustering performance. It is clear that our method has demonstrated competitive performance against all baseline methods across all evaluation metrics for both datasets. We note that it is generally difficult to identify clusters without the presence of label information as indicated by low NMI and RI scores. However, TC-EMNet outperforms baseline by a large margin in terms of NMI and RI scores when clustering with label. Training under supervised setting yields significantly better clustering performance compared to training under unsupervised setting. This is due to fact that the correlation between diagnosis and input features is encoded into each hidden representation. Although AC-TPC has better performance in terms of RI on the PPMI dataset. The method relies on pre-training the model with over 1000 epochs, which could result in the model memorizing the input data. Both Dipole and StageNet have comparable performance. However, it is worth mentioning that both models have leveraged attention over multi-layer

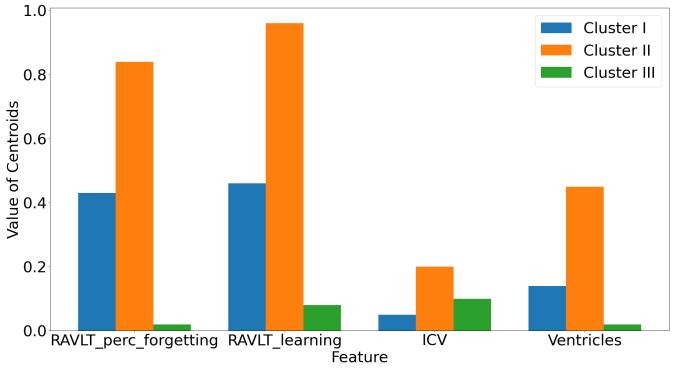


Fig. 4: Significant feature values of cluster centroids on ADNI dataset. The distribution of clusters is very different, which means distinct subtypes.

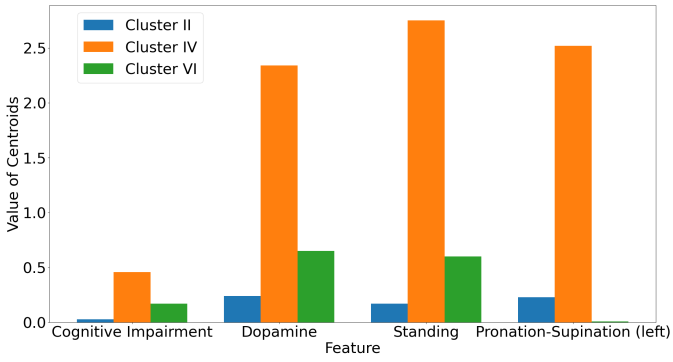


Fig. 5: Significant feature values of cluster centroids on PPMI dataset. The distribution of clusters is very different, which means distinct subtypes.

RNNs, which introduces additional complexity to the model. A detailed comparison between the trainable parameters is shown in table V. Furthermore, we find that when training with label information, RI score can be negatively impacted compared to training without labels. Such phenomena are observed for multiple baseline methods. One explanation could be directly leveraging label information overwhelms the training process since labels possess strong prediction power compared to input features, making the model more biased towards dominated class when dealing with imbalanced datasets; thus, RI may drop as there are more false positives and false negatives. It also can be observed that leveraging external memory effectively captures long-term information and the TC-EMNet is capable of learning complexity from the input data. The patient-level memory network constructively binds with the global-level memory network to produce more comprehensive memory representations.

B. Disease Stage

In order to interpret the disease stages and progression patterns found by TC-EMNet. We first selected three baseline models that have comparable performance against TC-EMNet and visualized the hidden representations in $2D$ space using PCA [37]. The results are shown in Fig 3. We observed that in general most methods can produce distinct clusters for the ADNI dataset. However, for PPMI dataset, most baseline methods failed at producing effective clusters, whereas TC-EMNet produces distinct clustering results. This shows that TC-EMNet is able to constructively model long-term information between each visit in order to find effective representations. Next, we compute feature importance for every cluster based on the weights from the last layer of the network. The results are shown in table IV. It can be observed that for both datasets, each cluster is determined by a diverse range of features, which means it is easier to identify each patients' progression patterns through observation. We also compute the centroid values for each cluster and plot the distribution in Fig 4, 5 for ADNI and PPMI datasets respectively. For ADNI dataset, our proposed model has determined significant features such as RAVLT_learning, RAVLT_perc_forgetting, ICV, ventricles. Rey's Auditory Verbal Learning Test (RAVLT) scores are helpful in testing episodic memories and are very important indicators in identifying a patient's progression in Alzheimer's disease [38]. In particular, the learning test (RAVLT_learning) and percent forgetting test (RAVLT_perc_forgetting) are highly correlated and thus become crucial biomarkers for early detection in AD. It can be observed in Fig 4 that three clusters produced by our model have wide distribution for RAVLT testing values, which suggests three different patient subtypes. As for PPMI dataset, our model has found that the dopamine dysregulation syndrome (Dopamine) is a significant feature in identifying clusters. Studies have discovered that under clinical settings early characterization of Dopamine can aid the treatment for motor and non-motor complications for Parkinson's disease [39]. There are also studies that showed that cognitive impairment (Cognitive impairment) is a strong indicator for Parkinson's disease. Difference in cognitive impairment scores can reflect advanced progression in PD [40].

VI. Conclusion

In this paper, we propose TC-EMNet for disease progression modeling on time-series data. TC-EMNet leverages VAE to model data irregularity and an external memory network to capture long-term dependency. We developed TC-EMNet to perform patient clustering/subtyping under both supervised and unsupervised settings. Under supervised setting, TC-EMNet leverages a dual memory network architecture to extract target-aware information from diagnosis to compute patient representations. Throughout the experiment on two real-world datasets, we showed that our model outperforms state-of-the-art methods and is able to identify interpretable disease stages that are clinically meaningful. TC-EMNet yields competitive clustering performance with limited complexity. In the real-world clinical setting, we hope that our model could help physicians identify patients' progression patterns and discover potential disease stages to gain more understanding about chronic and other heterogeneous diseases.

ACKNOWLEDGMENT

This paper was funded in part by the National Science Foundation under award number CBET-2037398.

REFERENCES

[1] A. A. Kehagia, R. A. Barker, and T. W. Robbins, "Neuropsychological and clinical heterogeneity of cognitive impairment and dementia in patients with parkinson's disease," *The Lancet Neurology*, vol. 9, no. 12, pp. 1200–1213, 2010.

[2] M. Ferrer, J. Alonso, J. Morera, R. M. Marrades, A. Khalaf, M. C. Aguar, V. Plaza, L. Prieto, and J. M. Anto, "Chronic obstructive pulmonary disease stage and health-related quality of life," *Annals of internal Medicine*, vol. 127, no. 12, pp. 1072–1079, 1997.

[3] S. Auer and B. Reisberg, "The gds/fast staging system," *International Psychogeriatrics*, vol. 9, no. S1, pp. 167–171, 1997.

[4] C. Lee and M. Van Der Schaar, "Temporal phenotyping using deep predictive clustering of disease progression," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5767–5777.

[5] J. Gao, C. Xiao, Y. Wang, W. Tang, L. M. Glass, and J. Sun, "Stagenet: Stage-aware neural networks for health risk prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 530–540.

[6] T. Ma, C. Xiao, and F. Wang, "Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction," in *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 2018, pp. 261–269.

[7] Z. Sun, S. Ghosh, Y. Li, Y. Cheng, A. Mohan, C. Sampaio, and J. Hu, "A probabilistic disease progression modeling approach and its application to integrated huntington's disease observational data," *JAMIA open*, vol. 2, no. 1, pp. 123–130, 2019.

[8] X. Zhang, J. Chou, J. Liang, C. Xiao, Y. Zhao, H. Sarva, C. Henchcliffe, and F. Wang, "Data-driven subtyping of parkinson's disease using longitudinal clinical records: a cohort study," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[9] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 85–94.

[10] V. Fortuin, M. Hüser, F. Locatello, H. Strathmann, and G. Rätsch, "Somvae: Interpretable discrete representation learning on time series," *arXiv preprint arXiv:1806.02199*, 2018.

[11] L. Mou, P. Zhao, H. Xie, and Y. Chen, "T-lstm: A long short-term memory neural network enhanced by temporal information for traffic flow prediction," *Ieee Access*, vol. 7, pp. 98 053–98 060, 2019.

[12] A. M. Alaa and M. van der Schaar, "Attentive state-space modeling of disease progression," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 338–11 348.

[13] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1195–1206, 2018.

[14] X. Teng, S. Pei, and Y.-R. Lin, "Stocast: Stochastic disease forecasting with progression uncertainty," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 3, pp. 850–861, 2020.

[15] J. M. Dennis, B. M. Shields, W. E. Henley, A. G. Jones, and A. T. Hattersley, "Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data," *The lancet Diabetes & endocrinology*, vol. 7, no. 6, pp. 442–451, 2019.

[16] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 65–74.

[17] C. Yin, R. Liu, D. Zhang, and P. Zhang, "Identifying sepsis subphenotypes via time-aware multi-modal auto-encoder," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 862–872.

[18] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[19] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[20] J.-T. Kuo and K.-T. Chien, "Variational recurrent neural networks for speech separation." INTERSPEECH, 2017.

[21] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2017.

[22] E. Jun, A. W. Mulyadi, and H.-I. Suk, "Stochastic imputation and uncertainty-aware attention to ehr for mortality prediction," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–7.

[23] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," *arXiv preprint arXiv:1503.08895*, 2015.

[24] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot," *Sensors*, vol. 17, no. 9, p. 1967, 2017.

[25] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *Icml*, vol. 1, 2001, pp. 577–584.

[26] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.

[27] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury *et al.*, "The parkinson progression marker initiative (ppmi)," *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[28] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, no. 3. Makuhari, 2010, pp. 1045–1048.

[29] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[30] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *arXiv preprint arXiv:1608.05745*, 2016.

[31] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1903–1911.

[32] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1945–1954.

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[34] Y. Zhao, Z. Qiao, C. Xiao, L. Glass, and J. Sun, "Pyhealth: A python library for health predictive models," *arXiv preprint arXiv:2101.04209*, 2021.

[35] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[36] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[37] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[38] E. Moradi, I. Hallikainen, T. Hänninen, J. Tohka, A. D. N. Initiative *et al.*, "Rey's auditory verbal learning test scores can be predicted from whole brain mri in alzheimer's disease," *NeuroImage: Clinical*, vol. 13, pp. 415–427, 2017.

[39] A. H. Evans and A. J. Lees, "Dopamine dysregulation syndrome in parkinson's disease," *Current opinion in neurology*, vol. 17, no. 4, pp. 393–398, 2004.

[40] D. Verbaan, J. Marinus, M. Visser, S. M. van Rooden, A. M. Stiggelbout, H. A. Middelkoop, and J. J. van Hilten, "Cognitive impairment in parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 78, no. 11, pp. 1182–1187, 2007.