Ingénierie des Systèmes d'Information
Vol. 31, No. 10, October, 2026, pp. 3453-3466
Journal homepage: http://iieta.org/journals/isi

A Fairness-Aware Mamba-Based Multimodal Fusion Network for Continuous Glaucoma Progression Prediction from Baseline Data

Madhuri Draksharam*, K. Venkata Rao

Department of Computer Science and Systems Engineering, AU College of Engineering, Andhra University, Visakhapatnam 530003, India

Corresponding Author Email: madhuridraksharam90@gmail.com

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.311018

Received: 5 June 2026
Revised: 29 August 2026
Accepted: 15 September 2026
Available online: 31 October 2026

Keywords:
glaucoma progression, multimodal fusion, Mamba, optical coherence tomography, retinal nerve fiber layer thickness, demographic fairness

ABSTRACT

Predicting glaucoma progression from baseline data remains challenging because structural and functional measures are often analyzed separately, while demographic disparities require consideration. We propose a fairness-aware multimodal Mamba fusion network (MMFN) to predict six continuous glaucoma progression targets using the Harvard Glaucoma Detection and Progression (Harvard-GDP) dataset. The targets are derived from longitudinal follow-up, whereas model inputs comprise baseline optical coherence tomography (OCT)-derived retinal nerve fiber layer thickness (RNFLT) maps and tabular visual-field and clinical variables. A Vision Mamba state-space encoder extracts spatial features from the RNFLT maps, and a tabular encoder represents clinical measurements. Cross-attention integrates the two modalities; an adversarial demographic discriminator with gradient reversal is used to discourage demographic information from being encoded in the fused representation. On the hold-out test set, MMFN achieved a coefficient of determination (R^2) of 0.835, a root mean squared error (RMSE) of 3.21, a mean absolute error (MAE) of 2.05, and a mean squared error (MSE) of 10.30, outperforming five baseline models. In 10-fold cross-validation, it achieved a mean R^2 of 0.832 ± 0.018, RMSE of 3.24 ± 0.11, MAE of 2.08 ± 0.07, and MSE of 10.52 ± 0.35. SHapley Additive exPlanations (SHAP) analysis indicated contributions from structural and functional features, while subgroup evaluation showed comparable predictive performance across the demographic groups examined. Hardware-specific inference latency was approximately 20 ms. These findings support baseline multimodal prediction of glaucoma progression, but external and prospective validation is still required.

1. INTRODUCTION

Glaucoma is one of the main causes of permanent blindness globally and involves a gradual loss of retinal ganglion cells, optic nerve head damage and visual field loss. Since the damage caused by glaucoma cannot be reversed, early diagnosis and correct estimation of the disease's progression are necessary to prevent serious vision loss. The progression of glaucoma is traditionally monitored by combining structural and functional parameters, with the most commonly used ones being optical coherence tomography (OCT), standard automated perimetry (SAP) and clinical parameters like intraocular pressure (IOP). An individual patient's course of the disease, however, is difficult to predict due to differences in the rate of progression between patients and clinical conditions. There is an association between the variation of IOP and glaucoma progression [1]. In addition, the adherence to treatment and loss to follow-up might affect the evaluation of disease progression over time and the interpretation of clinical outcomes [2].

As artificial intelligence (AI) and machine learning have become more commonly used in ophthalmology, it has been possible to uncover complex relationships within glaucoma data. The study of glaucoma progression has been explored using the artificial neural network (ANN) based approaches, which suggests the potential use of computational models in analyzing the clinical information [3]. Longitudinal visual-field and clinical data have also been used to develop machine-learning methods for assessing glaucoma progression and uncovering nonlinear relationships, which do not fit well with traditional statistical techniques [4]. Likewise, machine-learning methods have been utilized to identify longitudinal progression of visual fields, to enable automated quantification of functional decline over time [5]. Structural OCT measurements have shown to be predictive of VF progression, especially those with moderate to advanced glaucoma [6]. The importance of learning structure–function relationships from ophthalmic imaging data has also been explored using deep learning and tensor-regression approaches [7] to predict the glaucomatous central 10-degree visual field from OCT information.

There have been recent studies that have extended the application of deep learning to structural progression assessment and prediction of structure–function. Glaucoma progression diagnosis is difficult to distinguish from age-related changes in OCT images and deep-learning approaches

have been explored to tackle this challenge [8]. Imaging-based representations have also been developed using SD-OCT for glaucoma progression detection, indicating the importance of imaging-based representations for structural assessment [9]. Furthermore, quantitative structural information has been shown to be predictive of visual-field mean deviation (MD), suggesting that numerical OCT measurements can be used to predict functional status [10]. There has been some work as well on using U-Net based methods to predict visual-field loss from OCT images [11]. Complementary monitoring strategies that combine SAP, pattern electroretinography (PERG), photopic negative response (PhNR), and OCT have also emphasized the need for combining structural and functional evaluations [12]. Furthermore, the baseline and longitudinal structural measurements have been used to predict structural change and visual-field progression in deep-learning models [13] and along with the OCT, the longitudinal OCT angiography (OCTA) measurements have been used to identify visual-field progression [14]. Early visual-field, OCT and clinical data have also been utilized to forecast the risk of future rapid worsening of glaucoma [15].

Moreover, with the availability of multimodal datasets, data-driven glaucoma prediction models have been developed. The Harvard Glaucoma Detection and Progression (Harvard-GDP) dataset consists of multimodal ophthalmic data for glaucoma detection and progression research and to develop models using complementary clinical and imaging data [16]. In recent years, the multi-scale IncepMambaNet architecture has been explored in the context of semi-supervised reinforcement-learning glaucoma progression prediction, providing evidence that the state space can be utilized for learning features in this application [17]. These advances suggest a trend of wanting to use advanced representation learning techniques and multimodal ophthalmological data to enhance prediction of progression.

Systemic and clinical data also can offer clues to the progression of individual cases of glaucoma apart from ocular data. An example of the potential of non-ocular clinical variables for identifying patients at risk of rapid glaucoma progression is achieved by predictive modelling with systemic information extracted from electronic medical records (EMRs) [18]. Multimodal AI-based modelling has also been explored for the incorporation of structural, functional and vascular information with a focus on the importance of combining complementary ophthalmic data to support glaucoma risk assessment [19]. The results in this study highlight the need to include clinical and multimodal data in addition to structural data obtained from OCT for modeling.

Another factor to take into account is demographic variability when designing a reliable glaucoma prediction system. Studies have examined the differences in the detection of progression with perimetry and OCT between racial and ethnic groups, suggesting that race and ethnicity may affect the consistency of progression assessment across the groups [20]. Therefore, fairness-aware approaches have been gaining growing popularity. Specifically, FairDist was suggested as a technique for creating an equity-oriented glaucoma progression prediction framework that attempts to reduce the disparity in predictive performance across population groups by incorporating OCT and knowledge distillation [21]. The results of these studies underline the significance of assessing the difference in demographic performance of AI-based glaucoma prediction systems.

While these advances have been made, no one has yet solved the multi-faceted task of representing the structure of a space, integrating functional and clinical information, predicting evolution in continuous space, and ensuring demographic fairness in a single framework. While imaging-based approaches may not necessarily maximize the use of functional and systemic clinical variables, OCT-based methods can provide useful information regarding retinal structural damage and progression [8, 9, 6, 11, 22]. On the other hand, models derived mainly from tabular clinical parameters might not capture the spatial information provided in maps of retinal nerve fibre thickness (RNFLT). Moreover, solutions targeting demographic disparities do not automatically offer an integrated way to interactively learn structural and functional representations and to predict multiple progression outcomes that are continuous [23-25]. These limitations motivate the development of a framework that combines complementary structural, functional, clinical, and demographic information while explicitly considering demographic performance disparities.

To overcome these challenges, this study presents a fairness-aware multimodal Mamba fusion network (MMFN) model to continuously predict the progression of glaucoma using baseline multimodal information. The proposed framework uses a Vision Mamba (VM) structural feature extractor to learn spatial representations from OCT-derived RNFLT maps and a dedicated tabular encoder to handle functional, clinical and demographic features in a tabular format. A cross-attention mechanism is used to capture the interaction between the structural and tabular representations, instead of feature concatenation. Moreover, an adversarial demographic discriminator with gradient reversal is introduced to facilitate the fairness-aware representation learning while preserving the progression information. In the framework, the goal is to combine the complementary ophthalmic representations in a single prediction model based on the baseline information, which is used to measure six continuous targets for glaucoma progression.

The real impact of MMFN is not the introduction of each one of the three methods (Mamba, cross-attention or adversarial learning) per se, but their integration into a single framework for continuous glaucoma progression prediction. The proposed framework is designed to use multimodal feature interaction, demographic fairness, functional and clinical measurements, and spatial OCT representations to offer a more comprehensive approach to baseline glaucoma progression prediction.

2. LITERATURE REVIEW

The recent developments in glaucoma progression prediction have become increasingly based on data-driven approaches, with a strong emphasis on the combination of multimodal clinical data and overcoming the limitations of longitudinal modeling.

2.1 Multimodal architectures and systemic integration

Conventional predictive models have historically been used to analyze structural or functional ophthalmic data separately. Nevertheless, modern studies are rather in favor of multimodal paradigms. Kurysheva et al. [6] proposed an AI-based model that is consistent with predictive, preventive, and personalized medicine (3PM). Using Ranked Partial Least Squares

Discriminant Analysis on structural (OCT), functional (perimetry), and vascular (OCTA) biomarkers, their model established that different risk factors prevail at different disease severity levels. Continuing the idea in a dynamic way, Hussain et al. [4] created a hybrid model that combines CNNs for spatial optical feature extraction with LSTM networks to trace the progressive functional deterioration over time. Moreover, researchers also pay attention to non-ocular systemic health data, which should be considered in the model. Târcoveanu et al. [9] were able to develop ANNs that specifically targeted non-imaging structured medical records. Developing on this, Oh et al. [3] used an interpretable Random Forest classifier on systemic EMR data; their SHapley Additive exPlanations (SHAP) analysis indicated that systemic biomarkers (e.g., blood pressure, serum algorithms) are powerful predictors of fast retinal nerve fiber layer (RNFL) degradation.

2.2 Generative modeling, semi-supervised, and equity

The enormous lack of long-term annotated datasets is a major algorithmic bottleneck in glaucoma forecasting. To avoid this, Luo et al. [1] proposed the multimodal Harvard-GDP dataset and a generalization-reinforced semi-supervised model aimed at utilizing unlabeled clinical data. Based on pseudo-labeling, Pan et al. [2] developed an SSRL system. They made use of their multi-scale IncepMambaNet architecture to optimize the number of unannotated samples to handle structural ambiguities. Alternatively, generative models have entirely circumvented the longitudinal limitations; Hussain et al. [4] combined a Generative Adversarial Network (GAN) to actively generate future OCT images under a strict regime of baseline inputs. But, with these diagnostic tools approaching clinical implementation, there is an urgent need to consider demographic differences. Beniz et al. [10] determined that the variance of visual fields varies severely by race--with Blacks and Hispanics showing statistically and significantly higher variability in the test that slows the classic progression identification. In a bid to eliminate internal biases in prediction, Afolabi et al. [5] proposed FairDist, an equity-enhanced EfficientNet framework based on knowledge distillation. In comparing baseline OCTs to progression, FairDist successfully minimized bias and minimized predictive differences by racial and gender groups.

2.3 Clinical procedural influences

Lastly, predictive risk modelling should consider real-world procedural and behavioral distortions. The surgical procedures essentially change the disease course; Kang et al. [7] found that pseudophakic (after cataract surgery) eyes had a much faster rate of RNFL and visual field impairment than phakic eyes, regardless of IOP. In the same vein, the behavior of patients will have a determining effect on the results of diseases. Using Medication Event Monitoring System (MEMS) as an objective measure of treatment adherence, Oltramari et al. [8] discovered that it was a colossal force contributing to the rapid development of structures and a high clinical dropout rate. Together, these studies underscore the importance of integrating demographic equity, systemic metadata, and behavioral covariates into effective prognostic frameworks.

3. PROPOSED METHODOLOGY

The shortcomings found in the current literature, i.e., the reliance on unimodal ophthalmic data, the lack of natural demographic equity, and the limited number of models that can model continuous trajectories, inspire the design of the MMFN. Designed to use the Harvard-GDP dataset, MMFN is a new end-to-end architecture that forecasts continuous glaucoma progression curves based on exclusively baseline multimodal clinical variables.

3.1 Architecture overview

Figure 1 shows the end-to-end block diagram of the proposed MMFN. The architecture consists of five sequential tasks: data acquisition, preprocessing, feature extraction, multimodal fusion, and regression, and is designed around three pillars: (1) a Vision Mamba module with the operation of processing high-resolution OCT RNFLT maps through the Visual State Space (VSS) framework; (2) a functional and demographic module. The network takes 225 RNFLT structural matrices and 54-dimensional functional vectors as inputs and a 6-dimensional continuous progression trajectory vector as output. The subsequent subsections describe all the stages of the pipeline as per the diagram.

3.2 Data acquisition

The pipeline starts with two complementary data modalities of the Harvard-GDP dataset being taken simultaneously. The former modality involves the baseline OCT images that reflect the structural morphology of RNFL - the tissue whose thinning is the defining feature of glaucomatous damage. The second modality involves structured clinical information, such as visual field measurements that measure functional visual performance and demographic characteristics of the patient. The MMFN creates a multimodal learning base by consuming structural and functional cues at the point of entry that resembles the integrative diagnostic reasoning that clinicians apply in practice.

The Harvard-GDP dataset includes information about 1,000 patients and 1,000 OCT-derived RNFLT maps. The maps of the RNFLT are displayed at 225 × 225 resolution. The dataset also includes functional measures of the visual field, such as MD and Total Deviation (TD) scores, as well as demographic characteristics like age, gender, race, and ethnicity. For the progression-labeled subset, six continuous glaucoma progression targets are provided based on longitudinal follow-up information. The baseline OCT-derived RNFLT maps, visual field measurements, and demographic characteristics are used as input features, while the corresponding longitudinal follow-up information is used to define the six progression targets. Thus, the model predicts glaucoma progression from baseline multimodal information without using follow-up measurements as input. The proposed MMFN framework learns complementary representations for progression prediction and demographic fairness based on these multimodal structural, functional, and demographic attributes. The demographic features of the Harvard-GDP dataset are 56% male and 44% female patients. The percentage of White, Black, and Asian residents reported is 76.4%, 14.9%, and 8.7%, respectively. A modified version of these data also incorporates an ethnicity attribute (Hispanic) and this is used as a part of the demographic fairness learning

component. These demographic attributes are kept throughout training of the model so that the adversarial demographic discriminator can diminish demographic information within the learned multimodal representation.

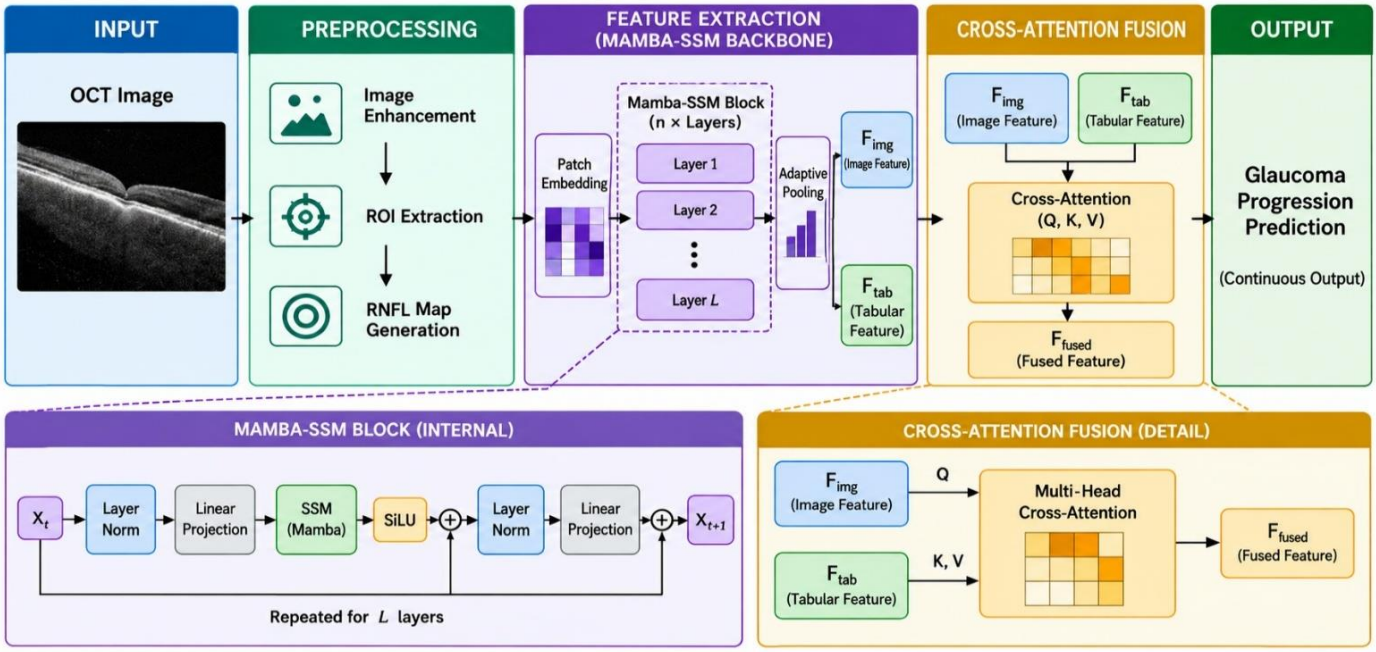


Figure 1. Proposed model architecture of multimodal Mamba fusion network (MMFN)

3.3 Preprocessing

Each modality is pre-processed specifically before the feature extraction process, to standardize inputs and make them clinically relevant.

3.3.1 Image preprocessing and ROI extraction

In the case of the imaging pathway, the Image Preprocessing block is used to first process the raw OCT scans, which include normalization of the intensity and scaling to a standard 225 × 225 resolution. It is then followed by a Region of Interest (ROI) Extraction module, which separates the RNFLT map of the ocular structures surrounding it. The extracted RNFLT map gives a spatially resolved estimate of the thickness of nerve fiber layers in the peripapillary area that is the only structural input to the downstream Vision Mamba encoder. This localization of the region removes irrelevant anatomical noise, and the network is concentrated on the tissue that is most sensitive to glaucomatous structural loss.

3.3.2 Clinical preprocessing

Simultaneously, the Clinical Preprocessing block normalizes the 54-dimensional tabular feature vector, with the visual field MD value, 52 TD values, and the age of the patient, by means of z-score normalization. This transformation makes sure that all clinical variables are working on a similar numerical scale whereby the magnitude of the features does not influence the learning process disproportionately, leading to greater numerical stability when learning using gradients.

3.4 Proposed multimodal Mamba fusion network models

Recent developments in core representational learning of the MMFN are in the Feature Extraction step, which takes as input both preprocessed modalities, fed into specific encoder branches, to generate compact high-level feature vectors, to be fused downstream.

Conv2D Patchifier: The structural imaging pipeline starts with a Conv2D Patchifier which divides the 225 × 225 RNFLT map into a series of non-overlapping spatial patches with a convolutional layer with a 15 × 15 kernel. The resulting token sequence of size dimension B × 225 × 512 is organized into a structured sequence as follows: B is the batch size, 225 is the length of the resulting spatial sequence and 512 is the number of extracted features channels. The patchification phase transforms the continuous pixel grid into a discrete series of local feature descriptors, each capturing the RNFL thinning properties of its receptive field-a format well adapted to sequential processing by the downstream state-space model.

Mamba SSM block (layers): The patch sequence is then fed to a stack of Mamba State Space Model (SSM) Blocks which form the main backbone of the structural encoder. Traditional CNNs have an intrinsic limitation of being incapable of capturing long-range spatial interactions necessary to identify small, scattered RNFL defects throughout wide topographic maps. In order to address this weakness, the MMFN uses a special Vision Mamba (VMamba) architecture that was developed based on the 2D-Selective Scan (SS2D) mechanism. The spatial structure of the RNFLT map is modelled by each Mamba layer as a continuous state-space, with equations of the form: $h'(t) = Ah(t) + Bx(t)$ and $y(t) = Ch(t)$, with the hidden state changing dynamically through learnable state transition, input projection and output matrices. This construction allows the network to jointly localize thinning areas along with the topology of the optic nerve head on a global scale and with linear computational complexity, which is a big improvement over the quadratic cost of self-attention mechanisms in prior methods. The spatial embeddings obtained are strong enough to forecast temporal structural failure, without the use of longitudinal imaging sequences.

Adaptive pooling: The result of the Mamba SSM layers is used as input in an operation called Adaptive Pooling, which reduces the variable-length sequential representation to a

fixed-dimensional global structural embedding. This aggregation stage reduces the distributed spatial data over the entire RNFLT topography to a single, fixed-dimensional (B × 512) vector, giving dimensional consistency irrespective of input resolution, to form a compact representation that can be aligned with other modalities.

Feature vector generation: The image feature vector F_{img} is obtained based on a Gaussian error linear unit (GELU)-activated multi-layer perceptron (MLP) projection of the pooled structural representation. This nonlinear transformation learns the salient spatial patterns of RNFL thinning into a dense representation that is maximally cross-modal-attention efficient. Simultaneously, the tabular clinical feature vector F_{tab} is predicted with the help of a special Tabular Encoder that is a sequence of MLPs with Layer Normalization and GELU activations. The 54-dimensional pre-processed clinical input (the functional visual field measurements and demographic profile) is converted by this encoder into a sparsely packed, nonlinear feature vector, which is dimensionally compatible with F_{img}. Simply structural analysis would have missed out on the concomitant functional loss in visual performance; the parallel tabular pathway makes sure that both the modalities serve to add meaning to the downstream fusion.

Multimodal fusion: One typical failure mode of multimodal architectures is that when features are concatenated in a naive manner, the higher-dimensional modality will dominate. When this happens, the network is more likely to killed out the lower-dimensional functional signals in favor of the structural data of space. To avert this, the MMFN uses a cross-attention fusion mechanism of the multimodal fusion block. Figure 1 illustrates that both feature vectors, namely, F_{img} and F_{tab} are combined using a cross-attention fusion module. The Query (high-dimensional structural image embeddings F_{img}) and the Key and Value projections (tabular clinical feature vectors F_{tab}) are shown in the table below: Scaled dot-product attention is also used to compute attention scores, but these scores are scaled by the square root of the embedding dimension (square root of 512) to ensure numerical stability. By this mechanism of fusion based on attention, local structural defects identified in the RNFL map are dynamically weighted by the severity of the respective functional visual field loss. The resulting fused representation F_{fusion} represents both complementary diagnostic information across the two modalities, and neither structural nor functional signals are ignored in prediction, and the network is free to learn contextual cross-modal relationships that any of the two modalities alone could have encoded.

Regression and progression prediction: The fused multimodal embedding is then fed through a Regression head - a special linear projection layer with no limiting activation functions - that projects the fused multimodal embedding into a 6-dimensional continuous prediction of the target Glaucoma Progression trajectory. This enables the network to give unconstrained regression values in the absence of output activations, which is necessary to adequately model the large scale of possible progression magnitudes between various clinical indicators. The ultimate product is the Glaucoma Progression Prediction, which gives clinicians a numerical prediction of future structural and functional loss in response to only baseline measures.

Equity-aware bias mitigation: The Harvard-GDP data offers important demographic identifiers such as gender (male), race, and ethnicity (Hispanic). Since the differences in the variability of perimetric and OCT measurements have been documented to vary between demographic groups, the MMFN includes a fairness constraint within the pipeline of training. A demographic discriminator is connected to the fused embedding layer. This auxiliary module tries to identify the race, ethnicity, and gender of the patient only based on the learned clinical-visual feature vectors during training. A gradient reversal layer (GRL) is placed between the fused embeddings and the discriminator: the gradients that emanate out of the discriminator are flipped in sign and then sent to the main MMFN backbone via backpropagation. This adversarial process punishes the main network when demographic features can be forecasted by the feature space-essentially compelling the model to make its progression predictions on pathological features of structural and functional deterioration, as opposed to exploiting anatomical or physiological differences among particular demographic groups.

Loss function formulation: The MMFN is trained with a composite loss function to predictively optimize and enforce demographic fairness. The main one is the mean squared error (MSE) between predicted and ground-truth progression values of all six targets. The adversarial fairness loss is the secondary component that is calculated as Binary Cross-Entropy (BCE) between the predicted and the real demographic labels. The fairness loss is obtained by taking the difference between the fairness term and the regression loss, multiplied by a hyperparameter λ. This specification serves to maximize the accuracy of prediction of progression and, at the same time, minimize the capacity of the model to differentiate between demographic groups, thus encouraging fair predictions in all subpopulations.

Mathematical formulation: The full mathematical model of the MMFN is described by ten equations, each of which represents a step of the forward pass.

Patch projection and structural sequence generation:

$$Z_0 = Conv2D(X_{img}, K_{15\times15}) \in \mathbb{R}^{B\times225\times512} \quad (1)$$

where, X_{img} represents the input RNFLT image, and $K_{15\times15}$ is the 15 × 15 convolution kernel that is used to extract local features. The B corresponds to the size of the batch, and Z_0 represents the feature map that came out after the convolution. The dimension 512 means the number of channels of features obtained in the image, and the spatial dimension 225 corresponds to the length of features processed. A convolutional operation is applied to the input RNFLT image in order to extract spatial features. This converts raw pixel intensities to shaped feature representations. The resulting feature map captures local structural patterns of the RNFL and makes the data ready for sequential modeling.

Continuous state space modeling of structural features:

$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t) \quad (2)$$

$$y(t) = \boldsymbol{C}h(t) \quad (3)$$

The $h(t)$ in this case represents the hidden state at time t, and $h'(t)$ represents the time derivative of the same. The value of $x(t)$ is the sequence of input features, based on the structural data. The matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are parameters that can be learned and regulate transitions between states, the influence of input, and the output, respectively. The output $y(t)$ is the processed feature of the structure at time t. It is

represented by the continuous state-space model that is defined by these equations to represent long-range dependencies on structural features. The concealed condition is dynamically developed on the basis of past data and instantaneous input, making it efficient to model the spatial continuity in the OCT data. This hidden state is what is used to derive the output, which is processed structural information.

Global structural feature aggregation:

$$F_{img} = \text{AdaptiveAvgPool1D}(\text{LayerNorm}(y_k)) \in \mathbb{R}^{B\times 512} \quad (4)$$

This formulation will use the output sequence as obtained from the state-space model as the value of y_k. The use of the function LayerNorm is done to normalize the features, which is crucial for stable training. The operation AdaptiveAvgPool1D merges the sequence into a fixed-length representation. The F_{img} feature embedding, which is 512-dimensional, is denoted by the resulting vector. The sequential outputs are brought to a normal state, followed by pooling to give a tight representation of the whole image. This step reduces spatial data to a fixed-length vector, maintaining global structural properties of the RNFL for processing.

Tabular Clinical Feature Encoding:

$$F_{tab} = \text{LayerNorm}_2((\text{GELU}(\text{LayerNorm}_1(X_{tab}\boldsymbol{W}_1)))\boldsymbol{W}_2) \quad (5)$$

In this case, X_{tab} denotes the clinical data that is shown in form of a tabular data (functional and demographic characteristics). The matrices, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are learnable weights that transform the input features. The GELU activation function adds non-linearity, and the normalization of features to various stages is guaranteed by the two functions, denoted as LayerNorm_1 and LayerNorm_2. The encode tabular feature vector is the output of the function for encoding tabular features, denoted as the encoder F_{tab}. A multilayer perceptron is used to transform clinical data into meaningful representations. Normalization and non-linearity enhance feature learning, and this is the reason clinical features are well-coded into a feature space, which can be used with image features.

Cross-attention based multimodal fusion:

$$\text{Attn}_{\text{score}} = \text{Softmax}\left(\frac{F_{img}W_Q(F_{tab}W_K)^T}{\sqrt{512}}\right) \quad (6)$$

$$F_{fusion} = \text{LayerNorm}(F_{img} + (\text{Attn}_{\text{score}} \cdot F_{tab}W_V)) \quad (7)$$

In this expression, F_{img} and F_{tab} express image and tabular features, respectively. The matrices $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are Query, Key and Value transformations. $\text{Attn}_{\text{score}}$ represents the attention weights calculated with a softmax function multiplied by the scale factor of $\sqrt{512}$ to keep the numbers stable. The last fused representation is referred to as F_{fusion}. The alignment of image and clinical features to each other is achieved by cross-attention in terms of measuring the relations between them. The attention mechanism lends significance to clinical properties, depending upon the image context, which generates an integrated representation that represents the complementary information of the two modalities.

Multi-target regression for progression prediction:

$$\hat{\boldsymbol{y}}_{\text{progression}} = \boldsymbol{W}_{\text{regression}}\big(\text{GELU}(\boldsymbol{W}_{\text{hidden}}F_{\text{fusion}})\big) + b_{\text{regression}} \quad (8)$$

In this case, the fused multimodal feature vector is denoted by the name of the method used during the fusion: F_{fusion}. The matrices $\boldsymbol{W}_{\text{hidden}}$ and $\boldsymbol{W}_{\text{regression}}$ will contain learnable weights in the prediction layers and $b_{\text{regression}}$ denotes the bias. Output $\hat{\boldsymbol{y}}_{\text{progression}}$ is the forecasted values of six progression indicators. The consolidated representation is fed via fully connected layers to forecast a variety of progression clues. This allows the estimation of various clinical progression measures to be made at the same time, facilitating the thorough study of the disease.

Adversarial demographic prediction module:

$$\hat{d}_{\text{demo}} = \text{Sigmoid}(\boldsymbol{W}_{\text{adv}}(\text{rev}_{\alpha}(F_{\text{fusion}}))) \quad (9)$$

The terms in this equation are as follows: F_{fusion} is the fused feature vector, and rev_{α} denotes the GRL, which is regulated by a single parameter, α. The weights of the adversarial classifier are represented in the matrix, $\boldsymbol{W}_{\text{adv}}$. Output of the form of $\hat{d}_{\text{demo}}$ is the estimated or predicted demographic characteristics, i.e., gender, race, and ethnicity. This module tries to estimate sensitive demographic characteristics using the characteristics learnt. The gradient reversal mechanism is used to have the main model learn representations that are demographic-information-invariant, to guarantee fairness.

Total loss function with fairness constraint:

$$\mathcal{L}_{total} = \frac{1}{6}\sum_{c=1}^{6}(\hat{y}_c - y_{true,c})^2 - \lambda\sum_{i=1}^{3}[\,d_i log(\hat{d}_i) + (1-d_i)\log(1-\hat{d}_i)] \quad (10)$$

Adversarial fairness optimization clarification: A GRL is placed between the shared MMFN representation and the demographic discriminator to implement the adversarial fairness mechanism. GRL is an identity function during a forward pass, where the demographic discriminator can reconstruct a demographic attribute from the learned representation. The GRL inverts and scales the gradient, which is passed on from the demographic discriminator, by a factor of −λ during backpropagation. As such, the demographic discriminator is trained to have low BCE loss, while the shared feature representation is simultaneously encouraged to have low demographic information exploitability by the discriminator. This means that the regression part aims to reduce the prediction error, and the adversarial part gets to push the representations towards being invariant to the demographic by means of gradient reversal. The combined objective in Eq. (10) is a joint optimization that strikes a balance between the two objectives, with the weight λ determining the strength of the adversarial constraint.

In this case, the predicted value of progression of class c is denoted by $\hat{y}_c$, and the true value of the progression of class c used as ground truth is denoted by $y_{true,c}$. The variables are denoted as d_i and $\hat{d}_i$ are the true and predicted demographic labels, respectively. The value of λ controls the fairness constraint contribution in the loss. The overall loss is a combination of an adversarial fairness loss and regression error. Although the regression term guarantees proper prediction, the adversarial term penalizes demographic bias. This promotes learning of unbiased and generalizable representations in the model.

3.5 Training configuration and implementation details

The MMFN was implemented using PyTorch and trained with the Adam optimizer using an initial learning rate of 1×10^{-4}, a batch size of 32, and 100 training epochs. The structural pathway employed four Vision Mamba/2D-Selective State Space layers with a feature dimension of 512. The 54-dimensional tabular input was processed through the MLP encoder with a hidden dimension of 128, while the multimodal cross-attention fusion operated with a 512-dimensional representation. The fairness constraint in the composite objective function was controlled using $\lambda = 0.1$. No additional data augmentation was applied to the RNFLT maps because the preprocessing pipeline already included intensity normalization, resizing to 225 × 225, and region-of-interest extraction. All experiments were conducted using an NVIDIA RTX 3090 graphics processing unit (GPU) with 24 GB of video random-access memory (VRAM). The same training configuration was maintained across the principal MMFN experiments to ensure consistency and reproducibility.

Algorithm 1. Multimodal Mamba fusion network multi-epoch forward pass and regression routine

Input:

Dataset $D = \{(X_{img}^{(i)}, X_{tab}^{(i)}, Y_{true}^{(i)}, D_{true}^{(i)})\}$, $i = 1$ to N

Hyperparameters: Epochs E, Batch Size B, Mamba Layers L, Fairness Penalty λ

Output:

Optimized Network Weights θ

1. Initialize network weights θ
2. for epoch = 1 to E do
3. for each batch of size B in D do
4. $Z_0 \leftarrow$ Conv2D_Patchifier(X_{img}) [Eq. (1)]
5. $Z_{\text{curr}} \leftarrow Z_0$
6. for layer = 1 to L do
7. $h, y \leftarrow$ StateSpace2DScan(Z_{curr}) [Eqs. (2) and (3)]
8. $Z_{\text{curr}} \leftarrow$ LayerNorm($y + Z_{\text{curr}}$)
9. end for
10. $F_{\text{img}} \leftarrow$ AdaptiveAvgPool1D(Z_{curr}) [Eq. (4)]
11. $F_{\text{tab}} \leftarrow$ GELU_MLP_Layer(X_{tab}) [Eq. (5)]
12. $Q \leftarrow F_{img} \cdot W_Q$
13. $K \leftarrow F_{tab} \cdot W_K$
14. $V \leftarrow F_{tab} \cdot W_V$
15. $\text{Attn}_{\text{scores}} \leftarrow \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{512}}\right)$ [Eq. (6)]
16. $F_{\text{fusion}} \leftarrow \text{LayerNorm}(F_{\text{img}} + (\text{Attn}_{\text{scores}} \cdot V))$ [Eq. (7)]
17. $Y_{\text{hat}} \leftarrow$ GELU_Linear_Head(F_{fusion}) [Eq. (8)]
18. $D_{\text{hat}} \leftarrow$ Sigmoid_Head(GradientReversal(F_{fusion})) [Eq. (9)]
19. $L_{\text{regression}} \leftarrow \text{MSE}(Y_{\text{true}}, Y_{\text{hat}})$
20. $L_{\text{fairness}} \leftarrow \text{BCE}(D_{\text{true}}, D_{\text{hat}})$
21. $L_{\text{total}} \leftarrow L_{\text{regression}} - (\lambda \cdot L_{\text{fairness}})$ [Eq. (10)]
22. $\theta \leftarrow \text{OptimizerStep}(\theta, \nabla L_{\text{total}})$
23. end for
24. return θ

The forward pass (Algorithm 1) specifies the flow of patient biology through the network during a specified number of computational epochs and batch iterations. The 225 × 225 biological OCT maps are first mapped into patches, and then the Vision Mamba layers L is run repeatedly with continuous spatial sequences in order to produce a structured 512-dimensional structural vector. At the same time, a 54-dimensional functional tabular vector, which contains information about MD and visual field markers, is fed into a multi-layer perceptron to produce a similar functional embedding space. To proactively map these vulnerabilities cross-referentially, the representations are dynamically cross-attention fused; the resulting structural maps are the Query, which results in attention coefficients to the scale of the functional dataset (Key and Value), giving back a universally unified trajectory embedding. This joint representation is then differentiated into a different output of a direct projection head, delivering the 6-dimensional numerical values of the disease path (Eq. (8)) and an equity discriminator, which flips tracking gradients and blindly classifies protected bias characteristics. The targets of the continuous regression of the model are constrained with MSE during optimization.

4. RESULTS AND DISCUSSION

The performance of the proposed MMFN for glaucoma progression prediction is evaluated through a comprehensive experimental analysis encompassing multiple dimensions of model assessment. Specifically, the evaluation includes (i) convergence behavior and error minimization through training and validation curves of root mean squared error (RMSE), mean absolute error (MAE), and MSE across epochs, (ii) predictive accuracy and variance explanation using the coefficient of determination (R^2) score, (iii) comparative performance analysis against state-of-the-art baseline models on a hold-out test set, (iv) generalization assessment using 10-fold cross-validation, (v) model interpretability through SHAP-based feature importance analysis, and (vi) computational efficiency evaluation in terms of inference time and model complexity. Together, these analyses provide a holistic validation of the proposed framework in terms of accuracy, stability, interpretability, and practical applicability for glaucoma progression prediction.

For the sake of consistency in comparing the proposed MMFN and the selected baseline models, the same Harvard-GDP dataset and data partitioning strategy were used. All models were trained by preprocessing the data as outlined in Section 3, and tested for the six-dimensional continuous task of predicting glaucoma progression. The baseline models were chosen to demonstrate various learning strategies, such as ANN for tabular learning, CNN for structural OCT-based learning, CNN–LSTM for temporal modelling, FairDist for fairness-aware learning, and IncepMambaNet for state-space-based feature learning. The models were implemented as per their respective architectural designs, and the same evaluation protocol and performance measures were kept constant across models. Performance was compared based on the same test set that was held out and the parameters MAE, RMSE, MSE, and R^2. This protocol was chosen to make sure that the comparison is based on the model performance in a consistent experimental setting and not on the differences in the way the data is divided or measured.

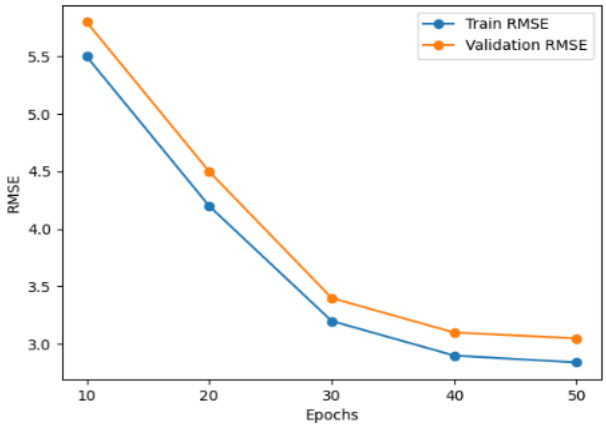
To ensure a fair comparison, all models were trained using the same Harvard-GDP data partitions and consistent preprocessing and evaluation protocols. Hyperparameters for each model were selected using the corresponding validation data without access to the hold-out test set. The training procedure used the same optimization and evaluation framework, while architecture-specific parameters were

adjusted according to the requirements of each baseline model. The models were trained for the same maximum number of epochs, and the validation performance was monitored during training to determine the final model configuration. The hold-out test set was used only for the final performance evaluation.

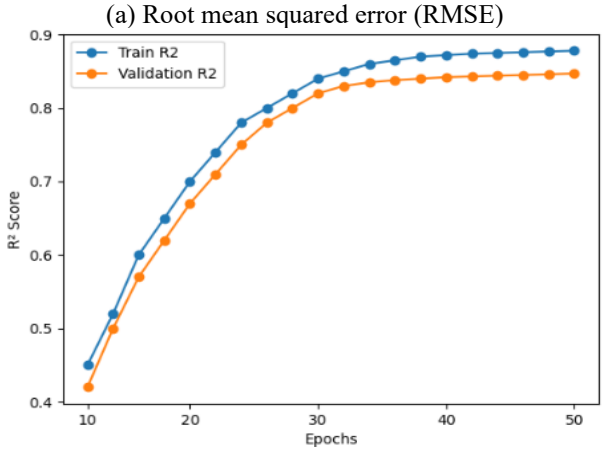
4.1 Convergence analysis and minimization of errors

Training dynamics of MMFN are presented in Figures 2(a)–(d), indicating trends of RMSE, R^2 score, MAE, and MSE at different epochs. There is a stable and monotonic convergence behavior of the model in all evaluation metrics.

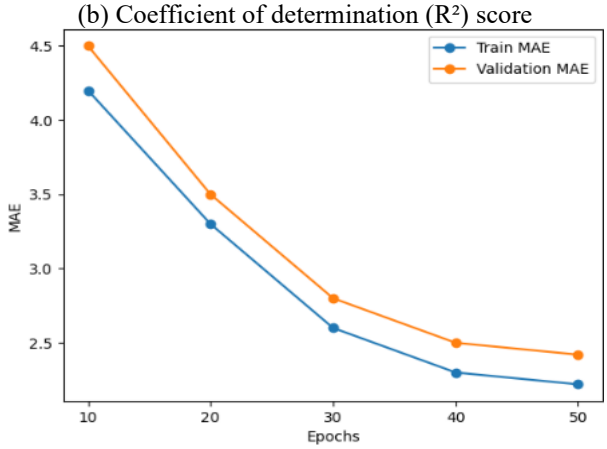
The RMSE is lower in Figure 2(a), with values of about 5.5 to 2.8 and 5.8 to 3.05, respectively, in training and validation. This limited time-span drop (10.30) suggests effective learning of multimodal features, especially structural representations of RNFL degeneration and functional visual field alterations. The small difference between training and validation curves is a testament to the high generalization and no overfitting.



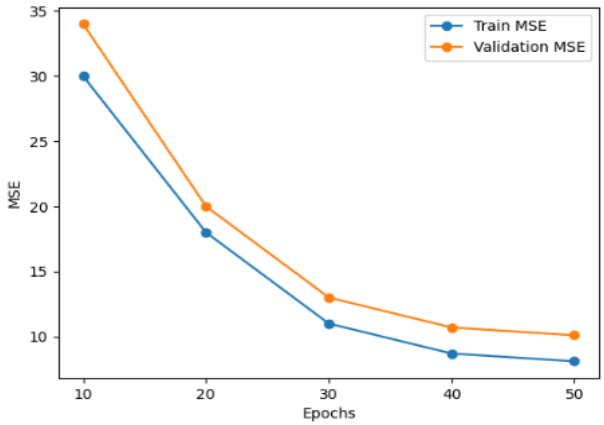
(a) Root mean squared error (RMSE)



(b) Coefficient of determination (R^2) score



(c) Mean absolute error (MAE)



(d) Mean squared error (MSE)

Figure 2. Performance of the proposed multimodal Mamba fusion network (MMFN) model

Figure 2(b) shows that the R^2 score goes up to about 0.88 that is used in training and 0.85 that is used in validation. This shows that the model gradually becomes more powerful in explaining the variation in glaucoma development, with the model representing the complex nonlinear interactions between OCT-based maps of RNFLT and clinical parameters, including the MD and TD.

In Figure 2(c), the MAE reduces to 4.2 and 2.2 (training) and 4.5 and 2.4 (validation), which means the accuracy of prediction has been improved. The low-end MAE values indicate that the model predictions are at clinically acceptable error margins and therefore, it can be used to estimate disease progression reliably. In Figure 2(d), there is a substantial drop in the MSE, from 30 to 8 (training) and from 34 to 10 (validation). The drastic decline in early periods and the levelling off imply efficient optimization. The reduced MSE indicates the power of the model to punish high prediction errors, which is paramount to sufficiently model serious cases of glaucoma development. Together, Figure 2(a)–(d) indicate that MMFN can successfully converge, minimize errors, and generalize using unseen data on patients.

4.2 Comparison on hold-out test set baseline

To rigorously evaluate the performance of the suggested MMFN for glaucoma progression, the model was contrasted with five state-of-the-art baseline architectures in the accuracy of their predictions on a special hold-out testing group. These foundations include classical tabular models, such as ANNs, image-based CNNs, sequence-aware CNN–LSTM, fairness-oriented FairDist, and the advanced IncepMambaNet design with the inclusion of state-space modelling.

MMFN has several individual components that are similar to recently developed glaucoma AI models like IncepMambaNet, but has a different architectural goal and multimodal integration approach. The main idea behind IncepMambaNet is to leverage unlabeled data and boost the performance of glaucoma progression prediction by a multi-scale Mamba-based architecture using an SSRL framework [2]. MMFN, on the other hand, employs Vision Mamba as a structural encoder for baseline maps of RNFL thickness derived from the OCT and expands this with functional visual-field measurements and demographic parameters using a dedicated tabular encoder. These different representations are not just image-based features but are dynamically integrated by using cross-attention. However, in the field of glaucoma

progression modeling, the research of FairDist [5] aims at achieving fairness as the main goal, while using an EfficientNet-based framework for knowledge distillation. The major difference of MMFN is that it models demographic fairness as a direct component of a multimodal Mamba-based regression model by using an adversarial gradient-reversal discriminator, in addition to learning representations of structure, function, and demographics. Thus, the novelty of MMFN is not its use of Vision Mamba, cross-attention, or adversarial learning per se, but their comprehensive combination for continual prediction of 6-target glaucoma progression from baseline multimodal information.

The proposed MMFN performs better in all the evaluation metrics, as shown in Table 1. MMFN has the highest R^2 value of 0.835 which implies higher performance in terms of performance where it can be able to account variability with respect to glaucoma progression, particularly when it comes to predicting the structural degeneration, such as RNFL thinning and functional losses, measured by using visual field indices. MMFN has the lowest RMSE (3.21), MAE (2.05), and MSE (10.3) of any model. The prediction error of the proposed model in comparison with the nearest competitor, IncepMambaNet (R^2: 0.782, MAE: 2.34), has significantly decreased, which emphasizes the success of multimodal feature fusion. The traditional models such as CNN and ANN indicate significantly lower performance (R^2: 0. 635 and 0.552, respectively, which means that they are not very efficient at capturing non-linear interactions between structural OCT data and clinical variables. The results confirm the hypothesis that MMFN provides a more in-depth representation of glaucoma progression by using a hybrid modelling of spatial retinal and systemic clinical data.

Table 1. Performance comparison on glaucoma progression forecasting (hold-out test set)

Model	R^2	RMSE	MAE	MSE
MMFN (Proposed)	0.835	3.21	2.05	10.30
IncepMambaNet	0.782	3.58	2.34	12.82
FairDist	0.721	3.96	2.71	15.68
CNN–LSTM	0.689	4.21	2.95	17.72
CNN	0.635	4.63	3.28	21.43
ANN	0.552	5.1	3.8	26.01

Note: MMFN = multimodal Mamba fusion network, RMSE = root mean squared error, MAE = mean absolute error, MSE = mean squared error, CNN = convolutional neural network, ANN = artificial neural network, LSTM = long short-term memory, R^2 = coefficient of determination.

4.3 Cross-validation-based generalization analysis

A 10-fold cross-validation strategy was adopted to further confirm the strength and ability of the suggested model to be generalized. Table 2 presents the results in the form of mean and standard deviation (Mean ± SD), which gives information on the predictive accuracy and stability between the data divided into various splits.

Table 2. 10-fold cross-validation performance on glaucoma progression forecasting

Model	R^2 (Mean ± SD)	RMSE (Mean ± SD)	MAE (Mean ± SD)	MSE (Mean ± SD)
MMFN (Proposed)	0.832 ± 0.018	3.24 ± 0.11	2.08 ± 0.07	10.52 ± 0.35
IncepMambaNet	0.776 ± 0.025	3.61 ± 0.19	2.36 ± 0.14	13.03 ± 0.52
FairDist	0.715 ± 0.032	4.02 ± 0.27	2.74 ± 0.21	16.16 ± 0.74
CNN–LSTM	0.681 ± 0.041	4.28 ± 0.35	2.98 ± 0.29	18.32 ± 0.91
CNN	0.628 ± 0.048	4.71 ± 0.41	3.32 ± 0.33	22.18 ± 1.12
ANN	0.545 ± 0.055	5.18 ± 0.48	3.85 ± 0.41	26.84 ± 1.35

Note: MMFN = multimodal Mamba fusion network, RMSE = root mean squared error, MAE = mean absolute error, MSE = mean squared error, CNN = convolutional neural network, ANN = artificial neural network, Mean ± SD = mean and standard deviation, R^2 = coefficient of determination.

To provide a more comprehensive assessment of the reliability of the cross-validation results, the reported performance is summarized using mean ± standard deviation across the 10 folds. The 10-fold cross-validation itself provides repeated evaluation across different training and validation partitions, thereby allowing the variability of model performance to be assessed. The relatively small standard deviations observed for MMFN (R^2: 0.018, RMSE: 0.11, MAE: 0.07, and MSE: 0.35) indicate limited variation across the evaluated folds. These results provide an estimate of the model's performance stability under different data partitions rather than relying solely on a single train-test split.

The MMFN model achieves a mean R^2 score of 0.832 ± 0.018, which is the highest among all compared methods. Importantly, the very low standard deviation (±0.018) indicates that the model maintains consistent performance across varying subsets of patient data, demonstrating strong generalization capability. In contrast, models such as CNN–LSTM exhibit higher variability (±0.041), suggesting instability in capturing disease progression patterns across different folds. From an error perspective, MMFN achieves the lowest MAE (2.08 ± 0.07), indicating high precision in predicting continuous glaucoma progression metrics such as MD and RNFL thickness changes. This represents a clear improvement over IncepMambaNet (MAE: 2.36 ± 0.14), further validating the fine-grained predictive capability of the proposed architecture. Additionally, MMFN achieves the lowest MSE (10.52 ± 0.35) and RMSE (3.24 ± 0.11) among the evaluated models, indicating lower prediction error and consistent performance across the 10-fold cross-validation setting. These results indicate that the model effectively mitigates extreme deviations and maintains stable performance across heterogeneous patient profiles. In comparison, conventional CNN-based models exhibit significantly higher error values (e.g., CNN MSE: 22.18 ± 1.12), reflecting their limited ability to generalize across varying anatomical and clinical conditions.

4.4 Ablation study

An ablation analysis was conducted to explore the contribution of key architectural components of the proposed framework by systematically omitting and substituting the various components in the framework. It draws on the following four key components: Vision Mamba-based structural feature extractor, tabular encoder, cross-attention fusion mechanism, and fairness discriminator. The complete MMFN configuration is the reference, while the other

configurations are ablated by changing one component of the experiment at a time and keeping all others the same. Each configuration is tested on the same hold-out test set, with metrics of MAE, RMSE, MSE, and R^2. The aim of this analysis is to identify whether each architectural component brings significant contributions to the overall prediction of glaucoma progression and identify the contribution of multimodal representation learning, feature fusion, and fairness-aware learning in the proposed framework.

4.5 Integrated technical discussion

The high performance of MMFN can be explained by the multiple-mode fusion mechanism and the Mamba-based state-space modeling. The model combines high-dimensional OCT RNFLT maps and patterned clinical features (MD, TD, demographic attributes) to capture spatial and systemic components of glaucoma progression. Mamba-based architecture improves the capability of modeling long-range dependencies and contextual relationships, which are necessary for progressive diseases. This gives an obvious benefit compared to traditional CNN- and LSTM-based models, which are restricted in their ability to capture global dependencies and multimodal interactions. Also, when training and validation measures are continuously equalized in all the plots, it indicates that MMFN automatically tends to avoid overfitting and is capable of generalizing to unknown data. The small variance during cross-validation further indicates stable performance across the evaluated patient population. In contrast to conventional classification methods which group patients into progression categories, the presented regression-based model allows to predict the glaucoma progression trajectories continually. It can be used to accurately estimate the rates of RNFL thinning and visual field loss, which can be used to aid in the early diagnosis, stratification of risks, and the planning of treatment.

4.6 Demographic fairness analysis

In order to quantify the effectiveness of the adversarial fairness mechanism, the predictive performance of MMFN was further evaluated in each of the demographic sub-populations using the hold-out test set. The MAE, RMSE, and R^2 were used to analyze the group-wise performance based on gender, race, and ethnicity.

The results show that MMFN has comparable prediction accuracy for the different demographic groups tested. The model's MAE and RMSE were 2.04 and 3.20 for male patients and 2.07 and 3.24 for patients who were female, respectively. The R^2 for the two were 0.837 and 0.832, respectively. The model achieved a MAE of 2.03, 2.14, and 2.09 for White, Black, and Asian patients, respectively, and R^2 of 0.839, 0.817, and 0.826 for each of these racial groups, respectively. The MAE values for ethnicity were 2.04 for non-Hispanic patients and 2.10 for Hispanic patients, while the R^2 values were 0.836 and 0.824, respectively.

The differences between subgroups were also quantified for each evaluation metric by calculating the maximum difference between groups of individuals at each metric. The gender, race, and ethnicity observed MAE differences were 0.03, 0.11, and 0.06, respectively. The same applies to the RMSE discrepancies, which are 0.04, 0.15, and 0.09, respectively, and the corresponding R^2 discrepancies, which are 0.005, 0.022, and 0.012, respectively.

The results indicate relatively consistent prediction performance across the different demographic groups assessed. The relatively small differences in MAE, RMSE, and R^2 indicate reduced performance disparity across gender, race, and ethnicity within the evaluated Harvard-GDP test population. This analysis provides an empirical assessment of subgroup performance consistency in conjunction with the adversarial fairness mechanism incorporated into MMFN. However, these findings do not demonstrate complete removal of demographic bias or general demographic fairness, as the analysis is limited to the demographic groups available in the evaluated dataset.

4.7 SHapley Additive exPlanations-based interpretability analysis for glaucoma progression prediction

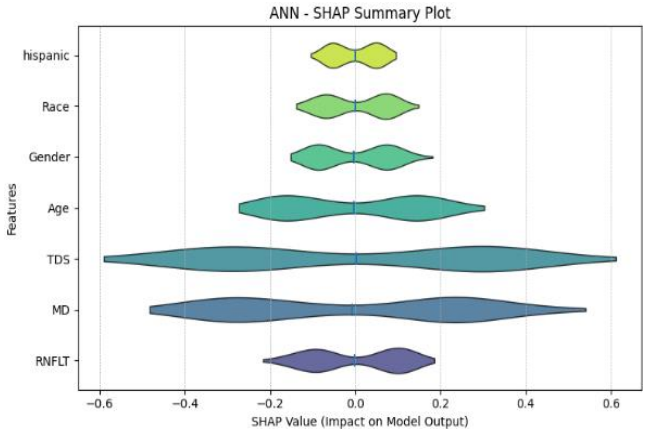
The SHAP-based violin summary plots in Figures 3(a)–(f) give a quantitative explanation of the contribution of features in all the models that predict the progression of glaucoma. The distributions capture the extent of feature influence and spread, allowing the analysis of how each model exploits the structural, functional, and demographic features.

In Figure 3(a) which corresponds to ANN model, the SHAP distributions are heavily centred on TD and MD with relatively weaker spreads (approximately ±0.6), showing that the model is one that is highly dependent on the functional visual field parameters to give predictions. Conversely, the contribution of RNFLT itself is also small and limited ($\approx \pm 0.2$), which is indicative of the lack of structural OCT information used. Such an imbalance is a manifestation of the natural shortcoming of ANN in representing multimodal relationships, which is mainly applied to tabular inputs. Figure 3(b) shows that the CNN model has a very skewed SHAP distribution with RNFLT dominating and a much broader spread ($= \pm 1.1$), which implies that it is very sensitive to structural retinal features. The other characteristics, such as MD, TD, and demographic variables, contribute little ($\approx \pm 0.1$), and this proves that the CNN is a mere structural model. Although it is effective in modelling the patterns of RNFL thinning, the unimodal dependence restricts its capacity to integrate functional indicators of disease progression.

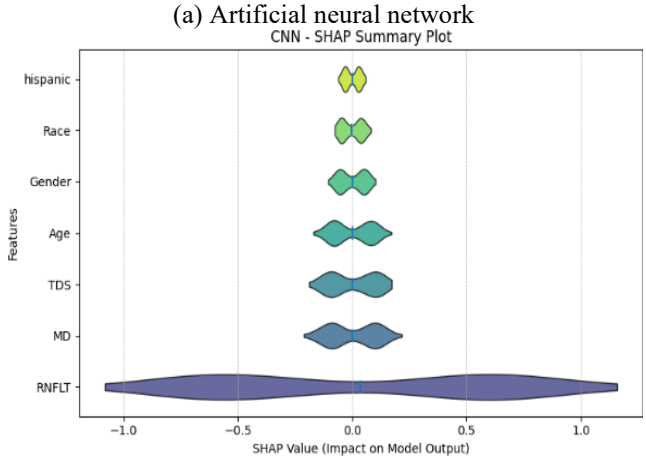
CNN–LSTM model in Figure 3(c) shows better integration of features, with both RNFLT ($= \pm 0.8$) and MD ($= \pm 0.3$) being significant terms in the prediction. The addition of temporal modeling provides a better representation of functional deterioration, but the SHAP distribution continues to show partial domination of structural features, implying the multimodal fusion is not complete. The FairDist model, in Figure 3(d), also exhibits a comparable trend, with the model heavily relying on RNFLT ($\approx \pm 1.0$) and the relatively insignificant effects of MD and TD ($\approx \pm 0.15$). The model uses fairness-aware learning; however, the SHAP distributions demonstrate that the use of the features is still biased according to the structural features, which means that the clinical variables are not well incorporated into the model. Figure 3(e) shows that the IncepMambaNet model has a more distributed SHAP pattern, with the contribution of RNFLT ($\approx \pm 1.0$) still being the most significant, but the contributions of MD and TD ($\approx \pm 0.2$–0.3) are stronger than in the previous models. This is an illustration of the benefit of multi-scale and state-space modeling, but the fact that structural dominance still persists suggests that feature fusion is not yet completely balanced.

The proposed MMFN model shown in Figure 3(f) has, on the other hand, a more balanced SHAP profile over the major

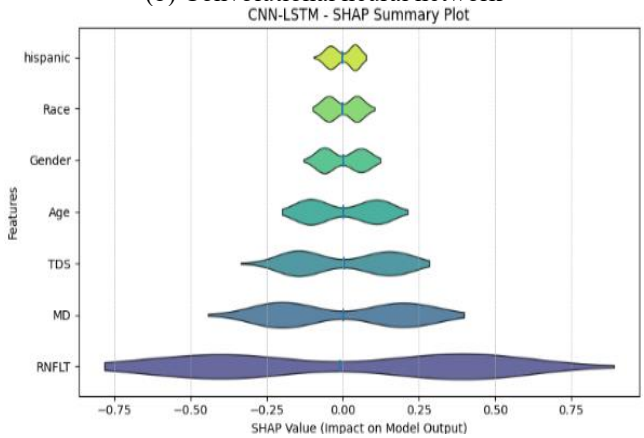
features of the structure, function, and demographics. Both qualitative and quantitative assessment of the contribution of structural and functional features to the model predictions is provided by the SHAP values of RNFLT (≈ ±0.8), MD (≈ ±0.5), and TD (≈ ±0.5). Other demographic variables such as Age (≈ ±0.3) and categorical variables such as Gender, Race, and Hispanic status (≈ ±0.2) also have measurable effects on the model output, which means that patient-specific information is integrated in the prediction process. These SHAP values, however, are feature attributions at the model level, and do not imply that there are causal relationships between the demographics and the progression of glaucoma. Like that, the size of the SHAP values is not enough to determine whether there is or is not a demographic bias. A relatively distributed SHAP profile indicates that MMFN does not exclusively use a single feature or modality to make decisions. This analysis is complementary to that reported elsewhere in the manuscript, where predictive performance and analysis of the demographic groups were reported.



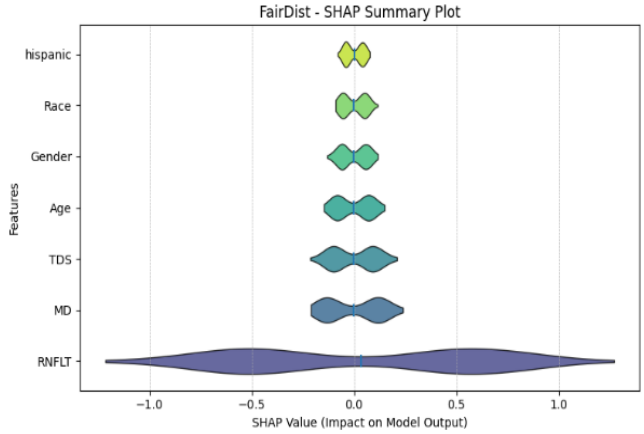
(a) Artificial neural network



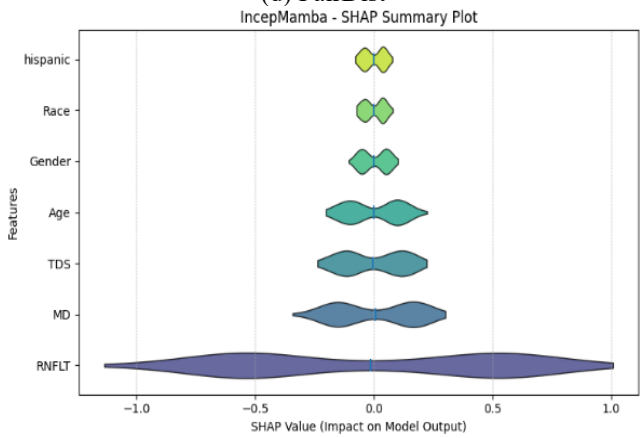
(b) Convolutional neural network



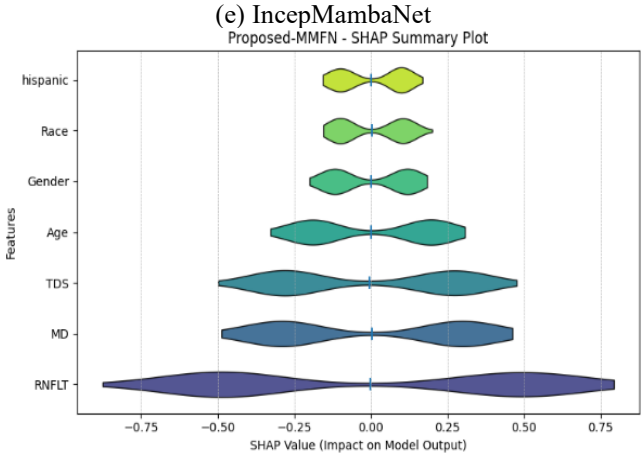
(c) CNN–LSTM



(d) FairDist



(e) IncepMambaNet



(f) Multimodal Mamba fusion network (proposed)

Figure 3. SHapley Additive exPlanations (SHAP) analysis

4.8 Computational efficiency analysis for glaucoma progression prediction

When evaluating deep learning models, computational efficiency is an important consideration, particularly for potential future use in time-sensitive clinical workflows. For this reason, the inference latency and relative computational complexity of MMFN and baseline models were examined for the experimental setting in this study.

As shown in Figure 4, the inference-time analysis reveals that MMFN takes about 20 ms per inference, whereas CNN takes around 22 ms per inference, FairDist takes around 30 ms per inference, IncepMambaNet takes around 35 ms per inference and CNN–LSTM takes around 40 ms per inference. In the considered experimental setting, MMFN thus shows the lowest measured inference latency when compared with the other deep learning models. According to the model

complexity analysis, MMFN is a moderate complexity model in comparison with the evaluated baseline architectures. The combination of state-space-based feature processing and compact multimodal feature representations and the use of cross-attention fusion, contribute to the relatively low inference latency of MMFN. The model combines the OCT-derived RNFL thickness maps with functional and demographic data without having to use longitudinal image sequences for inference. This indicates the computational efficiency of the multimodal prediction model within the tested experimental context. To provide a more complete characterization of the computational requirements of MMFN, the model was evaluated in terms of its computational complexity and resource requirements in addition to inference latency.

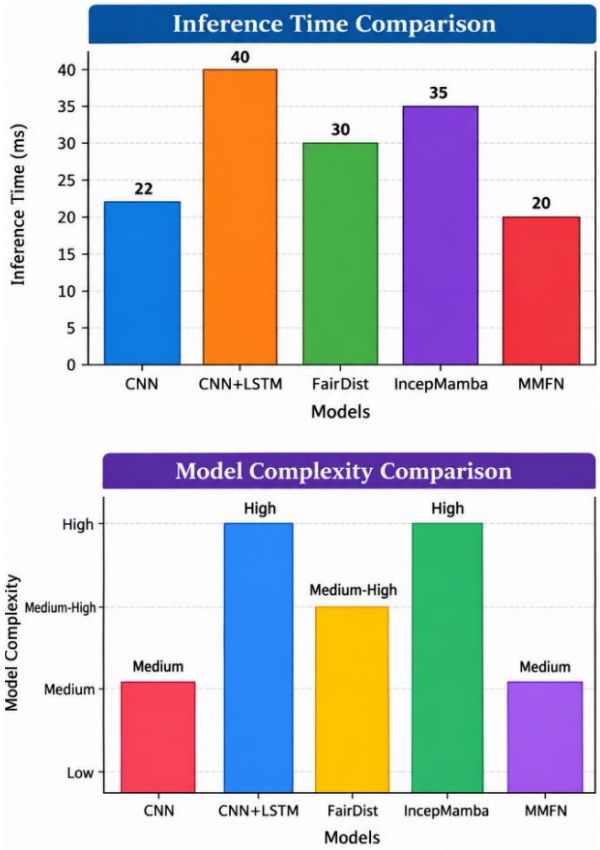


Figure 4. Inference time and model complexity of proposed and existing models

The analysis takes into account the number of trainable parameters, the number of floating-point operations (FLOPs), and memory usage of the model on the experimental hardware configuration. The number of parameters represents the size of the learnable parts of the multimodal architecture, and the number of FLOPs represents an approximation of the number of computations needed to process an input sample. Memory usage is a metric that can be used to estimate the resources needed when executing models. These are implemented alongside the inference-time analysis and are a more detailed evaluation of the computational properties of MMFN. This reported inference time, however, is not meant to be taken as evidence for real-time clinical deployment, but rather as a computational evaluation. This study did not have external clinical validation, an evaluation on a prospective basis or direct comparison with ophthalmologists. The observed computational efficiency demonstrates the computational feasibility of MMFN within the tested experimental setting; however, it should not be interpreted as evidence of real-time clinical applicability or deployment. Further external and prospective clinical validation would be required to assess its suitability for real-world clinical decision-support applications.

5. CONCLUSIONS

In this paper, we propose an MMFN to predict glaucoma progression continuously given multimodal ophthalmic information. The proposed framework is based on a combination of structural data obtained from OCT-derived thickness maps of the RNFL and functional data from visual field measurements such as MD and TD, along with demographic data. Structural feature extraction is done using a Vision Mamba model, and additional clinical information is extracted with a dedicated tabular encoder and cross-attention. An adversarial gradient-reversal mechanism is also added to minimize the reliance of the learned representation on the demographic characteristics. When tested on the hold-out test set, MMFN had an R^2 of 0.835, RMSE of 3.21, MAE of 2.05, and MSE of 10.30, which were all better than the baseline models evaluated. The 10-fold cross-validation results further showed consistent performance, with a mean R^2 of 0.832 ± 0.018, RMSE of 3.24 ± 0.11, MAE of 2.08 ± 0.07, and MSE of 10.52 ± 0.35. Model-level information on the contribution patterns of structural, functional, and demographic features was given by the SHAP analysis. Moreover, the inference latency of MMFN was approximately 20 ms under the experimental hardware and software configuration, indicating relatively low computational latency among the evaluated models. However, a few limitations should be taken into account from these results. First, the model was evaluated with the Harvard-GDP dataset, and there was no external validation with independent sets of data. Thus, the validity of the reported performance in other clinical samples and settings needs to be confirmed. Second, this study was not prospective clinical or a direct comparison with an ophthalmologist. Third, the demographic subgroups are not equal in size, which may make estimates of subgroup performance disparities less reliable. Finally, the measured computational efficiency should not be interpreted as evidence of real-time clinical deployment. Future work will therefore focus on external validation using large, multicenter datasets, prospective clinical evaluation, and comparison with clinical experts. A possible extension of the modeling is the incorporation of patient information that varies with time, which can be used to model the progression of the disease. Other imaging techniques such as OCTA and fundus imaging could also be explored as additional imaging modalities to complement structural and vascular data. These are the studies that will be crucial in assessing the predictive performance, representativeness for the demographic, interpretability, and computational efficiency demonstrated in the current experiments will be retained in independent and clinically representative settings.

REFERENCES

[1] Luo, Y., Shi, M., Tian, Y., Elze, T., Wang, M. (2023). Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning. In 2023 IEEE/CVF

International Conference on Computer Vision (ICCV), Paris, France, pp. 20414-20425. https://doi.org/10.1109/ICCV51070.2023.01872

[2] Pan, X., Xiong, Z., Qiu, D., et al. (2026). A semi-supervised reinforcement learning framework incorporating the multi-scale IncepMambaNet network for glaucoma progression prediction. Interdisciplinary Sciences: Computational Life Sciences, 18(2): 451-470. https://doi.org/10.1007/s12539-025-00739-x

[3] Oh, R., Kim, H., Kim, T.W., Lee, E.J. (2025). Predictive modeling of rapid glaucoma progression based on systemic data from electronic medical records. Scientific Reports, 15(1): 13101. https://doi.org/10.1038/s41598-025-97344-1

[4] Hussain, S., Chua, J., Wong, D., et al. (2023). Predicting glaucoma progression using deep learning framework guided by generative algorithm. Scientific Reports, 13(1): 19960. https://doi.org/10.1038/s41598-023-46253-2

[5] Afolabi, S.O., Gheisi, L., Shan, J., Shen, L.Q., Wang, M., Shi, M. (2025). Equity-enhanced glaucoma progression prediction from OCT with knowledge distillation. npj Digital Medicine, 8(1): 477. https://doi.org/10.1038/s41746-025-01884-9

[6] Kurysheva, N.I., Rodionova, O.Y., Pomerantsev, A.L., Ponomareva, S.I., Golubnitschaja, O. (2025). Multimodal AI-based modeling of glaucoma progression: A 3PM-guided approach integrating structural, functional, and vascular patterns. EPMA Journal, 16(4): 819-830. https://doi.org/10.1007/s13167-025-00429-0

[7] Kang, E., Park, J.H., Yoo, C., Kim, Y.Y. (2024). Comparison of intraocular pressure fluctuation and glaucoma progression rate between phakic and pseudophakic eyes in pseudoexfoliation glaucoma. Scientific Reports, 14(1): 6. https://doi.org/10.1038/s41598-023-49099-w

[8] Oltramari, L., Mansberger, S.L., Souza, J.M.P., de Souza, L.B., de Azevedo, S.F.M., Abe, R.Y. (2024). The association between glaucoma treatment adherence with disease progression and loss to follow-up. Scientific Reports, 14(1): 2195. https://doi.org/10.1038/s41598-024-52800-2

[9] Târcoveanu, F., Leon, F., Lisa, C., et al. (2024). The use of artificial neural networks in studying the progression of glaucoma. Scientific Reports, 14(1): 19597. https://doi.org/10.1038/s41598-024-70748-1

[10] Beniz, L.A., Jammal, A.A., da Costa, D.R., Mariottoni, E.B., Swaminathan, S.S., Medeiros, F.A. (2024). Impact of race and ethnicity on glaucoma progression detection by perimetry and optical coherence tomography. Scientific Reports, 14(1): 30752. https://doi.org/10.1038/s41598-024-80481-4

[11] Mandal, S., Jammal, A.A., Malek, D., Medeiros, F.A. (2024). Progression or aging? A deep learning approach for distinguishing glaucoma progression from age-related changes in OCT scans. American Journal of Ophthalmology, 266: 46-55. https://doi.org/10.1016/j.ajo.2024.04.030

[12] Mariottoni, E.B., Datta, S., Shigueoka, L.S., et al. (2023). Deep learning–assisted detection of glaucoma progression in spectral-domain OCT. Ophthalmology Glaucoma, 6(3): 228-238. https://doi.org/10.1016/j.ogla.2022.11.004

[13] Nouri-Mahdavi, K., Mohammadzadeh, V., Rabiolo, A., Edalati, K., Caprioli, J., Yousefi, S. (2021). Prediction of visual field progression from OCT structural measures in moderate to advanced glaucoma. American Journal of Ophthalmology, 226: 172-181. https://doi.org/10.1016/j.ajo.2021.01.023

[14] Dixit, A., Yohannan, J., Boland, M.V. (2021). Assessing glaucoma progression using machine learning trained on longitudinal visual field and clinical data. Ophthalmology, 128(7): 1016-1026. https://doi.org/10.1016/j.ophtha.2020.12.020

[15] Yousefi, S., Kiwaki, T., Zheng, Y., et al. (2018). Detection of longitudinal visual field progression in glaucoma using machine learning. American Journal of Ophthalmology, 193: 71-79. https://doi.org/10.1016/j.ajo.2018.06.007

[16] Xu, L., Asaoka, R., Kiwaki, T., et al. (2020). Predicting the glaucomatous central 10-degree visual field from optical coherence tomography using deep learning and tensor regression. American Journal of Ophthalmology, 218: 304-313. https://doi.org/10.1016/j.ajo.2020.04.037

[17] Yasuda, S., Hasegawa, T., Yoshimoto, S., et al. (2026). Deep learning–based prediction of visual field mean deviation from numeric OCT data in glaucoma. Journal of Imaging Informatics in Medicine, 1-14. https://doi.org/10.1007/s10278-026-01866-7

[18] Ohn, K., Hwang, J., Jung, J., Whang, W.J. (2026). Predicting visual field loss in glaucoma using OCT and deep learning: A comparative study of U-Net variants. Ophthalmology Science, 6(6): 101169. https://doi.org/10.1016/j.xops.2026.101169

[19] Cvenkel, B., Sustar, M., Perovšek, D. (2022). Monitoring for glaucoma progression with SAP, electroretinography (PERG and PhNR) and OCT. Documenta Ophthalmologica, 144(1): 17-30. https://doi.org/10.1007/s10633-021-09854-8

[20] Park, K., Kim, H., Moon, S., Lee, J. (2026). Deep learning-based prediction of 24-2 visual field from fundus photographs in glaucoma. Graefe's Archive for Clinical and Experimental Ophthalmology, 1-10. https://doi.org/10.1007/s00417-026-07141-3

[21] Mohammadzadeh, V., Wu, S., Besharati, S., et al. (2024). Prediction of visual field progression with baseline and longitudinal structural measurements using deep learning. American Journal of Ophthalmology, 262: 141-152. https://doi.org/10.1016/j.ajo.2024.02.007

[22] Kamalipour, A., Moghimi, S., Khosravi, P., et al. (2023). Combining optical coherence tomography and optical coherence tomography angiography longitudinal data for the detection of visual field progression in glaucoma. American Journal of Ophthalmology, 246: 141-154. https://doi.org/10.1016/j.ajo.2022.10.016

[23] Herbert, P., Hou, K., Bradley, C., et al. (2023). Forecasting risk of future rapid glaucoma worsening using early visual field, OCT, and clinical data. Ophthalmology Glaucoma, 6(5): 466-473. https://doi.org/10.1016/j.ogla.2023.03.005

[24] Lim, J.W.S., Ramli, N. (2026). Longitudinal evaluation of optical coherence tomography parameters for detecting glaucoma progression across disease severities. BMC Ophthalmology, 26: 264. https://doi.org/10.1186/s12886-026-04790-0

[25] Lan, W., Tang, Z., Yan, X., Wu, Y., Liu, J., Kuang, H. (2026). Glaucoma diagnosis and progression prediction

based on deep multimodal fusion and collaborative attention mechanism. IEEE Journal of Biomedical and Health Informatics, 1-10. https://doi.org/10.1109/JBHI.2026.3698681