Forecasting Irreversible Disease via Progression Learning

Botong Wu[1,2]*, Sijie Ren[7]*, Jing Li[1,6], Xinwei Sun[4](✉), Shi-Ming Li[5], Yizhou Wang [1,3]
[1] Dept. of Computer Science, Peking University [2] Adv. Inst. of Info. Tech, Peking University
[3]Center on Frontiers of Computing Studies, Peking University
[4]Microsoft Research, Asia [5]Beijing Tongren Hospital, Capital Medical University
[6] Deepwise AI Lab [7]Beijing Stars Universal Technology Co., Ltd
{botongwu, lijingg, yizhou.wang}@pku.edu.cn, rensijie@ooyby.com, xinsun@microsoft.com

Abstract

*Forecasting Parapapillary atrophy (PPA), i.e., a symptom related to most irreversible eye diseases, provides an alarm for implementing an intervention to slow down the disease progression at early stage. A key question for this forecast is: how to fully utilize the historical data (e.g., retinal image) up to the current stage for future disease prediction? In this paper, we provide an answer with a novel framework, namely **D**isease **F**orecast via **P**rogression **L**earning (**DFPL**), which exploits the irreversibility prior (i.e., cannot be reversed once diagnosed). Specifically, based on this prior, we decompose two factors that contribute to the prediction of the future disease: i) the current disease label given the data (retinal image, clinical attributes) at present and ii) the future disease label given the progression of the retinal images that from the current to the future. To model these two factors, we introduce the current and progression predictors in DFPL, respectively. In order to account for the degree of progression of the disease, we propose a temporal generative model to accurately generate the future image and compare it with the current one to get a residual image. The generative model is implemented by a recurrent neural network, in order to exploit the dependency of the historical data. To verify our approach, we apply it to a PPA in-house dataset and it yields a significant improvement (e.g., **4.48%** of accuracy; **3.45%** of AUC) over others. Besides, our generative model can accurately localize the disease-related regions.*

1. Introduction

The World Health Organization (WHO) estimates that 19 million children below the age of 15 were visually impaired [18; 6] (1% of the total population in this age group). Most of the eye diseases, such as myopia in children [14], glaucoma [26], retinal detachment, and dense cataract [11], are highly related to Parapapillary atrophy (PPA), which as a biomarker of above eye diseases, refers to outer retinal atrophy adjacent to the optic disc [23; 17; 3]. Due to the irreversibility of these eye diseases, forecasted PPA can be provided as an alarm to implement an intervention (*e.g.*, outdoor activities, or drug treatment) to prevent the rapid progression of eye diseases at the early stage. Due to the lack of future data when forecasting the future label, this forecasting task is equivalent to the following answer: *how to fully utilize the longitudinal/sequential data up to the current stage for future disease prediction, under the lack of future data?*

A series of works have recently been proposed to answer this question, such as [21; 15; 24]. Most of these works utilized the provided current data for generating the future medical data (*i.e.*, retinal images, clinical attributes), followed by an auxiliary classifier for disease prediction. However, these methods did not take the *irreversibility* medical prior into account.

This irreversibility prior overlooked in the above literature refers to that, the disease cannot reverse to healthy once diagnosed. That is, if diagnosed as diseased at present, the probability of disease at the future stage would be 100%. Inspired by such a *prior* in PPA[12], we decompose (according to the law of total probability) the disease label at future stage into two factors: **i)** the disease label at current stage given the medical data at present; and **ii)** the disease label at future stage given the progression from the current to the future stage. This factorization, in contrast to previous works that only leverage current data for the generation, claims an additional role of current data in determining the disease at present (*a.k.a the i)*). To effectively learn these two factors, we propose a novel framework, namely **D**isease **F**orecast via **P**rogression **L**earning (**DFPL**) which introduces two prediction modules: f_{cur} and f_{prog}, respectively. To further account for the degree of progression, we propose a temporal generative framework based on Generative Adversarial Networks (GAN), in which we incorporate the generator

* denotes equal contribution.

with the recurrent neural network that takes prior sequential data as input to predict the feature map in the next stage. By comparing this generated feature map with the one at the current stage, one can get the residual feature map, as a measurement of the degree of progression.

To validate the utility of our approach, we apply it to a in-house data which belongs to a longitude PPA protocol for clinical diagnosis for primary-school-aged children. The results show a large improvement over others in terms of prediction accuracy (ACC) and Area Under the ROC Curve (AUC): *e.g.*, **4.48%** of accuracy; **3.45%** of AUC. Besides, the visualization result shows that our DFPL equipped with temporal generative learning can localize the disease-related regions such as *optic disc*. An ablation study is further conducted to verify the contribution of each module of our framework. The main contributions can be summarized as follows:

- We are *the first* to point out the two-fold effects from the longitudinal data up to the current stage to the forecast for irreversible diseases: the disease status at present and the one based on progression from the current stage. We propose a novel framework to learn such two effects.
- We propose a temporal generative framework equipped with a recurrent neural network, to learn the dynamics of disease progression.
- Our method can achieve better prediction results than others on an in-house PPA data of primary-school-aged children; besides, the detected disease-related regions can be concentrated on the optic disc.

2. Related Work

Forecasting disease with longitudinal data refers to predicting the disease label at a future stage, given the sequential data up to the current stage. As a simple and effective approach, the deterministic-type method [5; 4; 19] adopted a two-step strategy: first, they extracted semantic features using the convolutional neural network; then they fed these features into a recurrent neural network to predict the future outcome. Alternatively, due to the ability to capture the temporal relation among the sequential data during generation, a series of generative-based methods [24; 15] have recently been proposed. As a typical example, the [24] proposed to generate future data (*e.g.* T) with the data at the current stage (*e.g.* $t < T$), via generative adversarial networks [8]; such a generated data, as a reflection of the progression from the current data to the future, was then fed into a classifier to predict the disease label. The [15] proposed to learn the smooth Riemannian manifold of the whole trajectory, from the low-dimensional latent space via the deep generative model. Compared to [24], the [15] additionally leveraged the information from the past (*e.g.* $\{\tilde{t} < t\}$). However, these methods did not exploit the irreversibility prior [24; 15] and the dependency among sequential data [24] during modeling.

Our Specification. Our method is better-motivated in that we exploit the fact that the disease cannot be reversed at any time in the future once diagnosed, to propose the two-fold effects for disease forecast: the current disease status and the progression. We formulate this proposition as a theoretical guideline of our learning framework, specifically the current and the progression predictors. To further account for the degree of progression, we propose a temporal GAN equipped with the recurrent neural network to generate the future feature maps; besides, we employ the high-order dynamics (*e.g.*, first-order $\boldsymbol{x}_{t_2} - \boldsymbol{x}_{t_1}$; $\boldsymbol{x}_{t_3} - \boldsymbol{x}_{t_2}$ and second-order $(\boldsymbol{x}_{t_3} - \boldsymbol{x}_{t_2}) - (\boldsymbol{x}_{t_2} - \boldsymbol{x}_{t_1})$) as input for prediction.

3. Methodology

Problem Setting & Notation. Our goal is to predict the disease label y_T at future stage T, given (a subset of) retinal fundus images $\boldsymbol{x}_{\leq t}$ and clinical attributes $\boldsymbol{a}_{\leq t}$ (*e.g., height, time for outdoor activities, myopia situation of parents, etc.*) at some time t with $t < T$. The $y_t \in \{\pm 1\}$ for any $t > 0$, with $+1, -1$ respectively denoting the disease and healthy status, without loss of generality. Our data for training this classifier contain N subjects: $\{\boldsymbol{s}^i\}_{i \in [N]}$, where $[N] := \{1, ..., N\}$ and $\boldsymbol{s}^i = \left(\boldsymbol{z}^i_{\leq t}, y^i_T\right)$ where $\boldsymbol{z}^i_t = \left(\boldsymbol{x}^i_t, \boldsymbol{a}^i_t\right)$. Note that due to labeling cost, we do not require the labels y_t before T (*i.e.*, $t < T$), except at initial time point $t = 1$ such that all samples are healthy, i.e., $y^i_1 = -1$ for all $i \in [N]$. We denote the K-order setting as the data of K time points are provided for training and testing, *i.e.*, $(\boldsymbol{x}_{t_{1:K}}, \boldsymbol{a}_{t_{1:K}})$.

Outline. We first introduce our roadmap in section 3.1, guided by our finding that the future disease is affected by the current stage and the progression, as formulated in Prop. 3.1. We then introduce our learning framework in section 3.2, with each module detailedly explained. Finally, we generalize our method to high-order cases (multiple points of images and attributes are observed) in section 3.3.

3.1. Roadmap with Theoretical Guideline

We consider the disease forecast, *i.e.*, $p(y_T = 1 | \boldsymbol{z}_{t_{1:K}})$ with $t_1 < ... < t_K < T$ denoting a sequence of K time points. For simplicity, in the following we consider 1-order case with $K = 1$ (with high-order case $K > 1$ introduced in section 3.3). Due to the inability of reverse the disease status without medical treatment [16], this future prediction should satisfy the following principles:

- *Irreversibility*: Once diagnosed as PPA, one would not transfer to healthy in the future, if no medical intervention is implemented.
- *Deterioration*: The probability of PPA is monotonic with respect to the time t.

Based on the *Irreversibility* principle, it can be induced that the disease status in the next stage is affected by *(i)* the situa-

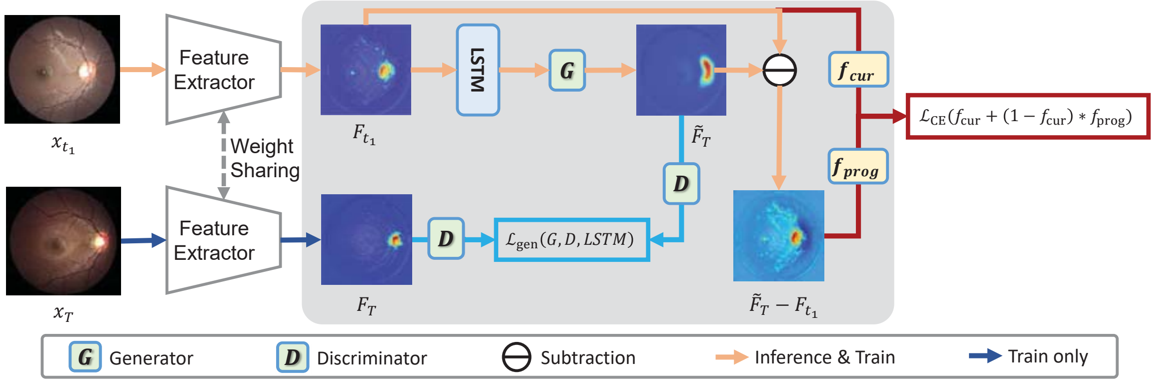

Figure 1. Illustration of our learning framework DFPL. We first pre-train a feature extractor and the extracted feature maps denoted as $\boldsymbol{F}_t$ (together with clinical attributes) are taken as inputs. The modules contained in the gray area are trained in an end-to-end scheme. Specifically the $\boldsymbol{F}_T$ generated by $\mathbb{G}(\text{LSTM}(\boldsymbol{F}_{t_1}, T - t_1), \boldsymbol{a}_{t_1})$ is trained to compete with discriminator $\mathbb{D}$ by adversarial loss. The final prediction is the combination of the current predictor f_{cur} and the residual predictor f_{prog} with residual feature maps $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}$ as input.

tion in the current stage and *(ii)* the progression speed as time grows, which is formulated as the following proposition:

Proposition 3.1. *Under the irreversibility principle, we have the following factorization for progression prediction:*

$$\begin{aligned} p(y_T = 1|\boldsymbol{z}_{t_1}) = &\underbrace{p(y_{t_1} = 1|\boldsymbol{z}_{t_1})}_{\text{Current}} + \\ &p(y_{t_1} = 0|\boldsymbol{z}_{t_1})\underbrace{p(y_T = 1|y_{t_1} = 0, \boldsymbol{z}_{t_1})}_{\text{Progression}}. \end{aligned} \tag{1}$$

Remark 1. *The Prop 3.1 shows that $p(y_T = 1|\boldsymbol{z}_{t_1}) \geq p(y_{t_1} = 1|\boldsymbol{z}_{t_1})$, agreeing with the Deterioration principle.*

As a guideline, we can correspondingly design two modules to respectively model the current disease prediction and the dynamic progression. Besides, the term “progression” can be obtained by

$$\begin{aligned} &p(y_T = 1|y_{t_1} = 0, \boldsymbol{z}_{t_1}) = \\ &\int_{\boldsymbol{x}_T} p(\boldsymbol{x}_T|y_{t_1} = 0, \boldsymbol{z}_{t_1}) * p(y_T = 1|y_{t_1} = 0, \boldsymbol{x}_T, \boldsymbol{z}_{t_1}) d\boldsymbol{x}_T. \end{aligned} \tag{2}$$

For $p(y_T = 1|y_{t_1} = 0, \boldsymbol{x}_T, \boldsymbol{z}_{t_1})$ that describes the extent of progression from the healthy status, we propose to approximate it using progression information which contains *i.e.*, $\boldsymbol{x}_T - \boldsymbol{x}_{t_1}$. We summarize the above conclusions as a roadmap for our learning framework.

RoadMap. We first pre-train a feature extractor to extract feature maps $\boldsymbol{F}_{t_1}$ from retinal fundus images $\boldsymbol{x}_{t_1}$. And the future feature maps $\tilde{\boldsymbol{F}}_T$ are estimated by a trainable temporal generative model with extracted $\boldsymbol{F}_{t_1}$. Then, we learn two prediction modules: f_{cur} and f_{prog}, respectively with the feature maps $\boldsymbol{F}_{t_1}$ and the residual feature maps $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}$ as inputs. The residual feature maps are calculated to measure the degree of progression, as which the current feature maps $\boldsymbol{F}_{t_1}$ are subtracted from the estimated feature maps in future stage $\tilde{\boldsymbol{F}}_T$. We will introduce our learning framework in details in the subsequent section.

3.2. Disease Forecast via Progression Learning

We introduce our learning framework, namely **D**isease **F**orecast via **P**rogression **L**earning (DFPL), with high-level spirit stated in the roadmap in the above section. In more detail, as illustrated in Fig. 1, we first pre-train a feature extractor Enc to extract feature maps from image at each time point. With extracted feature maps at different time steps, *i.e.*, $\boldsymbol{F}_{t_1}, ..., \boldsymbol{F}_{t_K}$ (here we set $K = 1$ for simplicity), we train a convolutional Long Short-Term Memory (LSTM) [10] followed by a generator $\mathbb{G}$ to generate the next stage feature maps, in an adversarial way via Generative Adversarial Networks (GAN) [8]. The extracted feature maps at current stage (*i.e.* $\boldsymbol{F}_{t_1}$) and the residual feature maps with estimated feature maps at the future stage (*i.e.*, $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}$), are respectively taken as inputs for classification modules f_{cur} and f_{prog}. The final prediction is given by $f_{\text{cur}} + (1 - f_{\text{cur}})f_{\text{prog}}$, which is optimized via cross-entropy loss in empirical risk minimization. In the following, we will explain all these modules in details: the pre-trained feature extractor Enc; generative model which is composed of generator $\mathbb{G}$, discriminator $\mathbb{D}$ and the recurrent neural network (here we adopt LSTM [10]); current predictor f_{cur} and progression predictor f_{prog}.

Pre-trained Feature Extractor (Enc)**.** Instead of training directly on images, we implement a pre-training strategy to obtain feature maps denoted as $\boldsymbol{F}$ as the input of classifiers (together with attributes $\boldsymbol{a}$), which has been found to be effective in the literature [7]. Specifically, we train a classifier on **(i)** $\{\boldsymbol{x}_t^i, y_t^i\}_{t \in \{1,T\}, i \in [N]}$ (recall that $y_{t=1}^i = -1$ for all i) to extract features representative of current disease status; and on **(ii)** $\{\boldsymbol{x}_t^i, y_T^i\}_{t<T, i \in [N]}$ to extract features that related to the progression. The bottom layers of neural networks after pre-training are (*e.g.*, the first two blocks for ResNet18

in experiment) denoted as feature extractor Enc. In the following, we take extracted feature maps as input of modules LSTM, $\mathbb{G}$, $\mathbb{D}$, f_{cur} and f_{prog} (the gray area in Fig. 1).

Generative Model. The goal is to generate the feature maps at future stage (*i.e.*, $\tilde{\boldsymbol{F}}_T$). By comparing it with the $\boldsymbol{F}_{t_1}$ at the current stage, the $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}$ measures the degree of progression and is thus fed into f_{prog} to predict the $p(y_T = 1|y_{t_1} = 0, \boldsymbol{z}_{t_1})$ in Eq. (1). For an accurate generation, we adopt the adversarial training strategy, specifically the Wasserstein GAN (WGAN) [1] with weight clipping, to train the generator $\mathbb{G}$ and a discriminator $\mathbb{D}$ in a competing way. To further capture the dependency of the historical feature maps, we additionally train a LSTM of which the output is then fed into the generator $\mathbb{G}$, as shown in Fig. 1. The generative loss function for the 1-order generation (with the higher-order generation introduced later) is $\mathcal{L}_{\text{gen}}(\mathbb{G}, \mathbb{D}, \text{LSTM})$. The generative loss is computed by the real future feature maps $\boldsymbol{F}_T^i$ and the generated future feature maps $\tilde{\boldsymbol{F}}_T^i = \mathbb{G}(\text{LSTM}(\boldsymbol{F}_{t_1}^i, T - t_1), \boldsymbol{a}_{t_1}^i)$.

Current Predictor. The f_{cur}, as the predictor of current disease status given $\boldsymbol{F}_t := \text{Enc}(\boldsymbol{x}_t)$[1], is trained via the empirical risk minimization (ERM) of labeled training data (initial time point and future time point) and generated future data:

$$\mathcal{L}_{\text{ERM}}(f_{\text{cur}}) = \sum_{i\in[N]} \left(\sum_{t\in\{1,T\}} \log \frac{1}{p_{f_{\text{cur}}}(y_t^i|\boldsymbol{F}_t^i, \boldsymbol{a}_t^i)} + \sum_{t_1<T} \log \frac{1}{p_{f_{\text{cur}}}(y_T^i|\tilde{\boldsymbol{F}}_T^i(\boldsymbol{F}_{t_1}^i, \boldsymbol{a}_{t_1}^i))} \right), \quad (3)$$

where $\tilde{\boldsymbol{F}}_T^i(\boldsymbol{F}_{t_1}^i, \boldsymbol{a}_{t_1}^i) := \mathbb{G}(\text{LSTM}(\boldsymbol{F}_{t_1}^i, T - t_1), \boldsymbol{a}_{t_1}^i)$. Therefore, the $\mathcal{L}_{\text{ERM}}$ also trains the generator $\mathbb{G}$ and the LSTM, which is omitted here for simplicity. Besides, we additionally regularize $p(y_T = 1|\boldsymbol{F}_T) \geq p(y_{t_1} = 1|\boldsymbol{F}_{t_1})$ for any $T > t_1$ according to the *Deterioration* principles, formulated as soft-margin regularization:

$$\mathcal{J}_{\text{cur}}(f_{\text{cur}}) = \sum_{i\in[N],t_1<T} \max\left(0, \text{diff}_i(\boldsymbol{F}_T^i, \boldsymbol{F}_{t_1}^i) + \theta\right), \quad (4)$$

where $\text{diff}_i(\boldsymbol{F}_T^i, \boldsymbol{F}_{t_1}^i) := p_{f_{\text{cur}}}(y_{t_1}^i = 1|\boldsymbol{F}_{t_1}^i) - p_{f_{\text{cur}}}(y_T^i = 1|\boldsymbol{F}_T^i)$ and $\theta > 0$ denotes the margin hyper-parameter. The overall loss function to train f_{cur} is:

$$\mathcal{L}_{\text{cur}}(f_{\text{cur}}) = \mathcal{L}_{\text{ERM}}(f_{\text{cur}}) + \alpha \mathcal{J}_{\text{cur}}(f_{\text{cur}}), \quad (5)$$

with $\alpha > 0$ denoting the hyper-parameter that balances the effects of prediction and the *Deterioration* principle.

Progression Predictor. As aforementioned in sec. 3.1, the "progression" term can be approximated by $p(y_T = 1|y_{t_1} = 0, \boldsymbol{z}_{t_1}) \approx \int p(\boldsymbol{F}_T|\boldsymbol{F}_{t_1}, \boldsymbol{z}_{t_1}) p(y_T = 1|\boldsymbol{F}_T - \boldsymbol{F}_{t_1}, \boldsymbol{a}_{t_1}) d\boldsymbol{F}_T$. The loss for f_{prog} taking the residual feature maps $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}$ as input and also f_{cur}, according to factorization of "current" and "progression" term in Prop. 3.1, is reformulated as:

$$\mathcal{L}_{\text{CE}}(f_{\text{prog}}, f_{\text{cur}}) = \sum_{i\in[N],t_1<T} \log \frac{1}{p_{f_{\text{cur}},f_{\text{prog}}}(y_T^i|\tilde{\boldsymbol{F}}_T^i, \boldsymbol{F}_{t_1}^i, \boldsymbol{a}_{t_1}^i)}, \quad (6)$$

$$p_{f_{\text{cur}},f_{\text{prog}}}(y_T^i = 1|\tilde{\boldsymbol{F}}_T^i, \boldsymbol{F}_{t_1}^i, \boldsymbol{a}_{t_1}^i) = p_{f_{\text{cur}}}(y_{t_1}^i = 1|\boldsymbol{F}_{t_1}^i, \boldsymbol{a}_{t_1}^i) + p_{f_{\text{cur}}}(y_{t_1}^i = 0|\boldsymbol{F}_{t_1}^i, \boldsymbol{a}_{t_1}^i) p_{f_{\text{prog}}}(y_T^i = 1|\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}, \boldsymbol{a}_{t_1}^i). \quad (7)$$

Note that the $\mathcal{L}_{\text{CE}}$ also depends on the generator $\mathbb{G}$ and the LSTM since that the $\tilde{\boldsymbol{F}}_T := \mathbb{G}(\text{LSTM}(\boldsymbol{F}_{t_1}, T - t_1), \boldsymbol{a}_{t_1})$.

Training & **Inference.** Combining separate losses for the modules mentioned above (specifically Eq. (5), $\mathcal{L}_{\text{gen}}(\mathbb{G}, \mathbb{D}, \text{LSTM})$ and Eq. (6)), the overall loss function is defined as:

$$\mathcal{L}(f_{\text{cur}}, f_{\text{prog}}, \mathbb{G}, \mathbb{D}, \text{LSTM}) := \mathcal{L}_{\text{gen}}(\mathbb{G}, \mathbb{D}, \text{LSTM}) + \lambda_1 * \mathcal{L}_{\text{cur}}(f_{\text{cur}}) + \lambda_2 * \mathcal{L}_{\text{CE}}(f_{\text{prog}}, f_{\text{cur}}). \quad (8)$$

During inference, given $(\boldsymbol{x}_{t_1}, \boldsymbol{a}_{t_1})$, we first obtain $\boldsymbol{F}_{t_1}$ via $\text{Enc}(\boldsymbol{x}_{t_1})$. Then we generate the $\tilde{\boldsymbol{F}}_T$ via $\mathbb{G}(\text{LSTM}(\boldsymbol{F}_{t_1}, T - t_1), \boldsymbol{a}_{t_1})$. Then we feed $(\tilde{\boldsymbol{F}}_T, \boldsymbol{F}_{t_1}, \boldsymbol{a}_{t_1})$ into $p_{f_{\text{cur}},f_{\text{prog}}}$ in Eq. (7) for prediction.

3.3. Extension to High-Order Prediction

We extend our loss in Eq. (8) to leverage high-order information (including the information from the past, *i.e.*, $\boldsymbol{z}_{t_{1:K-1}}$ and the current, *i.e.*, $\boldsymbol{z}_{t_K}$) into the generation of feature maps at future stage and hence the future disease, *i.e.*, $p(y_T = 1|\boldsymbol{z}_{t_{1:K}})$ with $K > 1$. The Prop. 3.1 for this case is presented similarly, with factorization of the current and the progression (please refer to supplementary for details). Therefore, the whole framework can be inherited and the extensions of K-order for current predictor, generative model and progression predictor are summarized as follows.

Current Predictor. We consider the $p(y_{t_K}|\boldsymbol{z}_{t_{1:K}})$ for any $t_1 < .. < t_K < T$. To leverage the information before t_K, *i.e.*, $\boldsymbol{z}_{t_{1:K-1}}$, we additionally train a classifier from $\boldsymbol{z}_t$ to y_T (the label only given at T), namely f_{fut} (with "fut" standing for the word "future"):

$$\mathcal{L}_{\text{fut}}(f_{\text{fut}}) = \sum_{i\in[N],t_1<T} \frac{1}{\log p_{f_{\text{fut}}}(y_T^i|\boldsymbol{F}_{t_1}^i, \boldsymbol{a}_{t_1}^i)}. \quad (9)$$

Based on the current predictor f_{cur} and future predictor f_{fut}, the $p(y_{t_K}|\boldsymbol{F}_{t_{1:K}}, \boldsymbol{a}_{t_{1:K}})$ is then modeled as:

$$p_{f_{\text{cur}},f_{\text{fut}}}(y_{t_K}|\boldsymbol{F}_{t_{1:K}}, \boldsymbol{a}_{t_{1:K}}) = \quad (10)$$
$$\frac{1}{K} \left(p_{f_{\text{cur}}}(y_{t_K}|\boldsymbol{F}_{t_K}, \boldsymbol{a}_{t_K}) + \sum_{j=1}^{K-1} p_{f_{\text{fut}}}(y_{t_K}|\boldsymbol{F}_{t_j}, \boldsymbol{a}_{t_j}) \right).$$

[1] The input does not contain $\boldsymbol{a}$, since it is unobserved at T for the 2nd term in Eq. (3)

Generative Model. To leverage the high-order information into the generation of the future maps in T, from the past K-length sequence (*i.e.* $\boldsymbol{F}_{t_1}, \boldsymbol{a}_{t_1}, ..., \boldsymbol{F}_{t_K}, \boldsymbol{a}_{t_K}$ for any $t_1 < t_2 < ... < t_K < T$), we iteratively feed the feature maps and related attributes into the LSTM up to the t_K, followed by the generator $\mathbb{G}$ that is trained by adversarial loss to compete the discriminator $\mathbb{D}$. The $\mathcal{L}_{\text{gen}}(\mathbb{G}, \mathbb{D}, \text{LSTM})$ is computed by the real feature maps and generated feature maps from t_2 to T. Equipped with the LSTM's ability of long-term memory, this high-order generation can capture the time-dependency.

Progression Predictor. For progression learning, the high-order residual information can be approximated by differentiation of the ones of lower-order (*e.g.*, the second-order residual at time t_2 can be approximated by difference of two first-order residuals $\boldsymbol{F}_{t_3} - \boldsymbol{F}_{t_2}$ and $\boldsymbol{F}_{t_2} - \boldsymbol{F}_{t_1}$ as $(\boldsymbol{F}_{t_3} - \boldsymbol{F}_{t_2}) - (\boldsymbol{F}_{t_2} - \boldsymbol{F}_{t_1})$). Generally speaking, the set of $\{j\}_{j \in [K]}$-order residual information denoted as $\text{prog}_K(\tilde{\boldsymbol{F}}_T, \{\boldsymbol{F}_{t_j}\}_{j \in [K]})$ is composed of **(i)** the first-order information $\{\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_K}, \{\boldsymbol{F}_{t_{K-i}} - \boldsymbol{F}_{t_{K-i-1}}\}_{i=0}^{K-2}\}$; and **(ii)** the ones related to the j-th order for $j \geq 2$, represented by $\{(\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_{K+2-j}}) - (\boldsymbol{F}_{t_K} - \boldsymbol{F}_{t_{K+1-j}}), \{(\boldsymbol{F}_{t_{K-i}} - \boldsymbol{F}_{t_{K-i+1-j}}) - (\boldsymbol{F}_{t_{K-i-1}} - \boldsymbol{F}_{t_{K-i-j}})\}_{i=0}^{K-2}\}$. The loss is the same with Eq. (6) except that the input of f_{prog} turns to $\text{prog}_K(\tilde{\boldsymbol{F}}_T, \{\boldsymbol{F}_{t_j}\}_{j \in [K]})$ $(K > 1)$ and $\boldsymbol{a}_{t_{1:K}}$, and the $p_{f_{\text{cur}}}(\boldsymbol{F}_{t_1})$ is replaced with Eq. (10) that additionally leverage the information . In summary, the $p_{f_{\text{cur}}, f_{\text{prog}}}$ in Eq. (7) (with $\boldsymbol{\zeta} := \{\boldsymbol{F}, \boldsymbol{a}\}$) is replaced by:

$$\begin{aligned} & p_{f_{\text{cur}}, f_{\text{prog}}, f_{\text{fut}}}(y_T^i = 1 | \boldsymbol{\zeta}_{t_{1:K}}^i) = \\ & \quad p_{f_{\text{cur}}, f_{\text{fut}}}(y_{t_K} = 1 | \boldsymbol{\zeta}_{t_{1:K}}^i) + p_{f_{\text{cur}}, f_{\text{fut}}}(y_{t_K} = 0 | \boldsymbol{\zeta}_{t_{1:K}}^i) \\ & \quad p_{f_{\text{prog}}}(y_T^i = 1 | \text{prog}_K(\tilde{\boldsymbol{F}}_T^i, \{\boldsymbol{F}_{t_j}^i\}_{j \in [K]}), \boldsymbol{a}_{t_{1:K}}^i). \end{aligned} \tag{11}$$

Training & Inference. The overall loss function on high-order setting is same as 1-order one Eq.(8) except that the future loss Eq.(9) need to be considered and the prediction for $\mathcal{L}_{\text{CE}}$ is computed by Eq. (11). The overall loss function on high-order setting is defined as:

$$\begin{aligned} & \mathcal{L}(f_{\text{cur}}, f_{\text{prog}}, f_{\text{fut}}, \mathbb{G}, \mathbb{D}, \text{LSTM}) := \mathcal{L}_{\text{gen}}(\mathbb{G}, \mathbb{D}, \text{LSTM}) \\ & + \lambda_1 * \mathcal{L}_{\text{cur}}(f_{\text{cur}}) + \lambda_2 * \mathcal{L}_{\text{CE}}(f_{\text{prog}}, f_{\text{cur}}, f_{\text{fut}}) \\ & + \lambda_3 * \mathcal{L}_{\text{fut}}(f_{\text{fut}}). \end{aligned} \tag{12}$$

During inference, the process is the same to 1-order setting except that feeding a sequential data $(\boldsymbol{x}_{t_{1:K}}, \boldsymbol{a}_{t_{1:K}})$ into Enc, $\mathbb{G}$ and LSTM to compute the feature maps $\boldsymbol{F}_{t_{1:K}}$ the high-order residual information set $\text{prog}_K(\tilde{\boldsymbol{F}}_T^i, \{\boldsymbol{F}_{t_j}^i\}_{j \in [K]})$. Then we feed above feature maps and related attributes into $f_{\text{cur}}, f_{\text{prog}}$ and f_{fut} in Eq. (11) for prediction.

4. Experimental Results

In this section, we evaluate our method on an in-house longitudinal dataset, which studies the PPA progression for primary-school-aged (from grade-1 to grade-6) children.

4.1. Dataset

The data contains 905 participants in primary school, with each containing 3-6 data records (retinal fundus images $\boldsymbol{x}$ and clinical information $\boldsymbol{a}$, *e.g.* height, time for outdoor activities, myopia situation of parents [2]). In total, there are 5,046 data. Due to the costly labeling process, only the labels for the images from 1st graders and 6th graders are provided, with all participants at grade-1 being healthy. The data is randomly split into 60% for training (543), 20% for validation (181), and 20% for testing (181) according to the index of participants. Our goal is to predict whether one would develop the disease at the final stage (*i.e.* at grade-6), for any samples in the test data at the early stage.

4.2. Baselines for Comparison

a) **R**es**N**et-18 (RN18) [9] which is trained to minimize cross entropy loss from $\boldsymbol{x}_t$ to y_T for any $t < T$. For the network structure, we replace the 7×7 kernels in the first convolutional layer replaced by the two convolutional layers with kernel size 3×3. We empirically find that this replacement can achieve better prediction results. For simplicity, we name it as RN18, without otherwise specified.

b) **M**ulti-**M**odality-**F**using (MM-F) [2], which proposed to fuse information of images $\boldsymbol{x}$ and clinical information $\boldsymbol{a}$ by concatenating features extracted from $\boldsymbol{x}$ via RN18 and those from $\boldsymbol{a}$ via a 3-layer (128→256→256) multilayer perceptron (MLP). It is also trained by minimizing cross entropy loss from $\boldsymbol{x}_t$ to y_T for any $t < T$.

c) **T**emporal **C**orrelation **S**tructure **L**earning (TCSL) [24]. It implemented GAN to learn the joint distribution of $(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}, y_{t+1})$ in order to capture the temporal relation between the adjacent points, followed by a classifier to predict the future label. Besides, it additionally trained a regression network to learn $\boldsymbol{x}_{t+1}$ from $\boldsymbol{x}_t$. For fair comparison, we adopt the same network structure ofthe generator and the discriminator as ours. We adopt the RN18 for the follow-up classifier and the U-Net [20] for the regression network. More implementation details can refer to [24]. Since it generated only with adjacent point, we only compare it with our method on 1-order setting.

d) **A**ttention **R**esidual **L**earning (ARL) [25]. It introduced an ARL-block which fuses input feature maps, residual feature maps, and attention feature maps to replace the traditional residual block in ResNet. The attention feature maps are computed by element-wise product of input feature maps and normalized residual feature maps. We replace the residual block with ARL-blocks for the MM-F method.

e) **R**iemannian **G**eometry **L**earning (RGL) [15]. It proposed a Riemannian manifold for the whole trajectory. As a

[2] For details please refer to supplementary information.

Table 1. The ACC, AUC (mean ± std) comparisons between our method and the baselines on the 1-order setting. $\delta t = T - t_i$ with the $\delta t = 1$ implying that the input of the test sample is from 5th graders since T represents $T = 6$. Average over ten runs.

Methods	RN18		MM-F		ARL		TCSL		Ours	
Num of Param	138.68M		137.19M		138.69M		157.51M		141.56M	
Metric	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
δt=5	60.53±2.55	63.56±2.51	63.46±1.77	64.38±1.62	63.14±1.90	65.58±3.24	58.56±0.56	56.53±1.01	**66.67**±1.94	**72.37**±0.82
δt=4	65.25±2.01	70.41±1.98	66.38±2.50	71.88±1.54	67.88±1.92	73.88±2.33	62.98±0.96	66.89±0.64	**69.80**±1.94	**76.88**±0.42
δt=3	62.80±2.77	66.80±1.64	63.07±1.65	67.52±2.36	67.09±1.02	71.32±2.52	66.12±1.69	69.20±0.92	**69.98**±1.69	**78.65**±1.02
δt=2	69.92±1.92	75.78±2.56	69.91±2.34	79.48±1.88	70.80±2.68	80.13±1.19	74.52±1.54	79.65±0.45	**77.16**±1.28	**83.52**±1.12
δt=1	73.05±3.36	82.74±1.51	75.50±2.86	86.70±1.04	75.30±3.02	86.68±1.57	77.53±1.94	84.88±0.27	**79.37**±2.09	**87.16**±1.12
Average	66.31±1.24	71.86±1.20	67.6.7±0.78	73.99±0.76	68.84±1.47	75.52±1.52	67.88±1.18	71.43±0.20	**72.60**±1.53	**79.72**±0.50

Table 2. The ACC, AUC (mean ± std) comparisons between our method and the baselines on the 2-order setting. $\delta t = T - t_i$ with the $\delta t = 1$ implying that the input of test samples are from 4th graders and 5th graders. Average over ten runs.

Methods	RN18		MM-F		ARL		RGL		Ours	
Num of Param	154.90M		155.46M		154.91M		150.28M		152.72M	
Metric	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
δt=4	62.54±1.81	68.09±1.67	68.84±1.80	74.21±1.35	67.96±2.24	74.23±1.47	69.24±1.25	76.34±1.25	**70.17**±1.11	**76.42**±1.20
δt=3	64.81±1.62	71.47±1.70	71.22±1.92	78.38±0.87	71.13±1.37	77.49±2.01	68.39±1.68	74.38±1.26	**72.75**±1.69	**80.16**±0.52
δt=2	67.07±3.53	74.07±1.39	73.43±1.86	81.14±0.67	70.99±0.62	80.15±2.51	75.26±2.14	81.98±0.64	**77.17**±1.69	**85.28**±0.51
δt=1	73.43±2.32	81.84±0.93	76.57±1.19	85.80±0.93	75.42±1.49	85.34±1.49	78.33±1.95	87.71±1.17	**78.64**±1.77	**90.04**±0.27
Average	66.96±1.47	73.87±0.99	72.51±1.04	79.88±0.57	71.37±1.10	79.30±1.74	72.81±0.98	80.10±0.57	**74.68**±1.25	**82.98**±0.52

high-order method, it implemented a deep generative model to map low-dimensional latent space to the high-dimensional observational data that lie in a geodesics of the manifold. We adopt the RN18 as the encoder and the same network structure of our generator as the decoder.

To compare with the high-order version of our method, we extend RN18, MM-F and ARL baselines to K-order version ($K > 1$). Specifically, as for K-order method, we optimize the sum of cross entropy losses with the k-th loss taking $\boldsymbol{x}_{t_k}$ as input. The final prediction is $\frac{1}{K} \sum_{k=1}^{K} p(y_T = 1 | \boldsymbol{x}_{t_k})$.

4.3. Implementation details

We first pre-train a RN18 from $\boldsymbol{x}_t$ to y_t for $t \in \{1, T\}$ and from $\boldsymbol{x}_t$ to y_T for $t < T$, to obtain the feature extractor Enc as the first two convolutional layers followed by two residual blocks. The output of the feature extractor is 128 feature maps with size 64×64. Then we concatenate features from (i) down-sampled 32×32 feature maps via two Conv-BN-ReLU blocks [3]; and (ii) up-sampled 32×32 feature maps obtained from four TransposeConv-BN-ReLU blocks (with the channel size: $106 \to 2048 \to 1024 \to 1024 \to 512$) with a concatenated vector of clinical attributes $\boldsymbol{a} \in \mathbb{R}^6$ and a 100-dimensional Gaussian noise vector. The concatenated feature maps are then fed into a one-layer convolutional LSTM (with channel size 256) to generate the feature maps with size 32×32 at the next time point, followed by a generator $\mathbb{G}$ with a TransposeConv-BN-ReLU block and a Conv-BN-ReLU block (with the channel size: $256 \to 256 \to 128$) that outputs 128 feature maps with size 64×64 (same as the size of extracted feature maps). The discriminator $\mathbb{D}$ composes of five Conv-BN-LeakyReLU blocks (with the channel size $128 \to 256 \to 512 \to 1024 \to 1024 \to 1024 \to 1$). The negative slope of LeakyReLU is set to 0.2. As for K-order ($K > 1$) version, the input (and output) of LSTM module are changed to the corresponding sequential data with length of K.

We adopt center-cropping on the original image and resize them to 256×256. Then, we apply random rotation with ≤ 30 degrees on each training image. All images are normalized with mean of 0.5 and std of 0.5. We respectively adopt RMSprop (with learning rate (lr) of 0.0001, weight decay (wd) of 0.0001) to train the generator and the discriminator and SGD (lr of 0.02, wd of 0.0001) to train the classification networks. We train the full model for 120 epochs and decay the lr by 0.2 every 60 epochs. The batch size is set to 20. The epoch number is optimized via the prediction accuracy on the validation set. The λ_1 and λ_2 in Eq.(8) are set to 0.1 and 1.0 for all order settings. The α in Eq.(5) is set to 0.1. The λ_3 in Eq.(12) is set to 1.0. During inference, we ensemble the models: i) Eq. (11), ii) $p_{f_{\mathrm{cur}}}$ in Eq. (10) and iii) $p_{f_{\mathrm{fut}}}$ in Eq. (9), *i.e.*, $\frac{1}{K+2} \Big(p_{f_{\mathrm{cur}}, f_{\mathrm{prog}}, f_{\mathrm{fut}}}(y_T = 1 | \boldsymbol{F}_{t_{1:K}}, \boldsymbol{a}_{t_{1:K}}) + p_{f_{\mathrm{cur}}}(y_T = 1 | \tilde{\boldsymbol{F}}_T, \mathbf{0}) + \sum_{j=1}^{K} p_{f_{\mathrm{fut}}}(y_T = 1 | \boldsymbol{F}_{t_j}, \boldsymbol{a}_{t_j}) \Big)$. $\mathbf{0} \in \mathbb{R}^6$ denotes the zero vector due to the attributes are not given at the future stage T. The average and standard deviation over 10 runs are reported.

4.4. Quantitative Results

We consider three evaluation settings: 1-order in Tab. 1, 2-order in Tab. 2 and 3-order in Tab. 3. The TCLS [24], which only leveraged adjacent point for generation is only compared with others on 1-order setting; and the RGL [15] which generates the trajectory is compared with on 2-order and 3-order settings. As shown in Tab. 1,2 and 3, our method perform better and comparable than others in terms of prediction accuracy (ACC) and AUC metrics on all settings.

[3] The "Conv", "BN" respectively stand for Convolution and Batch Normalization.

Table 3. The ACC, AUC (mean ± std) comparisons between our method and the baselines on the 3-order setting. $\delta t = T - t_i$ with the $\delta t = 1$ implying that the input of the test samples is from 3rd graders, 4th graders, and 5th graders. Average over ten runs.

Methods	RN18		MM-F		ARL		RGL		Ours	
Num of Param	154.90M		155.46M		154.91M		150.28M		152.72M	
Metric	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
δt=3	63.77±1.74	68.72±1.55	67.36±1.17	73.29±1.96	68.50±2.90	74.98±2.78	66.04±2.51	71.69±1.09	**74.03**±0.56	**81.47**±0.52
δt=2	68.67±2.34	74.41±1.16	73.54±1.29	80.26±1.20	70.07±2.67	80.07±2.19	73.57±1.73	79.81±1.48	**77.53**±1.15	**85.81**±0.51
δt=1	75.69±1.64	80.33±1.32	75.30±2.22	84.51±1.06	74.55±3.09	85.48±1.08	77.31±1.31	85.84±1.15	**79.56**±1.46	**88.92**±0.31
Average	69.38±1.06	74.49±1.04	72.07±0.81	79.35±1.20	71.15±2.57	80.01±1.72	72.30±1.02	79.12±0.96	**77.04**±1.01	**85.40**±0.17

Table 4. Ablation study on 1-order setting, to validate the effectiveness of each module. The Eq.(1) means that we train the model with loss Eq.(8) and predict by $f_{\text{cur}} + (1 - f_{\text{cur}})f_{\text{prog}}$. "MA" stands for Model Average with $f_{\text{cur}}(\tilde{\boldsymbol{F}}_T)$, Eq.(1) and $f_{\text{fut}}(\boldsymbol{F}_{t_i})$. f_{cur} denotes that we train the model with $\mathcal{L}_{\text{gen}} + \lambda_1 \mathcal{L}_{\text{cur}}$ and predict by $f_{\text{cur}}(\tilde{\boldsymbol{F}}_T)$. f_{prog} denotes that we train the model with $\mathcal{L}_{\text{gen}} + \lambda_2'(-\log p_{f_{\text{prog}}}(y_T|\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}, \boldsymbol{a}_{t_1}))$ and predict by $f_{\text{prog}}(\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}, \boldsymbol{a}_{t_1})$.

Predictor	LSTM	MA	δt=5		δt=4		δt=3		δt=2		δt=1		Average	
			ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
f_{cur}	√	×	64.64	63.73	67.96	75.52	66.85	73.64	71.82	83.02	79.56	88.03	70.17	76.79
f_{prog}	√	×	66.85	69.33	66.85	70.10	68.51	73.93	70.17	77.75	74.03	79.82	69.28	74.19
Eq.(1)	√	×	**69.06**	71.60	69.61	75.83	69.06	76.54	72.38	82.47	78.45	86.70	71.71	78.63
Eq.(1)	×	√	68.51	69.87	62.98	71.90	65.19	73.23	70.17	78.83	76.80	85.29	68.73	75.82
Eq.(1)	√	√	68.51	**72.14**	**71.82**	**77.36**	**71.82**	**78.49**	**77.90**	**83.56**	**81.77**	**87.38**	**74.36**	**79.79**

Table 5. Comparisons of different K-order (time steps) leveraged. All settings share the same set of sample indexes.

Task	1-order		2-order		3-order	
Metric	ACC	AUC	ACC	AUC	ACC	AUC
δt=3	65.75	74.69	68.51	76.79	**71.82**	**78.95**
δt=2	69.06	79.88	74.59	82.45	**75.69**	**83.37**
δt=1	75.14	84.26	76.24	**88.64**	**77.35**	87.49
Average	69.98	79.61	73.11	82.63	**74.95**	**83.27**

4.5. Ablation Study

We conduct an ablation study to validate the effectiveness of each module. The results are summarized in Tab. 4. As shown, the improvement of Eq. (1) (the 3rd row) over the first two rows (validate the effectiveness of $f_{\text{cur}}, f_{\text{prog}}$ in the disease forecast, as guided by Prop. 3.1. Besides, the incorporation of LSTM into our model can bring additional improvement (of the 5th row over the 4th row), due to the ability of LSTM to exploit the dependency of sequential data. Finally, implementing the model ensemble can achieve further improvement, as shown by the result in the 5th row compared to the one in the 3rd row.

Moreover, to validate the advantage of leveraging higher-order information to generate future image (hence residual feature map), we keep the samples with data on all-time steps ($t = 1 : 6$) provided for 1-order, 2-order, and 3-order settings. As shown in Tab. 5, the higher-order data we leverage, the better performance we can achieve (0.64% AUC of 3-order over 2-order; and 3.02% AUC of 2-order over 1-order).

4.6. Visualization

To verify that our method can learn interpretable features for disease forecast, we visualize the estimated feature maps by our method on 1-order and 2-order settings in Fig. 2 and Fig. 3, respectively. In Fig. 2 for one diseased case, the feature maps from top to bottom are, real images from 1st graders to 5th graders ($t_{i:1\to5}$), feature maps generated by feature extractor Enc on images from 1st graders to 5th graders (*i.e.*, $\boldsymbol{F}_{t_{i:1\to5}}$), the 5-time repeated feature maps generated by Enc on image from 6th graders (*i.e.*, $\boldsymbol{F}_T$), the future feature maps (*i.e.*, $\tilde{\boldsymbol{F}}_{T=6}(\boldsymbol{F}_{t_{i:1\to5}})$ estimated by our temporal generative model which respectively taking $\boldsymbol{F}_{t_1}$,...,$\boldsymbol{F}_{t_5}$ as inputs, residual feature maps $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_{i:1\to5}}$. In Fig. 3, from left to right are: real images (*i.e.*, $\boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2}, \boldsymbol{x}_T$ ($t_1 < t_2 < T$)), the corresponding feature maps via Enc (*i.e.*, $\boldsymbol{F}_{t_1}, \boldsymbol{F}_{t_2}, \boldsymbol{F}_T$ ($t_1 < t_2 < T$)), the estimated feature maps by our generative model (*i.e.*, $\tilde{\boldsymbol{F}}_{t_2}(\boldsymbol{F}_{t_1}), \tilde{\boldsymbol{F}}_T(\boldsymbol{F}_{t_{1:2}})$), the progression information in orange box which from left to right are: first-order estimated residual feature maps $\tilde{\boldsymbol{F}}_{t_2}(\boldsymbol{F}_{t_1}) - \boldsymbol{F}_{t_1}$, $\tilde{\boldsymbol{F}}_T(\boldsymbol{F}_{t_{1:2}}) - \boldsymbol{F}_{t_1}$, $\tilde{\boldsymbol{F}}_T(\boldsymbol{F}_{t_{1:2}}) - \boldsymbol{F}_{t_2}$ first-order residual feature maps $\boldsymbol{F}_{t_2} - \boldsymbol{F}_{t_1}$, $\boldsymbol{F}_T - \boldsymbol{F}_{t_1}$, $\boldsymbol{F}_T - \boldsymbol{F}_{t_2}$; second-order estimated residual feature maps: $(\tilde{\boldsymbol{F}}_T(\boldsymbol{F}_{t_{1:2}}) - \boldsymbol{F}_{t_2}) - (\boldsymbol{F}_{t_2} - \boldsymbol{F}_{t_1})$.

As shown in both Fig. 2 and 3, the high response regions (marked by the orange circle) in learned residual feature maps (the last row marked by the blue rectangle in Fig. 2, the last column in orange box in Fig. 3) are concentrated in the optic disc region (marked by the green circle in the third row of Fig. 2) which has been found to be highly correlated with PPA [13; 22]. Besides, it can be shown from the second row in Fig. 2 that the high-response regions in feature maps for 1st graders to 5th graders are more concentrated and similar to that of $\boldsymbol{F}_T$, which matches with our *Deteriorate principle*. Another interesting phenomena, as shown in Fig. 3, is that the high-response area in $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_2}$ (also $\boldsymbol{F}_T - \boldsymbol{F}_{t_2}$) is smaller than that of $\tilde{\boldsymbol{F}}_T - \boldsymbol{F}_{t_1}$ (also $\boldsymbol{F}_T - \boldsymbol{F}_{t_1}$), which validates the interpretability of our residual images in describing the degree of progression information.

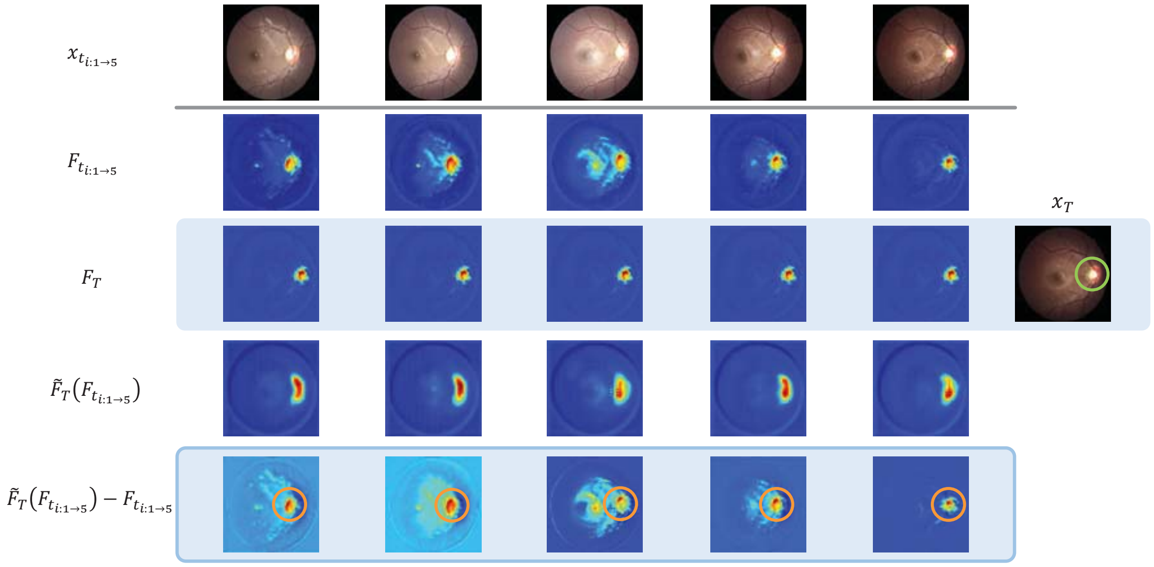
$x_{t_{i:1\to5}}$
$F_{t_{i:1\to5}}$
F_T
x_T
$\tilde{F}_T(F_{t_{i:1\to5}})$
$\tilde{F}_T(F_{t_{i:1\to5}}) - F_{t_{i:1\to5}}$

Figure 2. Visualization of learned feature maps on 1-order task. Feature maps from top to bottom are: $\boldsymbol{x}_{t_{i:1\to5}}$ that represents retinal funds images, $\boldsymbol{F}_{t_{i:1\to5}}$ denoting the feature maps extracted from images, the F_T, the $\tilde{\boldsymbol{F}}_T(\boldsymbol{F}_{t_{i:1\to5}})$ denoting the estimated feature maps via our generative model and the $\tilde{\boldsymbol{F}}_T(\boldsymbol{F}_{t_{i:1\to5}}) - \boldsymbol{F}_{t_{i:1\to5}}$ denoting the residual information.

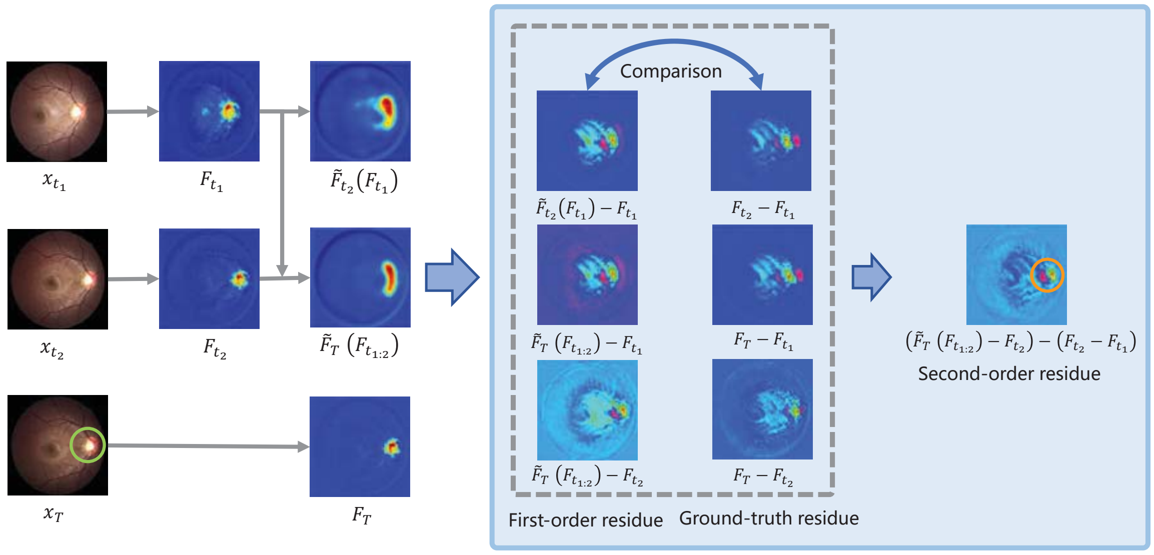
x_{t_1}
F_{t_1}
$\tilde{F}_{t_2}(F_{t_1})$
x_{t_2}
F_{t_2}
$\tilde{F}_T\ (F_{t_{1:2}})$
x_T
F_T
Comparison
$\tilde{F}_{t_2}(F_{t_1}) - F_{t_1}$
$F_{t_2} - F_{t_1}$
$\tilde{F}_T\ (F_{t_{1:2}}) - F_{t_1}$
$F_T - F_{t_1}$
$\tilde{F}_T\ (F_{t_{1:2}}) - F_{t_2}$
$F_T - F_{t_2}$
First-order residue
Ground-truth residue
$(\tilde{F}_T\ (F_{t_{1:2}}) - F_{t_2}) - (F_{t_2} - F_{t_1})$
Second-order residue

Figure 3. Visualization of estimated feature maps on 2-order task. Feature maps from left to right are: $\boldsymbol{x}_{t_{1:2}}, \boldsymbol{x}_T$ denoting retinal funds images, $\boldsymbol{F}$ denoting the corresponding feature maps, $\tilde{\boldsymbol{F}}_{t_2}(\boldsymbol{F}_{t_1})$, $\tilde{\boldsymbol{F}}_T(\boldsymbol{F}_{t_{1:2}})$ denoting the estimated feature maps, estimated first-order residual feature maps, first-order residual feature maps and estimated second-order residual feature maps. We use red circle to mark the high-response area (the response extent is from high to low for the color from the red to the blue).

5. Conclusions & Discussions

We present a framework to perform Disease Forecast via Progression Learning (DFPL) applied on an in-house sequential dataset for Parapapillary Atrophy forecast. To our knowledge, we are *the first* to identify the two-fold effects (disease at present and the progression) for disease forecasting. The high-order residual information is employed to achieve a more accurate prediction result. Equipped with a recurrent neural network in our temporal generative model, the disease-related region is localized accurately. In the future, we will apply our method to other irreversible diseases, such as Alzheimer's Disease.

6. Acknowledgements

This work was supported in part by MOST-2018AAA0102004, NSFC-61625201, NSFC-62061136001, NSFC-82071000, Beijing Natural Science Foundation (JQ20029), Capital health research and development of special (2020-2-1081).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Yidong Chai, Hongyan Liu, and Jie Xu. A new convolutional neural network model for peripapillary atrophy area segmentation from retinal fundus images. *Applied Soft Computing*, 86:105890, 2020.

[4] Ruoxuan Cui, Manhua Liu, Alzheimer's Disease Neuroimaging Initiative, et al. Rnn-based longitudinal analysis for diagnosis of alzheimer's disease. *Computerized Medical Imaging and Graphics*, 73:1–10, 2019.

[5] Ruoxuan Cui, Manhua Liu, and Gang Li. Longitudinal analysis for alzheimer's disease diagnosis using rnn. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1398–1401. IEEE, 2018.

[6] Ellen BM Elsman, Mo Al Baaj, Gerardus HMB van Rens, Wencke Sijbrandi, Ellen GC van den Broek, Hilde PA van der Aa, Wouter Schakel, Martijn W Heymans, Ralph de Vries, Mathijs PJ Vervloed, et al. Interventions to improve functioning, participation, and quality of life in children with visual impairment: a systematic review. *survey of ophthalmology*, 64(4):512–557, 2019.

[7] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208, 2010.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[11] Brien A Holden, Timothy R Fricke, David A Wilson, Monica Jong, Kovin S Naidoo, Padmaja Sankaridurg, Tien Y Wong, Thomas J Naduvilath, and Serge Resnikoff. Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology*, 123(5):1036–1042, 2016.

[12] Martha Kim et al. Longitudinal changes of optic nerve head and peripapillary structure during childhood myopia progression on oct: Boramae myopia cohort study report 1. *Ophthalmology*, 125(8):1215–1223, 2018.

[13] Tae-Woo Kim, Martha Kim, Robert N Weinreb, Se Joon Woo, Kyu Hyung Park, and Jeong-Min Hwang. Optic disc change with incipient myopia of childhood. *Ophthalmology*, 119(1):21–26, 2012.

[14] Hanxiang Li et al. Automatic detection of parapapillary atrophy and its association wif children myopia. *Computer methods and programs in biomedicine*, 183:105090, 2020.

[15] Maxime Louis, Raphael Couronne, Igor Koval, Benjamin Charlier, and Stanley Durrleman. Riemannian geometry learning for disease progression modelling. In *International Conference on Information Processing in Medical Imaging*, pages 542–553. Springer, 2019.

[16] Cheng-Kai Lu, Tong Boon Tang, Augustinus Laude, Ian J Deary, Baljean Dhillon, and Alan F Murray. Quantification of parapapillary atrophy and optic disc. *Investigative ophthalmology & visual science*, 52(7):4671–4677, 2011.

[17] Cheng-Kai Lu, Tong Boon Tang, Augustinus Laude, Baljean Dhillon, and Alan F Murray. Parapapillary atrophy and optic disc region assessment (pandora): retinal imaging tool for assessment of the optic disc and parapapillary atrophy. *Journal of biomedical optics*, 17(10):106010, 2012.

[18] Serge Resnikoff, Donatella Pascolini, Daniel Etya'Ale, Ivo Kocur, Ramachandra Pararajasegaram, Gopal P Pokharel, and Silvio P Mariotti. Global data on visual impairment in the year 2002. *Bulletin of the world health organization*, 82:844–851, 2004.

[19] David Edmundo Romo-Bucheli, Ursula Schmidt-Erfurth, and Hrvoje Bogunovic. End-to-end deep learning model for predicting treatment requirements in neovascular amd from longitudinal retinal oct imaging. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[21] Daniel Schmitter, Alexis Roche, Bénédicte Maréchal, Delphine Ribes, Ahmed Abdulkadir, Meritxell Bach-Cuadra, Alessandro Daducci, Cristina Granziera, Stefan Klöppel, Philippe Maeder, et al. An evaluation of volume-based morphometry for prediction of mild cognitive impairment and alzheimer's disease. *NeuroImage: Clinical*, 7:7–17, 2015.

[22] Min Kyung Song, Kyung Rim Sung, Joong Won Shin, Junki Kwon, Ji Yun Lee, and Ji Min Park. Progressive change in peripapillary atrophy in myopic glaucomatous eyes. *British Journal of Ophthalmology*, 102(11):1527–1532, 2018.

[23] Christopher C Teng, Carlos Gustavo V De Moraes, Tiago S Prata, Celso Tello, Robert Ritch, and Jeffrey M Liebmann. β-zone parapapillary atrophy and the velocity of glaucoma progression. *Ophthalmology*, 117(5):909–915, 2010.

[24] Xiaoqian Wang, Weidong Cai, Dinggang Shen, and Heng Huang. Temporal correlation structure learning for mci conversion prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 446–454. Springer, 2018.

[25] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Attention residual learning for skin lesion classification. *IEEE transactions on medical imaging*, 38(9):2092–2103, 2019.

[26] Zhuo Zhang et al. Automatic glaucoma diagnosis with mrmr-based feature selection. *J Biomet Biostat S*, 7:2, 2012.