StoCast: Stochastic Disease Forecasting with Progression Uncertainty

Xian Teng, Sen Pei, and Yu-Ru Lin†
for the Alzheimer's Disease Neuroimaging Initiative*

***Abstract*— Forecasting patients' disease progressions with rich longitudinal clinical data has drawn much attention in recent years due to its impactful application in healthcare and the medical field. Researchers have tackled this problem by leveraging traditional machine learning, statistical techniques and deep learning based models. However, existing methods suffer from either deterministic internal structures or over-simplified stochastic components, failing to deal with complex uncertain scenarios such as progression uncertainty (i.e., multiple possible trajectories) and data uncertainty (i.e., imprecise observations and misdiagnosis). To overcome these major uncertainty issues, we propose a novel deep generative model, Stochastic Disease Forecasting Model (StoCast), along with an associated neural network architecture StoCastNet, that can be trained efficiently via stochastic optimization techniques. Our StoCast model uses internal stochastic components to deal with departures of observed data from patients' true health states, and more importantly, is able to produce a comprehensive estimate of future disease progression trajectories. Based on two public datasets related to Alzheimer's disease and Parkinson's disease, we demonstrate how our StoCast model achieves robust and superior performance than deterministic baseline approaches, and conveys richer information that can potentially assist doctors to make decisions with greater confidence in a complex uncertain scenario.**

***Index Terms*— Disease Forecasting, Deep Generative Models, Progression Uncertainty, Neural Networks**

I. Introduction

Thanks to the rapid development of modern healthcare systems, Electronic Health Records (EHR) have been extensively used in smart healthcare applications. EHR data contain longitudinal health information, such as clinical tests, cognitive assessments, medication and procedures, allowing for tracking patient health status at each specific time throughout their medical history. With such rich EHR data, one important question is how to model patients' disease progressions and to effectively forewarn their future health states, so that early interventions may be undertaken to cope with chronic illness.

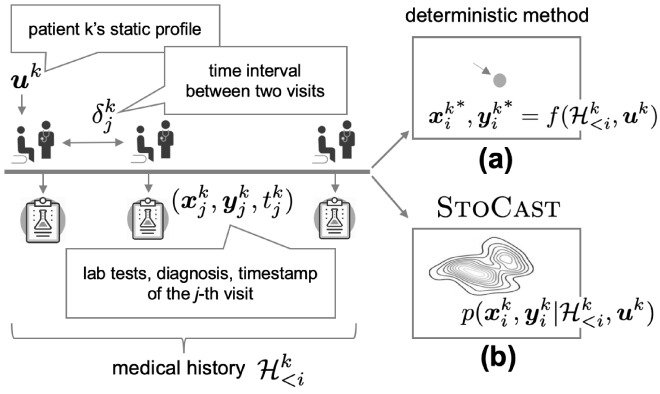

Fig. 1. A schematic illustration for predicting disease progression given a patient's prior information (e.g., patient profile and historic medical records). (a) Deterministic methods typically learn a parametric function that maps prior information to a unimodal outcome with an optimal point estimate, which fails to express a complex uncertain scenario where the output is multimodal and is often unable to inform a decision with confidence. (b) We propose StoCast that learn a generative model to predict the distribution over multiple possible outcomes to effectively capture the disease progression.

Until recent years, most of the techniques for disease prediction are traditional machine learning and statistical techniques. One paradigm of prior works have tackled this question by formulating a regression or classification problem[1], and the key idea can be illustrated in Fig. 1(a). From a probabilistic perspective, those models attempt to learn a *deterministic* parametric function that maps historic records into future outcomes through maximum likelihood – minimizing a loss function that captures the distance between the predictions and the observations. Examples in this paradigm include Wang et al. [1] and Xu et al. [2], where regression models are used to predict multiple cognitive scores from neuroimaging features

Manuscript received on October 14, 2019; revised on March 1, 2020 and May 12, 2020; Accepted on June 23, 2020. This work was funded in part by the grants from the PICSO Lab, including NSF Grants #1739413 and #2027713, DARPA and AFOSR awards. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

X. Teng is with School of Computing and Information, University of Pittsburgh, Pennsylvanian, PA 15213 USA (e-mail: xian.teng@pitt.edu).

S. Pei is with Mailman School of Public Health, Columbia University, New York, NY 10032 USA. (e-mail: sp3449@cumc.columbia.edu).

†Corresponding: Y.-R. Lin is with School of Computing and Information, University of Pittsburgh, Pennsylvanian, PA 15213 USA (e-mail: yurulin@pitt.edu).

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[1]We particularly focus on those classification models that learn a parametric function to estimate the conditional class probabilities.

for early recognition of Alzheimer's disease. Deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been introduced to predict disease progression [3]–[9]. These neural networks are designed to recognize a patient's sequential patterns and use temporal patterns to predict future probable scenarios such as diagnosis and prescriptions. Generally, these methods have advantages in modeling long-term temporal dependency and learning distributed representations; however they are still deterministic in nature since a specific mapping function is usually learned in the training process. Another line of effort is using statistical techniques to model the temporal progression of diseases [10]–[15]. A straightforward approach is to utilize Hidden Markov Model (HMM) to capture disease state transition and to predict state progression. Even though these approaches contain internally stochastic units (i.e., Markov chains over hidden variables), they often make strong assumptions about data generation process, have simplified discrete hidden state and are expensive to compute.

We argue that it is difficult to directly apply those prior approaches to address complex uncertain scenarios often seen in chronic and progressive diseases. We characterize two challenging aspects. (i) **Progression uncertainty**: multiple outcomes are possible as the diseases progress, in other words, the space of plausible outputs is *multimodal*. Take Parkinson's disease (PD) for example, it is generally characterized by different stages, ranging from mild to the most severe. However, ambiguities exit in determining the clinical stages due to the heterogeneity of symptoms and patient conditions, and thus multiple outcomes are possible. An internally deterministic model is unable to deal with such uncertainty because it assumes a unimodal output space and the source of uncertainty simply comes from the local conditional output distribution. A Markov chain based model might also fail this challenge due to its over-simplified internal stochastic structure. (ii) **Data uncertainty**: patient medical records are often high-dimensional and sometimes subject to errors (e.g., clinical assessments with errors, misdiagnosis), raising acute concern in such complicated progressive disease. Prior deterministic approaches that assume a Gaussian conditional distribution to account for measurement noises, or that presume all patients are correctly diagnosed, lack robustness to tolerate outliers [16], thus yielding undesired results.

Problem. Given these uncertainty issues, we argue that a more relevant and challenging question to ask is: *what is the probability distribution of a patient's health trajectory in the future?* That is, we seek to provide a comprehensive ensemble of future progression possibilities (Fig. 1 (b)), rather than extrapolating a single point estimate.

Method. To answer this, we propose a novel method, called Stochastic Disease Forecasting Model (STOCAST). Motivated from a generative perspective, we assume that future data is produced through a two-step generative procedure that involves an intermediate latent variable (see Fig. 2). The latent variable can specifically address the uncertainty challenge by acting as a stochastic bridge – its prior distribution conditioned on currently available information is used to express internal stochasticity, and a generation procedure conditioned on the latent variable is used to produce an ensemble of forecasts. On the other hand, the generative procedure can address the issue of data uncertainty, as the model allows potential departures of observed data from patients' true health states (or health "manifold") by generating output from a distribution conditioned on a latent variable. The objective for learning such a stochastic model is generally intractable – to overcome this, we leverage variational Bayesian approach and reformulate a *tractable* variational objective. Furthermore, we utilize reparameterization strategy to obtain an unbiased Monte Carlo estimator of the variational objective, which can be optimized efficiently by stochastic optimization techniques.

Neural Network Structure. We introduce a new neural network structure, called STOCASTNET, based on our STOCAST model. We use neural networks as a way to model it because it possesses powerful capabilities including *nonlinearity, long-term dependency, distributed representation and easy to be trained*. Specifically, the nonlinearity property enables us to learn complex nonlinear mapping functions, the long-term dependency is particularly useful since we need to leverage patients' past information to do prediction at a present time, the distributed representation is central for summarizing patients' health data into rich compact vectors, and the training can be done through stochastic gradient descent as neural networks are typically designed to be differentiable. The STOCASTNET, comprising three major components – a prior network, a generation network and a posterior network, is differentiable everywhere; therefore it can be trained end-to-end via stochastic optimization techniques.

The main contributions of this paper include:

- We formulate the problem of disease progression prediction from a novel generative perspective to account for progression uncertainty. Rather than producing a single point prediction under the unimodal assumption, we attempt to approximate the overall distribution of future disease progressions.
- We propose a deep generative model, called STOCAST, to solve the above problem. In contrary to deterministic approaches, our model consists of internal stochastic components, which makes it able to handle progression uncertainty, and robust to data distortion.
- We provide a neural network STOCASTNET based on the proposed model that can be trained efficiently end-to-end using stochastic optimization techniques.
- We conduct a set of comprehensive experiments on two benchmark datasets - Alzheimer's Disease Neuroimaging Initiative (ADNI) data and Parkinson's Progression Markers Initiative (PPMI) data. Our results demonstrate that STOCAST is able to achieve robust and superior performance compared to deterministic baselines approaches.

This paper is organized as follows. Section II reviews related works; Section III presents problem formulation; Section IV describes our proposed approach and technical details; Section V provides data descriptions and experimental results, and Section VI concludes this paper.

II. Related Work

We first briefly review existing literature for the task of disease progression forecasting, with special focus on this work's significance by accounting for uncertainty. We then introduce some basic knowledge about generative models and variational autoencoder.

A. Disease Progression Forecasting

The accumulation of EHR data has triggered great efforts of researchers in disease prediction [17]–[22]. Many existing works formulate the challenge of disease progression prediction as a regression or classification problem [1], [2], [23]. For example, to help identify Alzheimer's disease (AD) at an early stage, Xu et al. [2] propose a low-rank structured sparse regression model to foresee patients' cognitive scores based on current neuroimaging features, Wang et al. [1] develop a nonlinear Multi-layer Multi-target Regression (MMR) to achieve a similar goal. Recent years have witnessed the success of deep learning in various domains (e.g., public health and social crisis [24]–[26]), researchers have been applying these techniques to address disease prediction issues [5], [27]–[29]. Those methods' effectiveness is often attributed to the ability of neural networks to learn nonlinear and distribution representation of data, as well as to capture long-term dependency in sequences. For instance, Doctor AI [7] is a RNNs-based approach that assesses medical history of a patient to predict the next visit time as well as subsequent diagnosis. DeepCare [28] is built upon Long Short Term Memory (LSTM) units, meanwhile incorporating additional temporal decay and attention mechanisms to account for temporal irregularity and importance variation in hospital visits. Nevertheless, a major limitation in those methods is the deterministic internal structure, lacking considerations of progression uncertainties in disease forecasting. Besides, there have been active research in modeling the temporal disease progressions using machine learning and statistical techniques [10]–[12], [30]. For example, Wang et al [10] build an unsupervised probabilistic model that has a Markov Jump Process to characterize continuous-time disease state transitions and a set of Markov chains to capture the relations between disease states and comorbidity onsets. Jackson et al [11] develop a multistate Hidden Markov Model (HMM) to estimate disease state transition rates. Xiao et al [12] modifies HMM restricted by demographic data to model patient health trajectories. Unfortunately, such methods are typically limited by linear state transitions, over-simplified discrete hidden states and computational scalability [7]. In this paper, we solve the disease prediction task with consideration of progression uncertainty, and propose a novel StoCast model that contains internal stochasticity to approximate the distribution of future health states. Meanwhile, our model is build on neural networks, thus inherits most powerful capabilities of deep learning techniques.

B. Generative Models and Variational Autoencoder

Generative modeling is one type of unsupervised learning that deals with complicated data distributions. It could be interpreted as learning a generative process by which the observation data arose [31]. That implies, if we had learned a representative generative model M for a set of data points $\boldsymbol{x}$ distributed according to some unknown distribution $p(\boldsymbol{x})$, we can draw new samples from the model to obtain a distribution $p_M(\boldsymbol{x})$ that is similar to the true distribution. Training generative models, particularly for complicated high-dimensional data, is a challenging task: it might require strong assumptions about data, or have to adopt computationally expensive inference process like Markov Chain Monte Carlo. Recently, some progress has been made by leveraging neural networks into training generative models. One of the most popular deep generative networks is Variational Autoencoder (VAE) [32]. It has weak assumptions about generative process of data, and can be trained through stochastic optimization techniques in an efficient way. An extension of VAE is Conditional Variational Autoencoder (CVAE), which takes additional knowledge as extra inputs and builds the generative process conditioned on such inputs [33]. The proposed StoCast model is inspired by CVAE in the way that it learns a generative model conditioned on patients' medical history and profile information to forecast a set of future predictions.

III. Research Problem

We use a running example to explain our research problem. As shown in Fig. 1, a patient k has visited hospital from time to time irregularly. During each hospital visit at a certain date, his clinical data shall be collected, such as lab tests and symptoms (i.e., *features*) and diagnosis (i.e., *labels*). Given the sequence of his medical records, we might be curious about: what are the possible health progressions for patient k in the near future? Is his health condition getting better or worse? To formally define our research problem, we let $\mathcal{H}^k = \{(\boldsymbol{x}_j^k, \boldsymbol{y}_j^k, t_j^k) | j = 1, ..., T^k\}$ be the sequence of hospital visits for k, where $\boldsymbol{x}_j^k$ represents the feature vector, $\boldsymbol{y}_j^k$ represents the diagnosis vector, t_j^k is the timestamp of the j-th visit, and T^k is the total number of visits. The irregular time interval between two consecutive visits is denoted by $\delta_j^k = t_j^k - t_{j-1}^k$. Additionally, the patient k's profile information, such as demographic data, family disease history and gene information, is represented by a profile vector $\boldsymbol{u}^k$.

Problem. *Given a population of patients, denoted as $\mathcal{K}$, with data $\{(\boldsymbol{u}^k, \mathcal{H}^k) | k \in \mathcal{K}\}$ that comprise both static profiles and longitudinal clinical observations $\mathcal{H}_{<i}^k = \{(\boldsymbol{x}_j^k, \boldsymbol{y}_j^k, t_j^k) | j < i\}$, our question is: What is the overall distribution of the patient's possible health states at a future time point t_i^k? In other words, the goal is to predict, for each patient k, the distribution of possible health states at a future time point t_i^k (often the next hospital visit), i.e., $p(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k | \mathcal{H}_{<i}^k, \boldsymbol{u}^k)$.*

Unlike existing approaches that seek to find an optimal point $(\boldsymbol{x}_i^{k*}, \boldsymbol{y}_i^{k*})$ in the space of future health states, our proposed question requires to estimate the overall distribution of all possible of future health states $p(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k | \mathcal{H}_{<i}^k, \boldsymbol{u}^k)$. Such distribution conveys richer information that helps doctors to make decisions with greater confidence in a complex uncertain scenario, such as distribution modality (e.g., unimodal or

multimodal), the most likely future health states, and the extent of the credible regions for an estimate.

IV. Method

This section presents our proposed method, namely **Sto**chastic Disease Fore**cast**ing Model (**StoCast**).

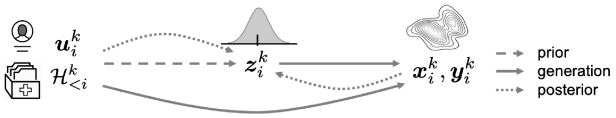

Fig. 2. A schematic illustration of the proposed StoCast approach. (i) The future observation $(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k)$ is generated according to a generative procedure corresponding to a particular latent variable $(\boldsymbol{z}_i^k)$. (ii) The latent variable $(\boldsymbol{z}_i^k)$ is drawn from a prior distribution conditioned on the previously available information $(\boldsymbol{\mathcal{H}}_{<i}^k, \boldsymbol{u}^k)$. (iii) The distribution of the latent variable is inferred from previously available information $(\boldsymbol{\mathcal{H}}_{<i}^k, \boldsymbol{u}^k)$ and newly observed data.

A. Two-Step Generative Procedure

In order to tackle the progression and data uncertainty issues, we approach the problem from a generative perspective. As shown in Fig. 2, we model the generative procedure of future health data conditioned on currently available information through two steps: (i) a latent variable $\boldsymbol{z}_i^k$ is drawn from the prior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{z}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k)$ conditioned on current knowledge $(\mathcal{H}_{<i}^k, \boldsymbol{u}^k)$ (dashed arrow); (ii) the output is generated by sampling from the distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k|\boldsymbol{z}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)$ conditioned on both current knowledge $(\mathcal{H}_{<i}^k, \boldsymbol{u}^k)$ and the latent variable $\boldsymbol{z}_i^k$ (line arrow). The $\boldsymbol{\theta}$ represents generative parameters. In our scenario, the health-related data could be of high-dimensional and involve complicated dependencies between the dimensions, while latent variable $\boldsymbol{z}_i^k$ lies in a hidden subspace of much lower dimensionality than that of the data space. The latent variable $\boldsymbol{z}_i^k$ is introduced to allow a generally complicated distribution (over observed data) to be constructed through a simpler conditional distribution. Its role is two-folds: (a) it captures the internal stochasticity from the data, through which the model can produce an ensemble of predictions as possible outcomes at a future time point; (b) the output generated at the second step, based on the distribution conditioned on the latent variable, offers greater flexibility in modeling the departure of data from the true health states.

B. (Variational) Objective Function

Given the two-step generative procedure, we shall be able to learn the parameters $\boldsymbol{\theta}$ using maximum likelihood estimation (MLE):

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \sum_k \sum_i \log \int \Big[\underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k|\boldsymbol{z}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)}_{\text{generation}} \times \underbrace{p_{\boldsymbol{\theta}}(\boldsymbol{z}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k)}_{\text{prior}} \Big] d\boldsymbol{z}_i^k. \tag{1}$$

However, the marginalization of $\boldsymbol{z}_i^k$ is generally intractable for complicated prior and generation functions (e.g., those described by neural networks). A widely-adopted strategy dealing with such intractability is variational Bayesian method, i.e., to derive a variational lower bound to approximate the logarithm of the marginal probability of the observation. The derivation is as follows:

$$\begin{aligned} &\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k) \\ &= \log \left(\mathbb{E}_{q_{\boldsymbol{\phi}}} \left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \boldsymbol{z}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^k|\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)} \right] \right) \\ &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \boldsymbol{z}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k) \right] \\ &- \mathbb{E}_{q_{\boldsymbol{\phi}}} \Big[\log \underbrace{q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^k|\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)}_{\text{posterior}} \Big] \triangleq \mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi}), \end{aligned} \tag{2}$$

where the inequality can be obtained using Jensen's inequality. Here $\mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi})$ is a single variational lower bound term for patient k at time t_i^k. Particularly, the lower bound $\mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi})$ involves a new probability $q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^k|\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)$ expressed by parameter $\boldsymbol{\phi}$. It is the variational distribution, or proxy posterior, introduced to approximate the intractable true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}_i^k|\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)$. This proxy posterior is associated with the posterior process in Fig. 2 (dotted arrow), that the distribution of latent variable can be inferred by combining available information and newly observed data. It follows that the KL divergence of the proxy posterior from the true posterior distribution is equal to the the difference between the original log likelihood and the variational lower bound:

$$\begin{aligned} &\mathrm{KL}\Big[q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^k|\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k) \| p_{\boldsymbol{\theta}}(\boldsymbol{z}_i^k|\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k) \Big] \\ &= \mathbb{E}_{q_{\boldsymbol{\phi}}} \Big[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^k|\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k) \Big] \\ &- \mathbb{E}_{q_{\boldsymbol{\phi}}} \Big[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \boldsymbol{z}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k) \Big] \\ &+ \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k) \\ &= -\mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k|\mathcal{H}_{<i}^k, \boldsymbol{u}^k). \end{aligned} \tag{3}$$

Equation (2) holds if and only if $q_{\boldsymbol{\phi}}$ is equal to the true posterior distribution, i.e., their KL divergence is zero. Thus, we leverage a *tractable* proxy posterior $q_{\boldsymbol{\phi}}$ to approximate the *intractable* true posterior, so as to rewrite the intractable objective function in (1) into a tractable variational objective function in (2). Then, the objective can be re-expressed as:

$$\boldsymbol{\theta}, \boldsymbol{\phi} = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \sum_k \sum_i \mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{4}$$

C. Solving Variational Objective Function

We present our solution to the optimization problem expressed in (4). In particular, we shall differentiate and optimize the variational objective function with respect to both parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$. However, according to (2), the lower bound $\mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi})$ contains two terms: the first term is the expected log joint probability, and the second term is the entropy of the proxy posterior. Both terms depends on $q_{\boldsymbol{\phi}}$, making the gradient with respect to $\boldsymbol{\phi}$ a little problematic. Following the prior works [32], [34], we employ reparameterization to obtain Monte Carlo gradients of $\mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi})$ with respect to both parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$. **Reparameterization and Monte Carlo Gradients.** The reparameterization trick expresses the latent variable $\boldsymbol{z}_i^k$ as an

invertible function of another set of random variables $\boldsymbol{\varepsilon}$ that does not depend on parameter $\boldsymbol{\phi}$, i.e., $\boldsymbol{z}_i^k = g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \bullet)$, where we use the black dot $\bullet$ to indicate all conditional elements $\{\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k\}$. In this way, the expectation with respect to the proxy posterior $q_{\boldsymbol{\phi}}$ of any function of $\boldsymbol{z}_i^k$, denoted as $f(\boldsymbol{z}_i^k)$, can be expressed as:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}}\big[f(\boldsymbol{z}_i^k)\big] = \mathbb{E}_{p(\boldsymbol{\varepsilon})}\big[f\big(g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \bullet)\big)\big] \simeq \frac{1}{L}\sum_l f\big(g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(l)}, \bullet)\big), \tag{5}$$

where we can randomly draw L samples $\boldsymbol{\varepsilon}^{(l)} \sim p(\boldsymbol{\varepsilon})$ to approximate the true expectation value. Accordingly, the gradient with respect to $\boldsymbol{\phi}$ can be pushed into the expectation, yielding:

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\mathbb{E}_{q_{\boldsymbol{\phi}}}\big[f(\boldsymbol{z}_i^k)\big] &= \mathbb{E}_{p(\boldsymbol{\varepsilon})}\big[\nabla_{\boldsymbol{\phi}} f\big(g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}, \bullet)\big)\big] \\ &\simeq \frac{1}{L}\sum_l \nabla_{\boldsymbol{\phi}} f\big(g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(l)}, \bullet)\big).\end{aligned} \tag{6}$$

Following (5) and substituting $f(\boldsymbol{z}_i^k)$ by the terms inside the expectation in $\mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi})$, we obtain an unbiased Monte Carlo estimator as follows,

$$\begin{aligned}\tilde{\mathcal{L}}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{L}\sum_l \Big[&\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \boldsymbol{z}_i^{k(l)} | \mathcal{H}_{<i}^k, \boldsymbol{u}^k) \\ &- \log q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^{k(l)} | \boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)\Big], \\ &\text{where } \boldsymbol{z}_i^{k(l)} = g_{\boldsymbol{\phi}}(\boldsymbol{\varepsilon}^{(l)}, \bullet),\ \boldsymbol{\varepsilon}^{(l)} \sim p(\boldsymbol{\varepsilon}).\end{aligned} \tag{7}$$

Similarly, we can push the gradient operator $\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}$ into expectation to obtain Monte Carlo estimates of the gradients with respect to $\boldsymbol{\theta}, \boldsymbol{\phi}$, as:

$$\begin{aligned}\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\tilde{\mathcal{L}}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{L}\sum_l \Big[&\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \boldsymbol{z}_i^{k(l)} | \mathcal{H}_{<i}^k, \boldsymbol{u}^k) \\ &- \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^{k(l)} | \boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)\Big].\end{aligned} \tag{8}$$

In this way, we are able to use traditional stochastic optimization (e.g., standard gradient ascent or Adagrad) to update the parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ until their convergence. It has been shown that the Monte Carlo gradients obtained via reparameterization exhibit relatively low variance [34], and typically only one sample (i.e., $L = 1$) is needed to estimate a noisy gradient, making the algorithms very efficient [32], [34].

D. StoCast *Neural Network:* StoCastNet

So far, we have presented our solution to maximize the variational objective function $\mathcal{L}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi})$ by stochastic optimization techniques, here we present StoCastNet, a novel neural network architecture based on our StoCast framework. We decompose the unbiased Monte Carlo estimator in 7 into four parts, yielding:

$$\begin{aligned}\tilde{\mathcal{L}}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{L}\sum_l \Big[&\underbrace{\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k | \boldsymbol{z}_i^{k(l)}, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)}_{①} \\ &+ \underbrace{\log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i^k | \boldsymbol{z}_i^{k(l)}, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)}_{②} + \underbrace{\log p_{\boldsymbol{\theta}}(\boldsymbol{z}_i^{k(l)} | \mathcal{H}_{<i}^k, \boldsymbol{u}^k)}_{③} \\ &- \underbrace{\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^{k(l)} | \boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \mathcal{H}_{<i}^k, \boldsymbol{u}^k)}_{④}\Big].\end{aligned} \tag{9}$$

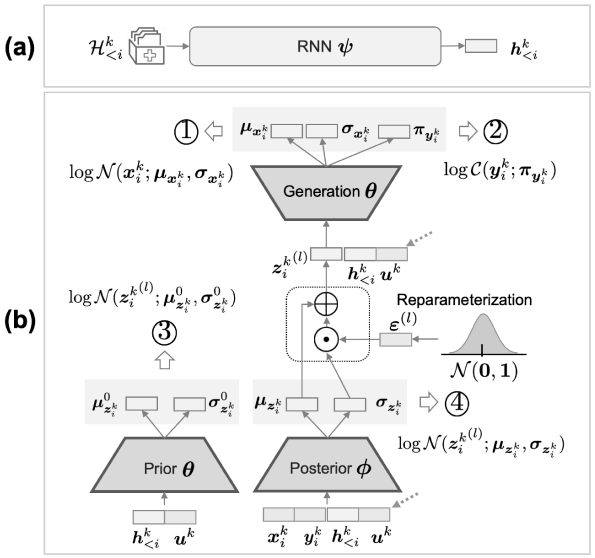

Fig. 3. Training-time StoCastNet structure. (a) An auxiliary RNN summarizing history data into a representation. (b) Three major components in StoCast structure: (i) prior, (ii) generation, (iii) posterior. A reparameterization layer in between posterior and generation is used to guarantee data flow as continuous.

Fig. 3(b) illustrates the overall training-time StoCastNet architecture that produces the above four parts as outputs from three networks - prior, generative, posterior. We additionally introduce an auxiliary RNN network parameterized by $\boldsymbol{\psi}$ that maps a patient's medical history $\mathcal{H}_{<i}^k$ into a summary representation $\boldsymbol{h}_{<i}^k$ (Fig. 3(a)). Besides, we restrict the latent variable to be a multivariate Gaussian distribution, thus its posterior should also follow the same distribution. We further assume the output distribution of features to be Gaussian, and the distribution of discrete diagnosis Categorical. Below, we discuss three main networks: prior, generation and posterior.

Prior Network. The prior network should construct a multivariate Gaussian distribution over the latent variable conditioned on history and profile data. Therefore, it takes the concatenation of history representation and profile vector as input, and outputs two prior parameter vectors, i.e., the mean vector as well as the standard deviation vector for prior Gaussian distribution (Fig. 3(b)):

$$\boldsymbol{\mu}_{\boldsymbol{z}_i^k}^0 = \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\text{prior}}([\boldsymbol{h}_{<i}^k, \boldsymbol{u}^k]), \boldsymbol{\sigma}_{\boldsymbol{z}_i^k}^0 = \boldsymbol{\sigma}_{\boldsymbol{\theta}}^{\text{prior}}([\boldsymbol{h}_{<i}^k, \boldsymbol{u}^k]), \tag{10}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\text{prior}}$ and $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{\text{prior}}$ are two deterministic functions built by feed-forward neural networks with parameter $\boldsymbol{\theta}$.

Posterior Network. The proxy posterior $q_{\boldsymbol{\phi}}$ is used to approximate the intractable true posterior of latent variable. Thus the posterior network shall take existing knowledge and also newly observed data as input, and outputs two posterior parameter vectors, i.e., the mean vector and the standard deviation vector for posterior Gaussian distribution (Fig. 3(b)):

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{z}_i^k} &= \boldsymbol{\mu}_{\boldsymbol{\phi}}^{\text{post}}([\boldsymbol{h}_{<i}^k, \boldsymbol{u}^k, \boldsymbol{x}_i^k, \boldsymbol{y}_i^k]),\\ \boldsymbol{\sigma}_{\boldsymbol{z}_i^k} &= \boldsymbol{\sigma}_{\boldsymbol{\phi}}^{\text{post}}([\boldsymbol{h}_{<i}^k, \boldsymbol{u}^k, \boldsymbol{x}_i^k, \boldsymbol{y}_i^k]),\end{aligned} \tag{11}$$

where $\boldsymbol{\mu}_{\boldsymbol{\phi}}^{\text{post}}$ and $\boldsymbol{\sigma}_{\boldsymbol{\phi}}^{\text{post}}$ are posterior functions described by feed-forward neural networks parameterized by $\boldsymbol{\phi}$. If we directly draw samples $\boldsymbol{z}_i^{k(l)}$ from posterior Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}_i^k}, \boldsymbol{\sigma}_{\boldsymbol{z}_i^k})$, which is a non-continuous operation and thus has no gradient, it would result in an unwanted scenario that we cannot back-propagate errors through this layer. This is exactly where reparameterization should come into play. As shown in Fig. 3(b), we first draw samples for the auxiliary variable $\boldsymbol{\varepsilon}^{(l)}$ from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$, and then rely on the following reparameterization transformation to obtain a set of samples,

$$\boldsymbol{z}_i^{k(l)} = \boldsymbol{\mu}_{\boldsymbol{z}_i^k} + \boldsymbol{\sigma}_{\boldsymbol{z}_i^k} \odot \boldsymbol{\varepsilon}^{(l)}, \text{ where } \boldsymbol{\varepsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{1}), \tag{12}$$

where $\odot$ indicates element-wise multiplication. In this way, we can guarantees smooth data flow from the posterior network to the generation network.

Generation Network. Given $\boldsymbol{z}_i^{k(l)}$ samples, we concatenate it with medical history and profile vector to generate features and diagnosis. The outputs comprise three parameter vectors, i.e., the mean vector and the standard deviation vector Gaussian distribution, and the $\boldsymbol{\pi}_{\boldsymbol{y}_i^k}$ for Categorical distribution (Fig. 3(b)), yielding

$$\boldsymbol{\mu}_{\boldsymbol{x}_i^k} = \boldsymbol{\mu}_{\boldsymbol{\theta}}^{\text{gen}}([\boldsymbol{z}_i^{k(l)}, \boldsymbol{h}_{<i}^k, \boldsymbol{u}^k]), \boldsymbol{\sigma}_{\boldsymbol{x}_i^k} = \boldsymbol{\sigma}_{\boldsymbol{\theta}}^{\text{gen}}([\boldsymbol{z}_i^{k(l)}, \boldsymbol{h}_{<i}^k, \boldsymbol{u}^k]),$$
$$\boldsymbol{\pi}_{\boldsymbol{y}_i^k} = \boldsymbol{\pi}_{\boldsymbol{\theta}}^{\text{gen}}([\boldsymbol{z}_i^{k(l)}, \boldsymbol{h}_{<i}^k, \boldsymbol{u}^k]),$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{\text{gen}}$, $\boldsymbol{\sigma}_{\boldsymbol{\theta}}^{\text{gen}}$ and $\boldsymbol{\pi}_{\boldsymbol{\theta}}^{\text{gen}}$ are generative functions described by feed-forward neural networks with parameter $\boldsymbol{\theta}$. Putting all together, the StoCastNet in Fig. 3(b) takes the inputs $\{\boldsymbol{\varepsilon}^{(l)}, \mathcal{H}_{<i}^k, \boldsymbol{u}^k, \boldsymbol{x}_i^k, \boldsymbol{y}_i^k\}$ and produces the four outputs $\{①-④\}$ corresponding to the four parts in (4):

$$\begin{aligned}
&① : \log \mathcal{N}(\boldsymbol{x}_i^k; \boldsymbol{\mu}_{\boldsymbol{x}_i^k}, \boldsymbol{\sigma}_{\boldsymbol{x}_i^k}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k | \boldsymbol{z}_i^{k(l)}, \boldsymbol{h}_{<i}^k, \boldsymbol{u}^k),\\
&② : \log \mathcal{C}(\boldsymbol{y}_i^k; \boldsymbol{\pi}_{\boldsymbol{y}_i^k}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i^k | \boldsymbol{z}_i^{k(l)}, \boldsymbol{h}_{<i}^k, \boldsymbol{u}^k),\\
&③ : \log \mathcal{N}(\boldsymbol{z}_i^{k(l)}; \boldsymbol{\mu}_{\boldsymbol{z}_i^k}^0, \boldsymbol{\sigma}_{\boldsymbol{z}_i^k}^0) = \log p_{\boldsymbol{\theta}}(\boldsymbol{z}_i^{k(l)} | \boldsymbol{h}_{<i}^k, \boldsymbol{u}^k),\\
&④ : \log \mathcal{N}(\boldsymbol{z}_i^{k(l)}; \boldsymbol{\mu}_{\boldsymbol{z}_i^k}, \boldsymbol{\sigma}_{\boldsymbol{z}_i^k}) = \log q_{\boldsymbol{\phi}}(\boldsymbol{z}_i^{k(l)} | \boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \boldsymbol{h}_{<i}^k, \boldsymbol{u}^k).
\end{aligned} \tag{13}$$

We note that the entire StoCastNet is continuous and differentiable, thus it can be trained end-to-end using stochastic optimization techniques. We present the minibatch training process in **Algorithm 1**. Specifically, in each iteration, we can estimate the Monte Carlo gradients in terms of all population $\mathcal{K}$, denoted as $\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi})$, using the randomly sampled minibatch of patients $\mathcal{K}'$, yielding:

$$\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \simeq \frac{\sum_{k\in\mathcal{K}} T^k}{\sum_{k\in\mathcal{K}'} T^k} \sum_{k\in\mathcal{K}'} \sum_i \nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \tilde{\mathcal{L}}_i^k(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{14}$$

Algorithm 1 Minibatch training of the StoCast network.

1: Initialize parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$
2: **repeat**
3: Draw a batch of patients $\mathcal{K}'$ from all population $\mathcal{K}$
4: Draw a set of samples $\boldsymbol{\varepsilon}^{(l)}$ from $\mathcal{N}(\mathbf{0}, \mathbf{1})$
5: Compute Monte Carlo gradients $\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}} \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ following (14)
6: Update $\boldsymbol{\theta}, \boldsymbol{\phi}$ using the above gradients
7: **until** convergence of parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$

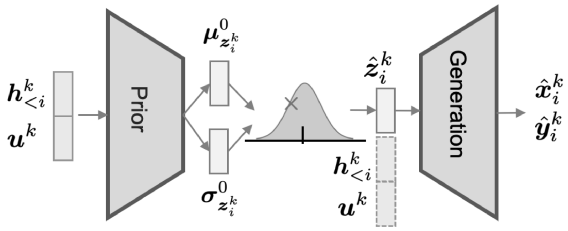

Fig. 4. The forecasting procedure based on trained StoCastNet.

E. Forecasting based on Trained Network

After the StoCastNet has been trained, we want to use it to predict the distribution of a patient's future health condition. Here we note that, although we cannot give an explicitly analytical expression in terms of the distribution, we can produce a set of forecasts to approximate the distribution $\hat{p}(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k | \mathcal{H}_{<i}^k, \boldsymbol{u}^k)$. As shown in Fig. 4, at prediction time we only need the trained prior and generation networks to produce forecasts $\{\hat{\boldsymbol{x}}_i^k, \hat{\boldsymbol{y}}_i^k\}$, corresponding to the two-step generative procedure discussed in Fig. 2. We use the prior network to get the mean and standard deviation vectors, then draw a set of samples $\hat{\boldsymbol{z}}_i^k$, and finally use the generation network to produce a set of forecasts $(\hat{\boldsymbol{x}}_i^k, \hat{\boldsymbol{y}}_i^k)$.

V. Experiments

We evaluate our proposed models using two public datasets from Alzheimer's Disease Neuroimaging Initiative (ADNI) database[2], and Parkinson's Progression Markers Initiative (PPMI) database[3].

A. Datasets

1) ANDI Dataset: Alzheimer's disease (AD) is a chronic neurodegenerative disease that usually causes problems with memory, thinking and behaviors. ADNI is a longitudinal multi-center study that aims to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of AD. The study has three phases[4]: an initial 5-year study from 2004 (ADNI-1), a 2-year extended study from 2009 (ADNI-GO), and a 5-year study started from 2011 (ADNI-2). New participants were recruited during each phase

[2]www.adni-info.org
[3]www.ppmi-info.org
[4]According to the new updates, the 4th phase has started from 2016, but our experiment does not use data from this phase.

of the study, and they are followed and reassessed over time to track the pathology of the disease as it progresses. ADNI data contain a rich set of heterogeneous features, including demographics, clinical assessments, cognitive scores, genomic, neuroimaging biomarkers and biospecimen. In addition, it also includes diagnosis labels assigned by doctors, including Control Normal (CN), Mild Cognitive Impairment (MCI) and AD. The three levels of diagnosis indicate how severe a patient's AD symptoms have progressed. A merged ADNI 1/GO/2 data package, called "ADNIMERGE", has been developed which is downloadble from ADNI data archive. It loads all ADNI data (except genetic data), documentation, and analysis vignettes[5]. Our experiments depend on a unified dataset in this package, called "adnimerge", that contains a diversity of commonly used variables. Table I lists the detailed ADNI features we have used in our experiment.

2) PPMI Dataset.: Parkinson's disease (PD) is a long-term degenerative disorder of the central nervous system that mainly affects motor system. PPMI is a 5-year landmark observational clinical study aimed at comprehensively evaluating cohorts of significant interest (e.g., patients with PD, people with high risk, and those who are healthy) using advanced imaging, biologic sampling and clinical and behavioral assessments to identify biomarkers of Parkinson's disease progression. The PPMI study takes places at clinical sites sites throughout the Unites States, Europe, Israel and Australia, and have collected data in a standardized manner under strict protocols developed by the steering committee. Because PPMI data do not provide per-visit diagnosis, we consider Hoehn and Yahr (NHY) score as a proxy label in our experiments. NHY is a widely used system for the purpose of describing how the symptoms of PD progress, with discrete scores ranging from 0 to 5. We recode score 0 into Control Normal (CN) label, score 1 into Unilateral Involvement (UI) label representing minimal or no functional disability (movement disorder is limited to one side of the body), and scores 2-5 into Bilateral Involvement (BI) label corresponding to severe PD symptoms (movement disorder affects both sides of the body). Table II lists the explanations of diagnosis labels.

3) Data Preprocessing: As patients might be given different tests in different hospital visits, there are missing values at the feature level. To prepare the data for further experiment, we first discard sparse features with a missing rate larger than 50% (since imputation might introduce undesirable bias), and then exclude patient sequences that contain no more than three hospital visits. Then we employ different imputation strategies to fill in missing data found in diagnosis as well as features: (a) Considering that the two diseases are irreversible progressive brain disorders, we do **diagnosis imputation** following the procedure: (i) if a patient's last diagnosis is the same as the next diagnosis, we replace the current missing point with the diagnosis; (ii) if a patient has already been diagnosed to be AD/PD, we carry forward such diagnosis and replace missing data thereafter; (iii) if a patient's first observed diagnosis is healthy, we carry backward such diagnosis and replace missing data prior to this visit. (b) For **feature imputation**, we employ the last occurrence carried forward and mean imputation. Specifically, if a missing record occurs in a patient's follow-up hospital visit, we impute it with the most recently observed value from the patient's history; if the records of a specific feature in a patient's medical history are all missing, we choose to use the mean value of that feature calculated from the cohort of patients having same diagnosis to impute it. Table III reports a series of data statistics after preprocessing.

B. Baseline Methods

We compare our STOCAST with three state-of-the-art deep learning approaches in healthcare domain, as well as three widely-used classifiers.

- **Doctor AI** [7] is a temporal predictive model built on recurrent neural networks (RNNs). Given longitudinal time stamped EHR data, it is able to predict the diagnosis and medication codes for a subsequent visit. In our experiment, we implement this baseline based on LSTM units.
- **T-LSTM** [9], also called "Time-Aware LSTM", is a LSTM networks augmented with a temporal decay mechanism to handle irregular time intervals in longitudinal patient records. In particular, the memory cell is decomposed into short- and long-term memories and the former one is adjusted in a way that longer the elapsed time, the smaller the effect of the previous memory to the current output.
- **RETAIN** [8] is an interpretable predictive model for healthcare based on reverse time attention mechanism. It learns to allocate attentions to individual hospital visits and clinical variables, so as to interpret importance of these factors in prediction task.
- **Logistic Regression (LR)** is a discriminative probabilistic model that uses a logistic (or softmax) function to model the class probabilities given feature variables. Due to its simplicity and effectiveness, LR has been widely applied in a diversity of domains.
- **Decision Tree (D-Tree)** is a commonly used non-parametric machine learning technique. In classification, it makes sequential and hierarchical decisions about the outcomes variable based on input data.
- **K-Nearest Neighbors (KNN)** is also a non-paremetric method that attempt to classify a data point by a plurality vote of its k nearest neighbors, with the data point being assigned to the class most common along its neighbors.

In particular, the deep learning based approaches take the entire longitudinal medical records along with patient profile as inputs and are trained by minimizing the distance between prediction and observation (i.e., continuous features and discrete labels). In contrast, the classifiers cannot capture long-term temporal dependency in disease modeling, therefore we only consider the current hospital visit along with patient profile data as inputs. Since the outputs of these classification methods are restricted to be discrete labels, we ignore the prediction of features in the training process of such classifiers.

[5]Referring to https://adni.bitbucket.io/index.html for more information

TABLE I
DIAGNOSIS LABELS AND FEATURE CATEGORIES OF ADNI DATA AND PPMI DATA IN OUR EXPERIMENT.

ADNI Data	
Category	**Features**
Profile	Age, gender, education, race, marital state
Neuropsychological	AD Assessment Scale (ADAS) ADAS Delayed Word Recall (ADASQ4) Mini-Mental State Examination (MMSE) Clinical Dementia Rating Scale (CDR) Everyday Cognition Participant Self Report Everyday Cognition Subject Partner Report Rey's Auditory Verbal Learning Test Logical Memory-Delayed Recall Montreal Cognitive Assessment (MoCA) Digit Symbol Substitution Preclinical Alzheimer's Cognitive Composite
Imaging-related	FDG-PET, PIB SUVR, MRI measurements
Biospecimen	Genetics, CSF biomarkers

PPMI Data	
Category	**Features**
Profile	Age, gender, race, family history, education
Non-motor	Benton Judgment of Line Orientation Test Hopkins Verbal Learning Test Letter Number Sequencing Test Montreal Cognitive Assessment (MoCA) Symbol Digit Modalities Test Geriatric Depression Scale The Questionnaire for Impulsive-Compulsive Disorders in Parkinson's Disease (QUIP) State-Trait Anxiety Inventory (STAI)) Scales For Outcomes In PD (SCOPA-AUT) UPenn Smell Identification Test (UPSIT) REM Sleep Behavior Disorder Questionnaire Epworth Sleepiness Scale (ESS))
Motor	Unified PD Rating Scale (MDS-UPDRS) Daily Living Scale (ADL) Physical Activity Scale for the Elderly (PASE)
Biospecimen	CSF biomarkers, Genetics

TABLE II
DIAGNOSIS EXPLANATION.

Diagnosis	**Explanation**
CN	Control normal
MCI	Mild cognitive impairment in AD
AD	Alzheimer's disease
UI	Unilateral involvement (PD affects one side of the body)
BI	Bilateral Involvement (PD affects both sides)

TABLE III
DATA STATISTICS AFTER PREPROCESSING.

Statistics	**ADNI**	**PPMI**
Number of subjects	1,574	1,093
Number of total visits	11,474	9,421
Feature/label dimension	42/3	56/3
Mean/max sequence length	7.29/19 visits	8.62/17 visits
Mean time interval	5.96 months	5.27 months
Label imbalance	32%,42%,26%	27%,20%,53%

C. Evaluation Metrics

In realistic scenario, people are more concerned about a patient's diagnosis in the future, the high-level indicator of health condition. Therefore, we will particularly focus on the prediction results of diagnosis, and examine the performances of distinct approaches. Here we stress that baseline methods output single point predictions, whereas our STOCAST would produce an ensemble of forecasts for each test case (100 predictions in our experiments). To facilitate a fair comparison of STOCAST and baselines, we replicate the ground truth label for 100 times to match it with each prediction point of STOCAST. In this way, some commonly used classification metrics can be applied in our scenario, including accuracy, precision, recall, F1 score and Area Under the Receiver Operating Characteristic Curve (ROC AUC). In particular, they evaluate different aspects of a model: accuracy is calculated as the fraction of correct classifications with respect to the total test cases, precision measures the ability of a model to identify only the relevant data points, recall assesses the ability of a model to find all the relevant cases within a dataset, F1 score can be interpreted as a weighted harmonic mean of the precision and recall, and ROC AUC is a performance measurement for classification problem at various threshold settings. As seen in Table III, the three labels in our datasets are not balanced, i.e., CN: 32%, MCI: 42%, AD: 26% in ADNI data, and CN: 27%, UI: 20%, BI: 53% in PPMI data. Therefore, under such imbalanced multiclass circumstance, we calculate a micro-average value and a weighted-average value for each of these metrics, including precision, recall, F1 and AUC. A micro-average will aggregate the contributions of all classes to compute the average metric, whereas a weighted-average will compute the metric independently for each class and then take the average by accounting for class imbalance. Besides, we also compute a single ROC AUC value for each of the classes.

D. Experimental Setting

Neural network structure. In our implementation of STOCAST, all mapping functions - Gaussian means and standard deviations - in the prior, posterior and generative are three-layered feed-forward neural networks. The standard deviations have softplus output layer to ensure non-negativity. Besides, we note that the deep generative models used in our work have "latent variable collapse" problem, i.e., the posterior is very close to the prior thus does not actually listen to the input data, making it unable to learn a faithful representation of observation data. To avoid such pitfall, we implement a "skip" version of STOCASTNET by adding connections that attach latent variables to multiple layers in the generation network [35], forcing the generation network to maintain a strong connection between latent variables and observation data. In terms of deep learning based baselines, we construct

TABLE IV
PERFORMANCE COMPARISON IN DISEASE FORECASTING EVALUATED ON DIAGNOSIS USING ADNI DATA.

Methods	Accuracy	ROC AUC					Precision		Recall		F1	
		CN	MCI	AD	Weighted	Micro	Weighted	Micro	Weighted	Micro	Weighted	Micro
StoCast	**0.9163**	**0.9913**	**0.9625**	**0.9804**	**0.9764**	**0.9808**	**0.9164**	**0.9163**	**0.9163**	**0.9163**	**0.9162**	**0.9163**
Doctor AI	0.8645	0.9757	0.9304	0.9791	0.9577	0.9648	0.8676	0.8645	0.8645	0.8645	0.8651	0.8645
RETAIN	0.8360	0.9576	0.9042	0.9766	0.9403	0.9488	0.8365	0.8360	0.8360	0.8360	0.8362	0.8360
T-LSTM	0.8458	0.9662	0.9153	0.9801	0.9486	0.9574	0.8460	0.8458	0.8458	0.8458	0.8458	0.8458
LR	0.8589	0.9760	0.9313	0.9800	0.9587	0.9643	0.8590	0.8589	0.8589	0.8589	0.8578	0.8589
D-Tree	0.7973	0.8741	0.7917	0.8847	0.8426	0.8482	0.7969	0.7973	0.7973	0.7973	0.7970	0.7973
KNN	0.7659	0.8960	0.8169	0.9485	0.8768	0.8898	0.7680	0.7659	0.7659	0.7659	0.7653	0.7659

TABLE V
PERFORMANCE COMPARISON IN DISEASE FORECASTING EVALUATED ON DIAGNOSIS USING PPMI DATA.

Methods	Accuracy	ROC AUC					Precision		Recall		F1	
		CN	UI	BI	Weighted	Micro	Weighted	Micro	Weighted	Micro	Weighted	Micro
StoCast	**0.7968**	**0.9876**	**0.8054**	**0.8975**	**0.9030**	**0.9385**	**0.7759**	**0.7968**	**0.7968**	**0.7968**	**0.7779**	**0.7968**
Doctor AI	0.7725	0.9807	0.7253	0.8769	0.8739	0.8913	0.6260	0.7725	0.7725	0.7725	0.6877	0.7725
RETAIN	0.7590	0.9775	0.7434	0.8657	0.8708	0.8962	0.7341	0.7590	0.7590	0.7590	0.7162	0.7590
T-LSTM	0.7672	0.9791	0.6762	0.8623	0.8558	0.8700	0.6219	0.7672	0.7672	0.7672	0.6829	0.7672
LR	0.7178	0.9733	0.7631	0.8787	0.8806	0.8957	0.7507	0.7178	0.7178	0.7178	0.7279	0.7178
D-Tree	0.6722	0.8911	0.5580	0.7153	0.7305	0.7552	0.6720	0.6722	0.6722	0.6722	0.6721	0.6722
KNN	0.6946	0.9436	0.6286	0.8083	0.8081	0.8510	0.6651	0.6946	0.6946	0.6946	0.6744	0.6946

output layers to generate Gaussian mean and standard deviation vectors for feature prediction, and Categorical parameter vector for diagnosis prediction.

Training details[6] The data is shuffled and splitted into training and validation sets, then we train the model on training and keep track of loss function on validation. The training process will be stopped early if there are 10 consecutive steps not showing any loss reduction. Both StoCast and deep learning baselines are implemented by *tensorflow*[7], using the same Adam optimizer [36] with the same configurations, i.e., the dimension of latent variable is 32; the dimension of hidden state is 32; the batch size is 32; and the total number of epochs is 150. The traditional classifiers are implemented by *scikit-learn*[8] package, where we use optimized parameters:"newton-cg" optimization solver for logistic regression, the number of neighbors is 5 for KNN.

E. Performance Evaluation

1) Performance in Next-Step Prediction.: Tables IV and V report the performance comparison of StoCast and baselines in the next-visit prediction evaluated by different metrics for ADNI data and PPMI data, respectively. The best performance is marked in bold. We can see that our StoCast outperforms baseline approaches across distinct metrics for both datasets, suggesting that our StoCast method is a promising and robust approach in disease forecasting. In addition, we can see that deep learning methods achieves relatively better performance than traditional classifiers, potentially due to their capabilities such as distributed representation, long-term dependency and non-linearity. Our StoCast naturally inherits such capabilities as it is built based on neural networks, meanwhile it is able to capture internal uncertainty in disease evolution, making it a more effective approach.

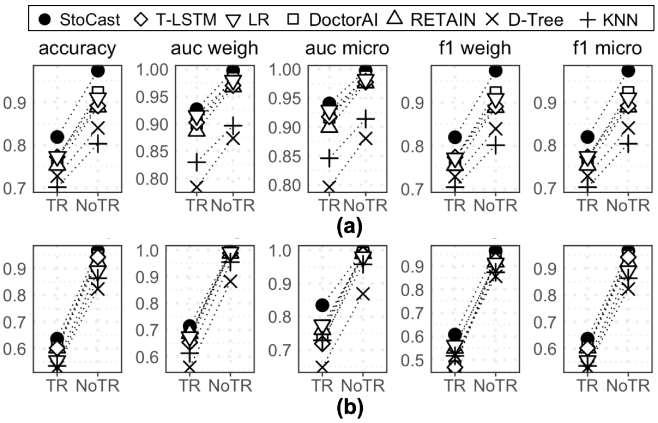

Fig. 5. Performance comparison on two groups of patients within (a) ADNI and (b) PPMI. Performance is evaluated on diagnosis. TR and NoTR are for the subset of patients with diagnosis transitions and without any transitions. The y-axis indicates performance score of different metrics.

2) Performance on Different User Groups.: To provide a comprehensive comparison for different approaches, we examine the performance of StoCast on different user groups within each of the two datasets. We divide the population into two groups – a TR group corresponding to the patients who have exhibited diagnosis transitions (approximately 33% for ADNI, and 32% for PPMI), and a NoTR group associated with the people who do not shown any diagnosis transitions. Such investigation can help understand the robustness of our model. Fig. 5 shows that our StoCast outperforms baselines

[6]Source code downloadable at https://github.com/picsolab/StoCast
[7]https://www.tensorflow.org/
[8]https://scikit-learn.org/stable/

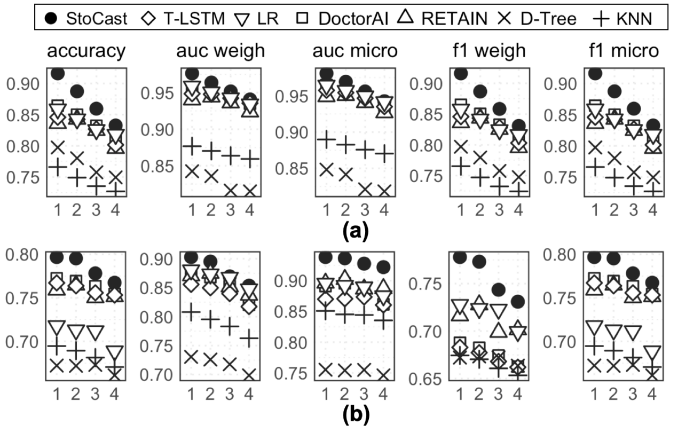

Fig. 6. Performance comparison in multi-step forward prediction for (a) ADNI and (b) PPMI. Performance is evaluated on diagnosis. The x-axis indicates forward step d (1-4), and y-axis indicates performance score.

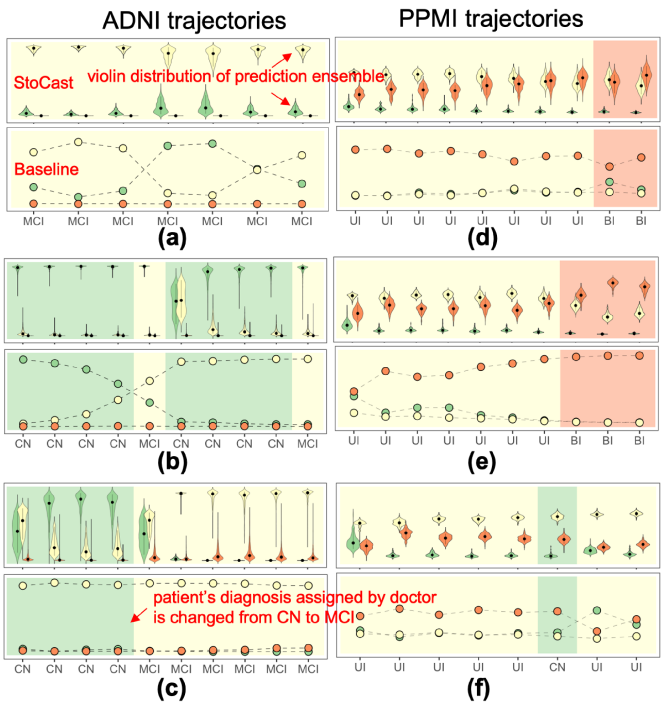

Fig. 7. Examples of patient disease trajectories described by diagnosis labels from (a,b,c) ADNI and (d,e,f) PPMI (better to view in color). The x-axis is longitudinal diagnosis assigned by doctors (background color also indicates doctor's diagnosis), and y-axis is predicted probabilities of different labels. Upper row (violins) shows StoCast's prediction ensemble and lower row (points) shows baseline's single-point prediction. The black dot in violins represents median and the thin line represents 5%-95% percentile.

on both TR and NoTR groups evaluated by different metrics for both ADNI and PPMI datasets. We note that, compared to the NoTR group, i.e., those with stable health condition, the prediction task for the TR group, i.e., those with unstable health conditions, is generally more challenging – in terms of obtaining better scores. The results demonstrate that our StoCast is able to effectively and robustly capture both the stable and unstable disease trajectories, potentially as it listens to observation data and also maintains certain internal stochasticity.

3) Performance in Multi-Step Forward Prediction.: We further examine whether StoCast can maintain such superior performance if we conduct multi-step forward prediction tasks - to predict a patient's diagnosis in the next d-th visit. Fig. 6 shows the performance of different methods as a function of forward step d assessed by different metrics. It can be seen that StoCast can still maintain its superior position in the near future, validating the effectiveness of our method. The decreasing trend in StoCast's performance implies that disease forecasting becomes harder as we foresee a farther future. Here we note that baseline approaches exhibits some fluctuations in their performances tested on PPMI data. It can be explained by their biased tendency towards the prevailing label (i.e., BI) in the dataset. We will offer more concrete interpretations by showing examples in the next subsection.

F. Qualitative Examination

Here we show qualitative results to provide a detailed case study about the difference between StoCast and baseline method (taking Doctor AI as an example). Fig. 7 displays three patients' disease trajectories per dataset, i.e., the plots (a-c) are ADNI examples, and (d-f) are PPMI examples. The x-axis records a patient's longitudinal diagnostic labels assigned by doctors in each visit, and the y-axis indicates the predicted probabilities for different labels. We use violin plots to describes the distributions of the ensemble of predictions outputted by StoCast, and use individual markers to indicate the baseline's single-point predictions. Hues and background colors indicate the diagnosis labels assigned by doctors.

Fig. 7(a) corresponds to a patient who has a relatively stable MCI state, but the baseline incorrectly predicts that this patient would recover to CN state in the 4th and 5th visit. In contrast, our StoCast gives correct predictions at every step, and even successfully detect the fluctuations of this patient's health states in the 4th and 5th visits, as shown in the shape variations of violin plots. Fig. 7(b) plots a normal person's trajectory who has certain risk of developing AD as doctors assigned MCI in the 5th and 10th visits. For most of the visits, our StoCast makes correct decisions confidently (indicated by the shape of CN's violins - they are densely clustered around 1.0), except for the 5th visit after doctors assigning a MCI label, where the violins of CN and MCI are stretched and overlap heavily. We point out that this is the critical time that a confident forecast decision is difficult to make, as the doctor has changed the diagnosis from MCI back to CN in the subsequent five visits. In contrary, the baseline starts giving a series of wrong predictions – it keeps predicting MCI in subsequent visits. Fig. 7(c) is a patient who exhibits a degenerating health trend towards developing AD. Our StoCast method is able to closely capture such deterioration inclination (as CN's violin is moving downwards and MCI's violin is rising upwards), but baseline's outputs maintain to be a constant MCI state, failing to forewarn that this patient's health is deteriorating.

We can obtain similar observations in Fig. 7(d-f) based on PPMI data, among which two patients exhibit gradual transition from UI to BI (Fig. 7 (d), (e)), and one maintains a

relatively stable UI state (Fig. 7 (f)). Based on the results in Fig. 7 (d), (e), we see that our STOCAST is able to well capture the two patients' health trends from UI to BI, as BI's violin is rising upwards while UI's is going downwards. By contrast, the baseline produces incorrect BI predictions long before when doctors give such BI diagnosis. Similar to Fig. 7(b), the example in Fig. 7(f) contains an "unexpected" diagnosis CN in the 6th visit after several visits with UI diagnosis. It is probability a misdiagnosis or at least an ambiguous diagnosis. The two examples in Fig. 7(b)(f) demonstrate that our approach is more robust in handling data uncertainty whereas deterministic baseline fail to tolerant such potential misdiagnosis. Overall, we find that baseline approach exhibits a tendency to output BI label across all hospital visits in PPMI data, contrary to doctors' diagnosis. This might be because BI accounts for the majority of diagnosis labels in the data (approximately 53% as shown in Table III), therefore the baseline is trained to be biased towards BI label due to such imbalanced distribution. This phenomenon can to some extent explain our observation in Fig. 6(b) that baselines display fluctuations in their performances in multi-step forward prediction task. As we increases d (farther prediction), a biased prediction of BI might become a "good" prediction because patients would gradually enter into the most severe BI stage. However, such biased prediction results are not what we expect in realistic application. Therefore, susceptibility to data imbalance is another disadvantage that prevents deterministic baselines from achieving good prediction performances. In contrary, STOCAST demonstrates a superior prediction capability: (i) it can successfully identify patients' disease progressions through the smooth movements of violin plots overtime; (ii) it can offer insights regards the level of difficulty in prediction tasks through the relative positions of violin plots for distinct labels (i.e., the closer the harder), and the level of forecast confidence through the shapes of violins (i.e., the more compact the more confident).

VI. DISCUSSION

We developed a novel generative model named STOCAST to address the problem of disease prediction formulated in an uncertain context that considers progression uncertainty and data uncertainty. Application of this method to two longitudinal clinical datasets - ADNI and PPMI - shows that it achieves superior and robust performance across different scenarios (e.g., different sub-populations and multi-step forward predictions) assessed by various evaluation metrics.

This work has some limitations. First, it demands a complete longitudinal data which unfortunately might not be met in reality: missingness can take different types (e.g., missing at random, missing systematically) due to many reasons (e.g., patients drop out studies, data is collected improperly). Without consideration of missing data types, our preprocessing procedure might introduce bias or affect the representativeness of the results. To address this issue, future research is needed to develop methods that take full account of incomplete data without relying heavily on data preprocessing or imputation. Second, our analysis are based on two datasets that are processed through sophisticated study design and standardized data acquisition and quality control, which might not be representative for the poor quality problem in many real health data. Besides, we only examine the effectiveness of our method based on two particular progressive irreversible brain diseases. One future extension would be to test the generalizability of this method on a broader range of diseases (e.g., diseases with recurrent states, with a large number of potential diagnosis labels) and a variety of noisy data. Third, this research adopted traditional evaluation metrics to assess the performance of disease prediction in an uncertain context. Related studies used the minimum distance of the closest forecast to the ground truth [37], or estimated the likelihood of ground truth within the predicted distribution [38]. However, both criteria are flawed because the former unfairly selects the best guess among all predictions while the latter is hard to be applied to the case of single point prediction. Therefore, we call for the research community to address this open challenge in designing better criteria to compare methods with stochastic nature to deterministic algorithms.

Among the related research of disease prediction using ADNI/PPMI data [1], [2], [39], [40], the major contribution of this research is its focus on addressing progression uncertainty and data uncertainty, advocating further efforts towards this direction. Our work has important clinical implications as it provides richer information for doctors to make decisions with greater confidence in a complex uncertain scenario, which is critical in offering patients earlier and more tailored treatments to defer their health deterioration.

ACKNOWLEDGMENT

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic

Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners could be found at www.ppmi-info.org/fundingpartners.

REFERENCES

[1] X. Wang, X. Zhen, Q. Li, D. Shen, and H. Huang, "Cognitive assessment prediction in Alzheimer's disease by multi-layer multi-target regression," *Neuroinformatics*, pp. 1–10, 2018.

[2] J. Xu, C. Deng, X. Gao, D. Shen, and H. Huang, "Predicting Alzheimer's disease cognitive assessment via robust low-rank structured sparse model," in *IJCAI*, vol. 2017. NIH Public Access, 2017, p. 3880.

[3] 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 J Biomed Health Inform*, vol. 22, no. 5, pp. 1589–1604, 2017.

[4] 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 *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*. ACM, 2017, pp. 1903–1911.

[5] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk prediction on electronic health records with prior medical knowledge," in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1910–1919.

[6] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: a convolutional net for medical records," *IEEE J Biomed Health Inform*, vol. 21, no. 1, pp. 22–30, 2016.

[7] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, 2016, pp. 301–318.

[8] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *NeurIPS*, 2016, pp. 3504–3512.

[9] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*. ACM, 2017, pp. 65–74.

[10] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*. ACM, 2014, pp. 85–94.

[11] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto, "Multistate Markov models for disease progression with classification error," *J. Royal Stat. Soc: Series D (The Statistician)*, vol. 52, no. 2, pp. 193–209, 2003.

[12] H. Xiao, J. Gao, L. Vu, and D. S. Turaga, "Learning temporal state of diabetes patients via combining behavioral and demographic data," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining*. ACM, 2017, pp. 2081–2089.

[13] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-Markov models," *J. Mach. Learn. Res*, vol. 14, no. Feb, pp. 673–701, 2013.

[14] J. M. Lange, R. A. Hubbard, L. Y. Inoue, and V. N. Minin, "A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data," *Biometrics*, vol. 71, no. 1, pp. 90–101, 2015.

[15] R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman, "Disease progression modeling using hidden Markov models," in *2012 Annu. Int. Conf. of IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 2845–2848.

[16] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, 2006.

[17] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *2017 IEEE Int. Conf. Data Mining*. IEEE, 2017, pp. 787–792.

[18] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, A. Zhang, and J. Gao, "Personalized disease prediction using a CNN-based similarity learning method," in *2017 IEEE Int. Conf. Bioinformatics and Biomedicine*. IEEE, 2017, pp. 811–816.

[19] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive Hawkes process," in *2015 IEEE Int. Conf. Data Mining*. IEEE, 2015, pp. 721–726.

[20] Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang, "Deep patient similarity learning for personalized healthcare," *IEEE Trans Nanobioscience*, vol. 17, no. 3, pp. 219–227, 2018.

[21] Y. Yuan, G. Xun, F. Ma, Q. Suo, H. Xue, K. Jia, and A. Zhang, "A novel channel-aware attention framework for multi-channel eeg seizure detection via multi-view deep learning," in *2018 IEEE EMBS Int. Conf. Biomedical and Health Informatics*. IEEE, 2018, pp. 206–209.

[22] S. Minhas, A. Khanum, F. Riaz, S. A. Khan, and A. Alvi, "Predicting progression from mild cognitive impairment to Alzheimer's disease using autoregressive modelling of longitudinal and multimodal biomarkers," *IEEE J Biomed Health Inform*, vol. 22, no. 3, pp. 818–825, 2017.

[23] L. Nie, L. Zhang, L. Meng, X. Song, X. Chang, and X. Li, "Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 28, no. 7, pp. 1508–1519, 2016.

[24] A. M. Ertugrul, Y.-R. Lin, and T. Taskaya-Temizel, "Castnet: Community-attentive spatio-temporal networks for opioid overdose forecasting," in *ECML PKDD*. Springer, 2019, pp. 432–448.

[25] X. Teng, M. Yan, A. M. Ertugrul, and Y.-R. Lin, "Deep into hypersphere: Robust and unsupervised anomaly discovery in dynamic networks," in *IJCAI*, 2018.

[26] X. Teng, Y.-R. Lin, and X. Wen, "Anomaly detection in dynamic networks using multi-view time-series hypersphere learning," in *Proc. 2017 ACM Conf. Information & Knowledge Management*, 2017, pp. 827–836.

[27] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J Biomed Health Inform*, vol. 21, no. 1, pp. 4–21, 2016.

[28] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "DeepCare: A deep dynamic memory model for predictive medicine," in *Pacific-Asia Conf. Knowledge Discovery & Data Mining*. Springer, 2016, pp. 30–41.

[29] K. Zheng, W. Wang, J. Gao, K. Y. Ngiam, B. C. Ooi, and W. L. J. Yip, "Capturing feature-level irregularity in disease progression modeling," in *Proc. 2017 ACM Conf. Information & Knowledge Management*. ACM, 2017, pp. 1579–1588.

[30] K. P. Exarchos, T. P. Exarchos, C. V. Bourantas, M. I. Papafaklis, K. K. Naka, L. K. Michalis, O. Parodi, and D. I. Fotiadis, "Prediction of coronary atherosclerosis progression using dynamic bayesian networks," in *2013 35th Annu. Int. Conf. of IEEE Engineering in Medicine and Biology Society*. IEEE, 2013, pp. 3889–3892.

[31] N. M. Nasrabadi, "Pattern recognition and machine learning," *J Electron Imaging*, vol. 16, no. 4, p. 049901, 2007.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[33] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *NeurIPS*, 2015, pp. 3483–3491.

[34] F. R. Ruiz, M. T. R. AUEB, and D. Blei, "The generalized reparameterization gradient," in *NeurIPS*, 2016, pp. 460–468.

[35] A. B. Dieng, Y. Kim, A. M. Rush, and D. M. Blei, "Avoiding latent variable collapse with generative skip models," *arXiv preprint arXiv:1807.04863*, 2018.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] M. Henaff, J. Zhao, and Y. LeCun, "Prediction under uncertainty with error-encoding networks," *arXiv preprint arXiv:1711.04994*, 2017.

[38] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *European Conference on Computer Vision*. Springer, 2016, pp. 835–851.

[39] K.-H. Thung, P.-T. Yap, E. Adeli, S.-W. Lee, D. Shen, A. D. N. Initiative *et al.*, "Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion," *Med Image Anal*, vol. 45, pp. 68–82, 2018.

[40] B. Lei, W. Hou, W. Zou, X. Li, C. Zhang, and T. Wang, "Longitudinal score prediction for alzheimer's disease based on ensemble correntropy and spatial–temporal constraint," *Brain Imaging Behav*, vol. 13, no. 1, pp. 126–137, 2019.