This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/156941/

This is the author’s version of a work that was submitted to / accepted for publication.

Citation for final published version:

Publishers page: http://dx.doi.org/10.1109/TMI.2023.3243692

Please note: Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
GLIM-Net: Chronic Glaucoma Forecast Transformer for Irregularly Sampled Sequential Fundus Images

Xiaoyan Hu, Ling-Xiao Zhang, Lin Gao, Member, IEEE, Weiwei Dai, Xiaoquang Han, Yu-Kun Lai, Member, IEEE, and Yiqiang Chen, Senior Member, IEEE

Abstract—Chronic Glaucoma is an eye disease with progressive optic nerve damage. It is the second leading cause of blindness after cataract and the first leading cause of irreversible blindness. Glaucoma forecast can predict future eye state of a patient by analyzing the historical fundus images, which is helpful for early detection and intervention of potential patients and avoiding the outcome of blindness. In this paper, we propose a GLaucoma forecast transformer based on irregularly sampled fundus images named GLIM-Net to predict the probability of developing glaucoma in the future. The main challenge is that the existing fundus images are often sampled at irregular times, making it difficult to accurately capture the subtle progression of glaucoma over time. We therefore introduce two novel modules, namely time positional encoding and time-sensitive MSA (multi-head self-attention) modules, to address this challenge. Unlike many existing works that focus on prediction for an unspecified future time, we also propose an extended model which is further capable of prediction conditioned on a specific future time. The experimental results on the benchmark dataset SIGF show that the accuracy of our method outperforms the state-of-the-art models. In addition, the ablation experiments also confirm the effectiveness of the two modules we propose, which can provide a good reference for the optimization of Transformer models.

Index Terms—Glaucoma forecast, transformer, attention mechanism, fundus image.

This work was supported in part by the Science and Technology Service Network Initiative of the Chinese Academy of Sciences (No. KFJ-STS-QYZD-2021-11-001), in part by the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), in part by the Royal Society Newton Advanced Fellowship (No. NAF R2,192151), in part by the Science and Technology Innovation Program of Hunan Province (No. 2021GK4015) and in part by the Youth Innovation Promotion Association CAS. Xiaoyan Hu and Ling-Xiao Zhang contributed equally to this work.

Corresponding authors are Yiqiang Chen (yqchen@ict.ac.cn) and Lin Gao (gaolin@ict.ac.cn).

Xiaoyan Hu, Ling-Xiao Zhang, Lin Gao and Yiqiang Chen are with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences. Lin Gao and Yiqiang Chen are also with University of Chinese Academy of Sciences, Beijing, China (e-mail: yxhhxy1999@163.com; {zhangxingxiao, gaolin, yqchen}@ict.ac.cn).

Weiwei Dai is with the Changsha Aier Eye Hospital, Changsha, Hunan, China (e-mail: daiweiwei@aierchina.com).

Xiaoquang Han is with the SSE, The Chinese University of Hong Kong, Shenzhen, China (e-mail: hanxiaoquang@cuhk.edu.cn).

Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Wales, UK (e-mail: LaiY4@cardiff.ac.uk).

I. INTRODUCTION

Glaucoma is the collective term for a group of diseases that result in progressive damage to the optic nerve and causes loss of vision, primarily associated with pathological intraocular pressure elevation. According to a 2014 global meta-analysis [1], the prevalence of glaucoma for population aged 40–80 years is 3.5%, or approximately 64.3 million people. The number of people with glaucoma is expected to increase to 112 million by 2040 due to population growth and aging.

Many early works [2]–[5] proposed measurement-based methods for automatic screening of glaucoma, which first segmented the pathological regions, such as optic disc and optic cup, and then calculated the relevant clinical values for glaucoma diagnosis. However, the accuracy of glaucoma screening can be seriously affected by the segmentation results, which can be easily influenced by the pathological regions and low image quality.

Recently, with the booming of deep learning, effort has been increasingly devoted to utilizing deep learning methods for automatic glaucoma diagnosis. For example, references [6], [7] applied convolutional neural networks (CNNs) to extract features and detect glaucoma from fundus images directly. However, different from diseases such as cataracts and myopia, the loss of vision caused by glaucoma is irreversible, but most of the loss of vision caused by glaucoma can be avoided by early detection and treatment. Therefore, it is vital to detect the potential deterioration ahead of time for earlier intervention. Li et al. [8] first established a dataset of sequential fundus images, called SIGF, and proposed a deep learning approach (DeepGF) for glaucoma forecast. DeepGF primarily consists of a novel long short-term memory (LSTM) network to learn the spatial and temporal information from sequential fundus images of a person. However, DeepGF outputs the probability of developing glaucoma at the next time step, but cannot specify when the next time is. Furthermore, LSTM may suffer from information loss when passing information from previous steps to the current one and cannot process data in parallel.

In this paper, we propose the first transformer-based glaucoma forecast network named GLIM-Net for irregularly sampled sequential fundus images. There are two main challenges that need to be addressed. First, the transformer architecture lacks inductive bias like convolutions and thus requires a large
amount of training data [9], but the scale of the SIGF dataset is relatively small. In addition, the images in SIGF dataset were captured at irregular times as it is impractical to request patients to take medical examinations on a regular basis. The time interval of fundus images in a sequence varies from the minimum of only one day to the maximum of 13 years, which will be a challenge for models to learn the status transition over time. To overcome the first challenge, we first replace the input embedding method of simple linear transformation, which was used in recent works of vision transformers [10], [11], with the polar convolutional neural network used in [8] to extract low-level features to enable the model to converge even if the scale of dataset is small. Transformer is designed on the hypothesis that the intervals between samples in a sequence are the same. So to overcome the second challenge, we further propose two novel modules i.e., time positional encoding module and time-sensitive MSA (multi-head self-attention) module. We redesign the positional encoding used in [12] and propose a new time positional encoding (TPE) to enable our model to effectively learn the time distribution of sequential fundus images, which are not sampled regularly. To make full use of the prior knowledge that the longer time the fundus image is captured from now, the less impact the fundus image will have on the current diagnosis, we propose a novel time-sensitive MSA to learn more effective temporal features. Forecasting the probability of developing glaucoma conditioned on a specific time is more meaningful and useful than that of developing glaucoma for an unspecified time, as it provides richer information to make more informed decisions for monitoring, early detection and further timely treatment to slow or halt the glaucoma progression. We therefore introduce an extended model capable of predicting glaucoma for a specific time by feeding the time condition into both input and output labels. Extensive experiments on SIGF dataset [8] demonstrate the effectiveness of our proposed model and our two modules.

Our main contributions can be summarized as follows:

- We propose the first transformer-based glaucoma forecast network named GLIM-Net with time positional encoding and time-sensitive MSA modules to better address the irregularly sampled data.
- We evaluate our model on SIGF dataset and experimental results demonstrate our GLIM-Net achieves better performance than other state-of-the-art methods with a remarkable margin.
- We extend our model to be able to predict glaucoma conditioned on a specific time, which addresses a problem that cannot be handled by existing works, and experiments show the effectiveness of the extended model.

II. RELATED WORK

A. Early Prediction of Diseases

Early prediction of disease deterioration can help clinicians to better treat patients. It is estimated that 11% of patient deaths followed a failure of swift recognition and treatment [13]. Recently, many deep learning methods have been proposed for early prediction of diseases, such as predicting chronic lung disease, acute kidney injury, heart disease, Alzheimer’s disease (AD), dementia and glaucoma.

Cheng et al. [14] used convolutional neural networks (CNNs) and temporal fusion mechanisms to analyze electronic health records (EHRs) and predict the probability of suffering from chronic lung disease in the future. Tomavsev et al. [15] proposed a recurrent neural network that processes sequential EHR of patients and outputs a probability of acute kidney injury occurring at any stage of severity within the next 48 hours. Ali et al. [16] first extracted features from both sensor data and EHR, and combined them using a feature fusion method. An ensemble deep learning model was then implemented to perform heart disease prediction.

The above methods are based on EHRs, but the high-dimensionality, sparsity and irregularity of EHR [14], [17] severely affect the accuracy of prediction. Li et al. [18] proposed a deep learning method to implicitly extract features from hippocampal magnetic resonance imaging (MRI) data and established a time-to-event prognostic model to predict the progression of subjects who meet criteria for mild cognitive impairment to Alzheimer’s Disease (AD) dementia. Li et al. [8] used sequential fundus images that can better show the subtle pathological features of glaucoma to address the prediction of glaucoma. First, a CNN was applied to extract low-level features, which were then inputted into a long short-term memory (LSTM) network [19] that can better capture temporal features, and finally the probability of developing glaucoma in the future was outputted.

B. Transformers in Vision

Transformers [12] were first proposed for machine translation tasks and have made great success in many natural language processing (NLP) tasks, such as BERT [20], GPT [21], [22], and XLNet [23]. In recent years, self-attention mechanism has been extensively explored to build long-range dependencies. Reference [24] showed that attention modules can completely replace convolution operations and the study reported in [25] showed that self-attention layers deal with pixel-grid patterns analogously to CNN layers. Recently, many works have successfully applied self-attention to various vision tasks, such as image classification, object detection and image segmentation. DETR [26] regarded object detection as a direct set prediction task and reasoned about the relationships of the objects and the global image context to yield the final set of predictions straightforwardly in parallel. In ViT (Vision Transformer) [10], an image was split into patches and a pure transformer with MLP (Multi-Layer Perceptron) head was applied to perform image classification. On the basis of ViT, DeiT [11] proposed some training strategies and introduced the knowledge distillation (KD) [27] to make Transformer more efficient and achieved competitive results on small datasets. VisTR [28] regarded video instance segmentation as a direct coding/prediction task. With a video clip comprising a specified number of image frames as input, VisTR outputted a sequence of masks for individual instances in the video directly and in order.
The proposed network architecture is an encoder-decoder structure. The encoder inputs the embedding feature of a sequence of fundus images, which is then added with time positional encoding proposed to obtain temporal information. The decoder inputs the encoded feature and the embedding feature of the labels (0 for negative and 1 for positive) added with time positional encoding, and outputs the probability distribution of glaucoma. TMSA (Time-sensitive multi-head self-attention) is designed to make full use of the prior knowledge that the longer time the fundus image is captured from now, the less impact the fundus image will have on the current diagnosis.\( \Delta t_{i,j} \) denotes the time offset of the \( i \)-th fundus image from the \( j \)-th fundus image in the sequence. More details of the TMSA module are shown in Fig. 2. This illustration is based on taking the fundus images captured in the 5 visits as input and predict the probability of developing glaucoma at the time of the 6th visit, to match the setting of previous work [8]. Our method is general and can be applied to different numbers of images in the input sequence.

Some other studies [29]–[32] tried to apply self-attention mechanisms to the medical imaging field. Xiong et al. [29] proposed a reinforcement learning method, namely reinforced transformers for medical image captioning (RTMIC), to generate long and coherent medical imaging reports. Inspired by DETR, Prangemeier et al. [30] proposed a novel attention-based detection transformer called cell detection transformer (Cell-DETR) for faster cell instance segmentation. Valanarasu et al. [31] introduced a modified gated axial-attention by presenting an additional control mechanism and incorporated it as the building block of multi-head attention models for medical image segmentation. Cao et al. [32] introduced a Transformer-based U-shaped Encoder-Decoder architecture with skip-connections to extract local-global semantic features for medical image segmentation. However, most of these previous works cannot perform well on irregularly sampled data. In this paper, we propose two modules, namely time positional encoding and time-sensitive self-attention, to make full use of temporal information and enable the transformer to achieve good results on irregularly sampled data.

III. Method

In this section, we introduce our GLIM-Net in detail. We first discuss the vanilla transformer and our network architecture, and then introduce the two novel modules, time positional encoding and time-sensitive MSA, as well as the loss function. Finally, we demonstrate the extended model with the capability of forecasting glaucoma conditioned on a specific time.

A. Revisiting Transformer Architecture

The vanilla Transformer [12] is an encoder-decoder structure with stacked encoder and decoder layers. Positional encodings are added to the embedded input prior to the first layer of encoder and decoder to make use of the order of the input sequence. Each encoder layer is composed of two sub-layers i.e., a multi-head self-attention mechanism followed by a fully connected feed-forward network. In addition to the two sub-layers of encoder layer, the decoder layer has an additional masked multi-head self-attention mechanism before the two sub-layers to ensure that the prediction of time \( t_i \) can only depend on the known data at positions before \( t_i \). A residual connection [33] followed by layer normalization [34] is employed around each pair of sub-layers.

B. GLIM-Net Framework

The architecture of GLIM-Net is illustrated in Fig. 1, which is also an encoder-decoder structure. For a given sequence consisting of \( n \) fundus images, each image is denoted by \( x_i, i \in [1, ..., n] \), and has a corresponding time stamp \( t_i \) and label \( y_i \in \{0, 1\} \), 0 for negative glaucoma and 1 for positive glaucoma. The output of GLIM-Net \( \hat{y}_{i+1} \) is the probability of positive glaucoma of the image \( x_{i+1} \). We denote the GLIM-Net by \( \mathcal{F} \), and \( \hat{y}_{i+1} = \mathcal{F}_{i} = \mathcal{F}(x_1, x_2, ..., x_i, y_1, y_2, ..., y_i) \). The final output is \( y_{n+1} \) corresponding to the \( n \)-th fundus image \( x_{n+1} \) captured in the next visit so unknown to the algorithm but used by clinicians to determine the ground truth label \( y_{n+1} \). Specifically, we first apply a polar transformation [35] and a convolution network [8] on the fundus image \( x_i \) to extract low-level features \( \tilde{f}_i \). Then, the time positional encoding (see Sec. III-C) is added to \( \tilde{f}_i \) to form \( \tilde{f}_i \in \mathbb{R}^{d_m} \) as the input of the encoder. The encoder consists of \( N \) identical layers. We set \( d_m = 512 \) and \( N = 6 \) in our paper. Similar to the work [12], each layer is composed of two sub-layers, i.e., time-sensitive MSA and multilayer perceptron, and a residual connection [33] followed by layer normalization [34] is employed around each pair of sub-layers.
masked time-sensitive MSA to ensure that the prediction of time \( t_i \) can only depend on the previous known data.

### C. Time Positional Encoding

The self-attention module in the original Transformer [12] is permutation invariant, so it needs positional encoding to combine input embedding and position information to enable the model to have the ability to learn the order of the sequence. Positional encoding can either be fixed or learnable, and either absolute or relative. A fixed sinusoidal absolute positional encoding is proposed in [12]. Based on sinusoidal positional encoding, we propose a time positional encoding (TPE) that can make better use of the temporal information of irregularly sampled data. TPE can be expressed by the following

\[
\text{TPE}(\Delta t_{i,1}, q) = \begin{cases} 
\sin(\omega_q \cdot \Delta t_{i,1}), & q = 2k \\
\cos(\omega_q \cdot \Delta t_{i,1}), & q = 2k + 1
\end{cases} \tag{1}
\]

where \( \Delta t_{i,1} \) is the time offset of the current fundus image \( x_i \) from the first fundus image \( x_1 \) in the sequence and \( \omega_q = 1/10000^{2q/d_m}, q \in [0, d_m] \). The wavelengths of functions increase from \( 2\pi \) to 10000 \(- 2\pi\). To make encoding richer, sin and cos functions are used alternately. The modified function can not only learn the relative position relationship of sequential fundus images, but also learn the time distance relationship between fundus images in the sequence.

### D. Time-Sensitive MSA

References [10], [11], [28] have proved that self-attention performs well in capturing long-range dependencies in the vision field. However, the data of these works is all regularly sampled, and thus models are designed based on the hypothesis that the intervals between samples in the sequence are equal, which makes self-attention sensitive to the position relationship of data but not sensitive to the time relationship of data. In this paper, we propose a novel time-sensitive self-attention module to enable the model to learn the time distance dependencies of sequential fundus images. Also we extend our time-sensitive self-attention to time-sensitive MSA to allow the model to attend to features from different representation subspaces.

#### 1) Time-Sensitive Self-Attention

On the basis of the self-attention mechanism proposed by Vaswani et al. [12], we propose a time-sensitive self-attention mechanism to tackle the variable time interval problem of this task, as illustrated in Fig. 2 (a), which can be expressed by the following

\[
G(Q, K, T, V) = \text{softmax}(\frac{QK^T \circ T}{\sqrt{d_m}})V \tag{2}
\]

\[
Q = \tilde{f}W^Q \tag{3}
\]

\[
K = \tilde{f}W^K \tag{4}
\]

\[
V = \tilde{f}W^V \tag{5}
\]

where \( \tilde{f} \) is the result of the input embedding, \( \tilde{f} \in \mathbb{R}^{n \times d_m} \). We obtain the queries \( Q \in \mathbb{R}^{n \times d_q} \), keys \( K \in \mathbb{R}^{n \times d_k} \) and values \( V \in \mathbb{R}^{n \times d_v} \) by multiplying \( \tilde{f} \) with learnable weight matrices

![Fig. 2. (a) Time-Sensitive Self-Attention](image-url)
Conditional Glaucoma Forecast Network

Forecasting the probability of developing glaucoma conditioned on a specific time is more meaningful and useful than that of developing glaucoma for an unspecified time, which provides richer information to help make more informed decisions for monitoring, early detection and further timely treatment to slow or halt the glaucoma progression. Therefore, we extend our GLIM-Net to conditional GLIM-Net to assist people who may develop glaucoma in a few years to take measures to prevent further vision loss. Specifically, the time condition is incorporated in both the encoder and the decoder as additional input. The architecture of conditional GLIM-Net is shown in Fig. 3. We denote C-GLIM-Net by $\mathcal{F}$, and $\hat{y}_{j,i} = \mathcal{F}_{j,i} = \mathcal{F}(x_1, x_2, \ldots, x_i, y_1, y_2, \ldots, y_j | \Delta t_{j,i})$ means that the network inputs a sequence consisting of $i$ fundus images and outputs the probability of glaucoma of $x_j$ ($j > i$). Different from GLIM-Net, the last layer of the convolution network [8] is changed to output a $(d_m - 1)$-dimension feature and it is concatenated with the time interval $\Delta t_{j,i}$ as the input of the encoder, and the embedding feature of the label is also changed to $(d_m - 1)$-dimension and is also concatenated with the time interval $\Delta t_{j,i}$ as the input of decoder. The decoder last outputs the probability of positive glaucoma $\hat{y}_{j,i}' \in [0,1]$.

To train the C-GLIM-Net, we need first $i$ images of a sequence to predict the $\hat{y}_{i+1}'$ and $\hat{y}_{i+2}'$, under the conditions $\Delta t_{i+1,i}$ and $\Delta t_{i+2,i}$. The loss function is modified to

$$L_C = -\frac{1}{n} \sum_{i=1}^{n} (y_{i+1} \log(\hat{y}_{i+1,i}') + (1 - y_{i+1}) \log(1 - \hat{y}_{i+1,i}')) + \frac{1}{n - 1} \sum_{i=1}^{n-1} (y_{i+2} \log(\hat{y}_{i+2,i}') + (1 - y_{i+2}) \log(1 - \hat{y}_{i+2,i}'))$$

(10)

We conduct experiments of C-GLIM-Net model on the SIGF dataset by forecasting glaucoma conditioned on the time of the next two fundus images in the sequence simultaneously. Experimental results are detailed in the next section.

IV. EXPERIMENTS AND RESULTS

A. Datasets

To evaluate the effectiveness of our proposed method, we first carry out experiments on a dataset consisting of sequential fundus images annotated with either positive or negative glaucoma named SIGF [8]. We further prove the effectiveness of our proposed method on a dataset named Tumor-CIFAR proposed by [36] that shares the same characteristics as the SIGF dataset.

1) SIGF: The SIGF contains 3671 fundus images in total and consists of 405 sequential fundus images from different eyes with an average of 9 images per eye ranging from 1986 to 2018. In the dataset, there are at least 6 fundus images for each eye. All the fundus images are annotated with positive glaucoma when they satisfy any of the three criteria, i.e., retinal nerve fiber layer defect, rim loss and optic disc hemorrhage. The sequences are divided into 2 types: time-variant and time-invariant as defined in [8]. Time-variant sequences are those that change from negative to positive
The fundus image in the Cartesian coordinate system (a).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Time-variant</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>27</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Validation</td>
<td>273</td>
<td>32</td>
<td>63</td>
</tr>
</tbody>
</table>

The parameters of time-sensitive MSA, A and B, are initialized as 0.9 and 2.0, respectively, according to our extensive experiments. The batch size is set to 4. Our model can gradually achieve the approximate global optimum after 400 epochs of iterations.

B. Experimental Setup

1) Implementation Details: We implement our GLIM-Net with Tensorflow [38] on a single NVIDIA TITAN X (Pascal) GPU with 12 GB memory. We employ the Adam algorithm [39] to update the parameters of our model with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1e-8$. Considering that the data distribution of SIGF dataset is extremely imbalanced, we adopt the active convergence strategy used in [8] to self-update the distribution of the training set actively and adaptively, i.e., discarding the sequences the training loss of which ranks the lowest. Specifically, we set learning rate $\tau_r = \{3 \times 10^{-8}, 10^{-7}, 6 \times 10^{-6}, 6 \times 10^{-5}, 10^{-5}, 5 \times 10^{-7}\}$ each for 5 epochs. The parameters of time-sensitive MSA, A and B, are initialized as 0.9 and 2.0, respectively, according to our extensive experiments. The batch size is set to 4. Our model can gradually achieve the approximate global optimum after 400 epochs of iterations.

To augment the data set and avoid the overfitting problem, following the work [8], we segment the 405 sequences into 1146 clips, with each clip comprising $T = 6$ time steps and an overlap of 5 frames is allowed in segmenting the sequences into clips. Before being input into the model, the fundus images are reshaped to $224 \times 224$ and a polar transformation is applied on them in order to enlarge the disc and cup structure, as illustrated in Fig. 4, which can be expressed by the following

$$r = \sqrt{u^2 + v^2},$$

$$\theta = \tan^{-1}(\frac{u}{v}),$$

where $(u, v)$ denotes a point in the original Cartesian coordinate system, as shown in Fig. 4 (a), $r$ and $\theta$ are the radius and directional angle in the polar coordinate system, as shown in Fig. 4 (b).

2) Evaluation Metrics: To compare our method with other state-of-the-art methods, we adopt four commonly-used metrics to assess our GLIM-Net and [8], [36], [40]–[42] on SIGF, including accuracy, sensitivity, specificity and AUC (Area Under Curve).

C. Comparison with State-of-the-art Methods

1) Comparison on the SIGF dataset: We use the vanilla transformer as our Baseline method, because transformer can build long-range dependencies and shows good performance in processing sequential data. We compare our model with the baseline method and 8 other state-of-the-art methods to verify the superiority of our proposed method, namely DeepGF [8], Deep CNN [40], AG-CNN [41], tLSTM [42], DLSTM [36], MIL-VT [43], CABNet [44] and CoG-Net [45]. Notice that in DeepGF, they convert image classification models to prediction models by supervising the models with the labels at the next time step, so we similarly try to use two sequential image classification models, tLSTM [42] and DLSTM [36], the setting of which is more similar to the setting of our task than that of the image classification models. The MIL-VT, CABNet and CoG-Net are designed for fundus image classification and we convert them to be prediction models by supervising the models with the labels at the next time step as DeepGF [8] does.

Table II tabulates the results of our model, baseline method and other state-of-the-art-methods. As shown, our baseline
method performs better than 6 other state-of-the-art methods in terms of accuracy, sensitivity, specificity and AUC, and our modified model significantly improves the performance of baseline method. Our GLIM-Net model considerably outperforms 8 other state-of-the-art methods and the baseline method by achieving 89.5%, 87.6%, 89.6% and 93.6% in accuracy, sensitivity, specificity and AUC, respectively.

The loss, ACC and ROC (Receiving Operating Characteristic) curves of the training and validation phases are shown in Figure 5. (a) and (b) show the ROC curves of the training and validation phases. (c) presents the loss curves and (d) shows the ACC curves of the training and validation phases. The blue curve is the result of training phases, and orange curve is the result of validation phases.

D. Ablation Study

To demonstrate the effectiveness of our each module, we conduct ablation studies on the convolutional input embedding, time positional encoding, time-sensitive MSA, and the setting of threshold $δ$.

1) Convolutional Input Embedding (CIE): We first evaluate the impact of convolutional input embedding compared to the simple linear transformation by flattening the original fundus image $x_{img} \in \mathbb{R}^{3 \times H \times W}$ to a vector $x'_{img} \in \mathbb{R}^{3HW}$ and then using linear transformation to embed it to $d_{in}$ dimensions. As indicated in Table IV, the convolutional input embedding can improve accuracy and specificity by 4.2%. Also, sensitivity and AUC increase considerably from 83.0% and 90.0% to 87.6% and 93.6%, respectively. These results verify the rationality of replacing linear transformation with ConvNet as the input embedding method.

2) Time Positional Encoding (TPE): To demonstrate the superiority of our time positional encoding, we replace it with the vanilla positional encoding. As shown in Table IV, TPE improves the performance of glaucoma forecast in terms of all four evaluation metrics. Specifically, TPE improves the accuracy and specificity by 1.6% and 1.5%. Sensitivity and AUC reach 87.6% and 93.6% from 81.1% and 88.6%, respectively. These results verify that it is an effective module to make use of temporal information.

### Table II
Comparison with the state-of-the-art methods on SIGF. ACC, SE and SP denote the accuracy, sensitivity and specificity, respectively. Our method outperforms other methods in all four metrics with a considerable margin. For each measure, we show the mean and standard deviation.

<table>
<thead>
<tr>
<th>Method</th>
<th>ACC (%)</th>
<th>SE (%)</th>
<th>SP (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG-CNN</td>
<td>47.5 ± 0.5</td>
<td>57.5 ± 4.6</td>
<td>47.1 ± 0.7</td>
<td>51.8 ± 2.5</td>
</tr>
<tr>
<td>Deep CNN</td>
<td>67.0 ± 5.5</td>
<td>62.1 ± 1.8</td>
<td>67.2 ± 5.7</td>
<td>63.0 ± 3.3</td>
</tr>
<tr>
<td>DeepGF</td>
<td>76.0 ± 4.8</td>
<td>79.4 ± 1.3</td>
<td>75.9 ± 5.0</td>
<td>85.0 ± 2.5</td>
</tr>
<tr>
<td>tLSTM</td>
<td>84.7 ± 1.4</td>
<td>72.5 ± 0.0</td>
<td>85.1 ± 1.4</td>
<td>83.6 ± 0.6</td>
</tr>
<tr>
<td>DLSTM</td>
<td>87.2 ± 1.1</td>
<td>84.3 ± 1.6</td>
<td>87.3 ± 1.2</td>
<td>92.4 ± 0.7</td>
</tr>
<tr>
<td>MIL-VT</td>
<td>79.7 ± 1.1</td>
<td>77.8 ± 3.4</td>
<td>79.8 ± 1.2</td>
<td>83.4 ± 1.6</td>
</tr>
<tr>
<td>CAGNet</td>
<td>73.9 ± 1.6</td>
<td>74.4 ± 1.6</td>
<td>73.9 ± 1.6</td>
<td>78.7 ± 2.6</td>
</tr>
<tr>
<td>CoG-Net</td>
<td>77.0 ± 2.0</td>
<td>72.5 ± 3.6</td>
<td>77.2 ± 1.7</td>
<td>81.8 ± 3.5</td>
</tr>
<tr>
<td>Baseline</td>
<td>84.7 ± 0.2</td>
<td>77.8 ± 0.9</td>
<td>84.9 ± 0.1</td>
<td>86.1 ± 1.5</td>
</tr>
<tr>
<td>Ours</td>
<td>97.5 ± 0.8</td>
<td>87.6 ± 0.9</td>
<td>89.6 ± 0.8</td>
<td>93.6 ± 0.3</td>
</tr>
</tbody>
</table>

### Table III
Comparison on the Tumor-CIFAR dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>ACC (%)</th>
<th>SE (%)</th>
<th>SP (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLSTM</td>
<td>96.1 ± 0.3</td>
<td>96.0 ± 0.6</td>
<td>96.1 ± 0.7</td>
<td>99.3 ± 0.1</td>
</tr>
<tr>
<td>DeepGF</td>
<td>94.0 ± 0.1</td>
<td>94.1 ± 0.1</td>
<td>93.8 ± 0.2</td>
<td>98.6 ± 0.1</td>
</tr>
<tr>
<td>Ours</td>
<td>97.5 ± 0.6</td>
<td>97.1 ± 0.3</td>
<td>97.4 ± 0.8</td>
<td>99.7 ± 0.1</td>
</tr>
</tbody>
</table>

### Table IV
Ablation studies on convolutional input embedding (CIE), time positional encoding (TPE) and time-sensitive MSA (TMSA). According to the results, the TMSA module is the most effective module of our method.

<table>
<thead>
<tr>
<th>Method</th>
<th>ACC (%)</th>
<th>SE (%)</th>
<th>SP (%)</th>
<th>AUC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o CIE</td>
<td>85.3 ± 1.1</td>
<td>83.0 ± 0.9</td>
<td>85.4 ± 1.1</td>
<td>90.0 ± 0.4</td>
</tr>
<tr>
<td>w/o TPE</td>
<td>87.9 ± 0.6</td>
<td>81.1 ± 0.9</td>
<td>88.1 ± 0.6</td>
<td>88.6 ± 1.8</td>
</tr>
<tr>
<td>w/o TMSA</td>
<td>85.0 ± 1.5</td>
<td>81.1 ± 0.9</td>
<td>85.1 ± 1.5</td>
<td>88.5 ± 2.0</td>
</tr>
<tr>
<td>Ours</td>
<td>89.5 ± 0.8</td>
<td>87.6 ± 0.9</td>
<td>89.6 ± 0.8</td>
<td>93.6 ± 0.3</td>
</tr>
</tbody>
</table>
Average probability changes compared to \( C_1 \) and then evaluate the impact of \( A \).

<table>
<thead>
<tr>
<th>( A )</th>
<th>0.3</th>
<th>0.6</th>
<th>0.9</th>
<th>1.2</th>
<th>1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>82.6 ± 0.8</td>
<td>85.9 ± 1.3</td>
<td>89.5 ± 0.8</td>
<td>87.1 ± 1.4</td>
<td>82.6 ± 1.2</td>
</tr>
</tbody>
</table>

Comparison of accuracy with different \( B \) in Eq. 6. We fix \( A = 0.9 \) and then evaluate the impact of \( B \).

<table>
<thead>
<tr>
<th>( B )</th>
<th>0</th>
<th>1.0</th>
<th>2.0</th>
<th>3.0</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>81.4 ± 0.9</td>
<td>85.0 ± 1.3</td>
<td>89.5 ± 0.8</td>
<td>86.5 ± 0.7</td>
<td>82.0 ± 1.0</td>
</tr>
</tbody>
</table>

3) Time-sensitive MSA (TMSA): We justify our time-sensitive MSA by simply removing the time related matrix \( (T) \) from our model. As shown in Table IV, TMSA significantly improves the performance of the model, and this module improves the network performance most. Accuracy, sensitivity, specificity and AUC are all increased by at least 4%, to be specific, 4.5%, 6.5%, 4.5% and 5.1%, respectively.

4) Hyper-parameters \( A \) and \( B \) in Eq. 6: As aforementioned, we need the model to pay more attention to the most recent fundus images and less on the older ones when processing a particular fundus image. Eq. 6 is the key component to fundus images and less on the older ones when processing we need the model to pay more attention to the most recent find that model performs best when \( \Delta E \). Conditional Glaucoma Forecast Network

E. Conditional Glaucoma Forecast Network

With the help of time condition, C-GLIM-Net can output the probability of developing glaucoma of a patient conditioned on a specific time. In the following experiment, we feed in first \( i = 4 \) images of a sequence into C-GLIM-Net, under the condition \( \Delta t_{i+1,i} \) and \( \Delta t_{i+2,i} \) to predict the labels of the 5th and 6th images. We denote the results of the 5th and 6th predictions as C-GLIM\(_1\) and C-GLIM\(_2\), respectively. Because DeepGF is used to predict probability at the next step and cannot predict the probability at a specified time, to compare with C-GLIM\(_1\), we feed in first \( i = 4 \) images to train the DeepGF to only predict the label of the 5th image. As shown in Table VIII, our C-GLIM\(_1\) outperforms the DeepGF by achieving 83.0%, 87.6%, 82.9% and 90.7% in accuracy, sensitivity, specificity and AUC respectively. The specificity is dropped slightly in C-GLIM\(_2\) because it is harder to predict positives at a longer time. To further investigate this, we show the prediction of time-steps in-between the 5th and 6th visits in Fig. 6. Although the probability of positives are increased stably, the false negatives are increased due to the probabilities of some positive samples do not exceed the 50% threshold at the 6th visit.

![Average probability changes compared to \( C_1 \) of our C-GLIM-Net performing on the time-steps between the 5th and 6th visits.](image)
V. Conclusion

In this paper, we present the first transformer-based glaucoma forecast network for irregularly sampled sequential fundus images named GLIM-Net. Instead of using linear transformation as input embedding method, we utilize a convolutional neural network to embed the input into specific dimension to address the problem that transformer architecture needs large-scale training data. We propose time positional encoding and time-sensitive MSA to harness the challenge that the time interval of fundus images in a sequence varies from the minimum of only one day to the maximum of 13 years. To make our model more practical in real life, both input and output are concatenated with a time condition so as to output the probability of developing glaucoma conditioned on the time we feed to the model. Extensive experiments on the SIGF dataset demonstrate that our proposed GLIM-Net greatly outperforms other state-of-the-art methods in the glaucoma forecast task. Furthermore, we prove that the transformer architecture, which has achieved great success recently due to its effectiveness in traditional computer vision tasks, can also learn representative features of fundus images and can provide an idea of representing transformer architectures in the medical imaging area. Besides, extensive experiments prove that our two proposed modules can help to better capture temporal information of sequential data, which may provide a good reference for the problems with similar data structures as that of our data. A limitation of our current work is that the dataset used has a limited number of glaucoma cases for training. It would be better to collect more cases for training the model with better generalization ability in the future.

REFERENCES


