

Appendix for Forecasting Irreversible Disease via Progression Learning

1. Dataset

The data contains 905 participants in primary school, in which there are 400 males and 505 females. The attributions we used are some clinical information selected from the whole medical examination results of the students: age, gender, height, myopia situation of parents, time for outdoor activities, axial length (AXL). The age, gender, height, and myopia situation of parents are obtained from an epidemiological survey for the students. Especially, the myopia situation of parents means the number of people wearing glasses in parents for each student, of which the value is defined as 0/1/2. AXL is measured by Intraocular lens (IOL) master, and the time for outdoor activities is obtained from a tracking measurement for each student, of which the unit is hours per week.

2. Prove of Proposition

2.1. 1-order Case

Note that the $P(y_T = 1|z_{t_1})$ can be expanded by total probability formula.

$$\begin{aligned}
 p(y_T = 1|z_{t_1}) &= p(y_T = 1, y_{t_1} = 1|z_{t_1}) + p(y_T = 1, y_{t_1} = 0|z_{t_1}) \\
 &= p(y_{t_1} = 1|z_{t_1})p(y_T = 1|y_{t_1} = 1, z_{t_1}) + p(y_{t_1} = 0|z_{t_1})p(y_T = 1|y_{t_1} = 0, z_{t_1}) \\
 &= p(y_{t_1} = 1|z_{t_1}) + p(y_{t_1} = 0|z_{t_1}) * p(y_T = 1|y_{t_1} = 0, z_{t_1}) \\
 &= p(y_{t_1} = 1|z_{t_1}) + [(1 - p(y_{t_1} = 1|z_{t_1})) * p(y_T = 1|y_{t_1} = 0, z_{t_1})],
 \end{aligned} \tag{1}$$

where the third equality is due to assumption that $p(y_T = 1|y_{t_1} = 1, z_{t_1}) = 1$ for $t_1 < T$.

The term of $p(y_T = 1|y_{t_1} = 0, z_{t_1})$ can be computed as:

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

The Eq. 1 shows that $p(y_T = 1|z_{t_1}) \geq p(y_{t_1} = 1|z_{t_1})$ for $t_1 < T$, agreeing with the *Deterioration* principle in main body. Besides, the “progression” of the disease can

be obtained by Eq. 2, in which $p(y_T = 1|y_{t_1} = 0, \mathbf{x}_T, z_{t_1})$ describes the extent of progression from the healthy status, we propose to approximate it using progression information which contains *i.e.*, $\mathbf{x}_T - \mathbf{x}_{t_1}$.

2.2. High-Order Case

Proposition 2.1 *Under the irreversibility principle, we have the following factorization for progression prediction on K -order setting ($K > 1$):*

$$\begin{aligned}
 p(y_T = 1|z_{t_{1:K}}) &= \underbrace{p(y_{t_K} = 1|z_{t_{1:K}})}_{\text{Current}} + \underbrace{p(y_{t_K} = 0|z_{t_{1:K}})p(y_T = 1|y_{t_K} = 0, z_{t_{1:K}})}_{\text{Progression}}.
 \end{aligned} \tag{3}$$

Proof of Proposition 2.1 Note that the $P(y_T = 1|z_{t_{1:K}})$ can be expanded by total probability formula.

$$\begin{aligned}
 p(y_T = 1|z_{t_{1:K}}) &= p(y_T = 1, y_{t_K} = 1|z_{t_{1:K}}) + p(y_T = 1, y_{t_K} = 0|z_{t_{1:K}}) \\
 &= p(y_{t_K} = 1|z_{t_{1:K}})p(y_T = 1|y_{t_K} = 1, z_{t_{1:K}}) + p(y_{t_K} = 0|z_{t_{1:K}})p(y_T = 1|y_{t_K} = 0, z_{t_{1:K}}) \\
 &= p(y_{t_K} = 1|z_{t_{1:K}}) + p(y_{t_K} = 0|z_{t_{1:K}}) * p(y_T = 1|y_{t_K} = 0, z_{t_{1:K}}) \\
 &= p(y_{t_K} = 1|z_{t_{1:K}}) + [(1 - p(y_{t_K} = 1|z_{t_{1:K}})) * p(y_T = 1|y_{t_K} = 0, z_{t_{1:K}})],
 \end{aligned} \tag{4}$$

where the third equality is due to assumption that $p(y_T = 1|y_{t_{1:K}} = 1, z_{t_{1:K}}) = 1$ for $t_1 < \dots < t_K < T$.

The term of $p(y_T = 1|y_{t_K} = 0, z_{t_{1:K}})$ can be computed as:

$$\begin{aligned}
 p(y_T = 1|y_{t_K} = 0, z_{t_{1:K}}) &= \int_{\mathbf{x}_T} p(\mathbf{x}_T|y_{t_K} = 0, z_{t_{1:K}}) * p(y_T = 1|y_{t_K} = 0, \mathbf{x}_T, z_{t_{1:K}})d\mathbf{x}_T.
 \end{aligned} \tag{5}$$

On high-order settings, the Eq. 3 shows that $p(y_T = 1|z_{t_{1:K}}) \geq p(y_{t_K} = 1|z_{t_{1:K}})$ for $t_1 < \dots < t_K < T$, also agreeing with the *Deterioration* principle in main body. Besides, the “progression” of the disease can be obtained by Eq. 5. In above equations, $p(y_T = 1|y_{t_K} = 0, \mathbf{x}_T, z_{t_{1:K}})$ describes the extent of progression from the healthy status.

Table 1. Ablation study on 2-order setting, to validate the effectiveness of each module. The Eq.(3) means that we train the model with high-order loss and predict by $f_{\text{cur}} + (1 - f_{\text{cur}})f_{\text{prog}}$. “MA” stands for Model Average with $f_{\text{cur}}(\tilde{F}_T)$, Eq.(3) and $f_{\text{fut}}(F_{t_i})$. f_{cur} denotes that we train the model with $\mathcal{L}_{\text{gen}} + \lambda_1 \mathcal{L}_{\text{cur}}$ and predict by $f_{\text{cur}}(\tilde{F}_T)$. f_{prog} denotes that we train the model with $\mathcal{L}_{\text{gen}} + \lambda'_2(-\log p_{f_{\text{prog}}}(y_T | \text{prog}_K(\tilde{F}_T, \{F_{t_j}\}_{j \in [K]}), \mathbf{a}_{t_{1:K}}))$ and predict by $f_{\text{prog}}(\text{prog}_K(\tilde{F}_T, \{F_{t_j}\}_{j \in [K]}), \mathbf{a}_{t_{1:K}})$.

Predictor	LSTM	MA	$\delta t=4$		$\delta t=3$		$\delta t=2$		$\delta t=1$		Average	
			ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
f_{cur}	✓	×	70.17	76.56	66.85	75.80	72.93	83.27	76.80	85.99	71.69	80.40
f_{prog}	✓	×	73.48	77.76	69.06	77.17	74.03	84.01	78.45	88.44	73.76	81.84
Eq.(3)	✓	×	69.06	77.63	71.82	78.83	74.59	83.29	81.22	90.50	74.17	82.56
Eq.(3)	×	✓	69.06	76.90	70.17	77.77	75.69	85.15	81.77	90.35	74.17	82.54
Eq.(3)	✓	✓	70.72	79.33	76.24	80.82	79.01	86.36	80.11	91.14	76.52	84.41

Table 2. Ablation study on 3-order setting, to validate the effectiveness of each module. The Eq.(3) means that we train the model with high-order loss and predict by $f_{\text{cur}} + (1 - f_{\text{cur}})f_{\text{prog}}$. “MA” stands for Model Average with $f_{\text{cur}}(\tilde{F}_T)$, Eq.(3) and $f_{\text{fut}}(F_{t_i})$. f_{cur} denotes that we train the model with $\mathcal{L}_{\text{gen}} + \lambda_1 \mathcal{L}_{\text{cur}}$ and predict by $f_{\text{cur}}(\tilde{F}_T)$. f_{prog} denotes that we train the model with $\mathcal{L}_{\text{gen}} + \lambda'_2(-\log p_{f_{\text{prog}}}(y_T | \text{prog}_K(\tilde{F}_T, \{F_{t_j}\}_{j \in [K]}), \mathbf{a}_{t_{1:K}}))$ and predict by $f_{\text{prog}}(\text{prog}_K(\tilde{F}_T, \{F_{t_j}\}_{j \in [K]}), \mathbf{a}_{t_{1:K}})$.

Predictor	LSTM	MA	$\delta t=3$		$\delta t=2$		$\delta t=1$		Average	
			ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
f_{cur}	✓	×	67.96	76.59	72.93	82.65	77.35	86.82	72.74	82.02
f_{prog}	✓	×	68.51	77.58	67.96	81.34	64.09	82.87	66.85	82.87
Eq.(3)	✓	×	71.27	78.80	74.38	84.14	80.11	88.32	75.14	83.75
Eq.(3)	×	✓	70.72	79.90	76.80	85.65	79.56	88.72	75.69	84.75
Eq.(3)	✓	✓	73.48	81.92	76.24	86.30	78.45	88.58	76.06	85.60

Especially, as introduced in Sec 3.3, the high-order progression information is defined as $\text{prog}_K(\tilde{F}_T, \{F_{t_j}\}_{j \in [K]})$, which includes the first-order information and the second-order information.

3. Ablation Study on High-Order Setting

Similar to the 1-order situation, the ablation study is also be implemented to validate the effectiveness of each module on high-order settings. The results are showed in Tab. 1 and 2. The 3rd row over the first two rows in both two tables has shown the improvement of Eq. 3 on high-order settings, from which the effectiveness of $f_{\text{cur}}, f_{\text{prog}}$ in the disease forecast is validated. Besides, in both tables, the effectiveness of LSTM is displayed on the difference of the 5th row and 4th row. As same as the 1-order results, further improvement is also achieved on high-order settings, which is showed from the comparison of the 5th row and 3rd row.

4. About experiment and visualization

The p -value of hypothesis test is $< 1e^{-4}$ comparing ours with RN18, $< 1e^{-3}$ with MM-F, $< 1e^{-4}$ with ARL and $< 1e^{-4}$ with TCLS. As shown in Fig. 1, our model is insensitive to each hyperparameter if it belongs to a reasonable range.

We randomly select one feature map among 128 feature maps for visualization, based on our observation of the high similarity among all 128 feature maps: the average pairwise cosine similarity ($\in [0, 1]$) among 128 feature maps is 0.909. Here we visualize a healthy case for comparison in Fig.

2. As shown, the generated future feature map $\tilde{F}_T(F_{t_1})$ is similar to the future one F_T . Compared to the diseased ones, the high response area in the residual feature map is smaller. The “ground-truth” is denoted as the residual feature map extracted by the pretrained feature extractor $F(x_T) - F(x_t)$. The high response area in our estimated residual feature map, although includes other regions, has a high overlap with the residual feature map extracted by the pretrained model in the optic disc region. In other words, it captures information of the disease-related regions. We will modify our words to make the above descriptions more clear and accurate.

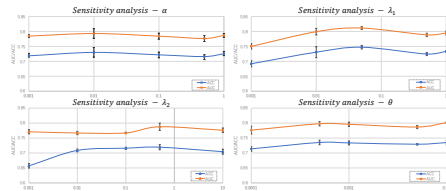


Figure 1. Sensitivity analysis of hyperparameters on 1-order setting.

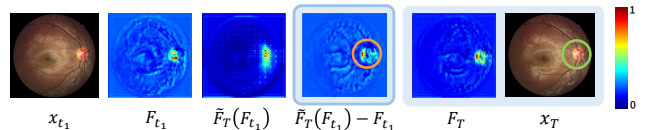


Figure 2. Visualization for a healthy case on 1-order setting.