Grenze International Journal of Engineering and Technology, June Issue

ChronoDR-Net: A Multimodal Temporal Transformer Framework for Longitudinal Prediction of Diabetic Retinopathy Progression

Gayathri. K[1] and A. Sasi Kumar[2]
[1]Research Scholar, School of Science and Computer Studies, CMR University Bangalore, India.
Email: k.gayathri@cmr.edu.in
[2]Associate Professor, School of Science and Computer Studies, CMR University Bangalore, India.
Email: sasikumar.a@cmr.edu.in

***Abstract*— Forecasting diabetic retinopathy progression cannot be reliably addressed as a sequence of isolated classification tasks, but instead requires modeling disease evolution as a longitudinal and patient-specific process.**

Most existing approaches continue to rely on static fundus image analysis or limited forms of temporal aggregation, which restrict their ability to capture long-range dependencies and conditionally evolving interactions between retinal pathology and systemic risk factors. In this study, diabetic retinopathy progression is treated as a multimodal temporal learning problem in which future disease severity depends jointly on historical retinal appearance and longitudinal clinical context.

ChronoDR-Net is introduced as a multimodal temporal transformer framework designed to integrate fundus image sequences with visit-level clinical metadata through structured cross-modal attention. Spatial retinal features are extracted using a Vision Transformer, while temporal self-attention is employed to model disease dynamics across irregularly sampled clinical visits.

Clinical variables are represented as contextual tokens that modulate the temporal saliency of retinal features through bidirectional attention, allowing patient-specific progression signals to emerge naturally from the data. Multi-horizon forecasting is performed under a jointly optimized objective that constrains ordinal progression, temporal smoothness, and cross-modal alignment. Evaluation on a longitudinal subset of the EyePACS dataset, followed by external validation on a multi-center cohort, indicates that ChronoDR-Net achieves a Quadratic Weighted Kappa of 0.915 for two-step-ahead severity prediction, outperforming recurrent and metadata-agnostic temporal baselines by a clear margin.

Ablation experiments further suggest that deep cross-modal interaction plays a central role in modeling heterogeneous progression trajectories. Taken together, these findings highlight the importance of explicit temporal and conditional multimodal modeling for reliable diabetic retinopathy forecasting and support the use of ChronoDR-Net in personalized longitudinal screening settings.

***Index Terms*— Diabetic Retinopathy, Longitudinal Disease Modeling, Fundus Imaging, Multimodal Learning, Temporal Transformers, Cross-Modal Attention.**

Grenze ID: 01.GIJET.12.2.64

I. INTRODUCTION

Diabetic retinopathy (DR) develops as a progressive microvascular complication which affects diabetes mellitus patients. The condition represents one of the main reasons which lead to avoidable vision impairment among adults who work throughout the world. The increasing number of diabetes cases results in higher DR detection rates which create ongoing demands for both ophthalmic screening initiatives and medical facilities. The available treatment options succeed when doctors identify disease progression before patients experience permanent retinal damage. The process which individual patients experience for DR evolution needs to be understood because it directly affects the achievement of better visual results throughout the time period which follows.

In clinical practice, DR assessment is inherently longitudinal. Patients undergo repeated retinal examinations, and clinicians interpret current retinal findings in relation to previous visits, considering both visible retinal changes and systemic risk factors such as glycemic control, blood pressure, and duration of diabetes. The medical team determines when to continue patient monitoring and start treatment and send patients to other doctors based on their assessment of how the patient will develop their medical condition. The medical assessment process uses temporal indicators to determine patient outcomes because they predict future results better than medical professionals who assess patients based on their condition at one moment.

The majority of large-scale screening programs together with their associated public datasets operate by design for cross-sectional studies despite the ongoing longitudinal decision-making process. The assessment of fundus images occurs through independent evaluation which results in each image receiving a severity label that represents one specific time point. Automated screening systems have achieved successful deployment through this paradigm but the system lacks capacity to forecast disease evolution which has become essential for delivering personalized treatment.

The precise forecasting of DR progression brings essential social benefits according to its wider healthcare effects. Accurate prediction capabilities will enable healthcare systems to implement adaptive screening methods which will decrease unnecessary medical appointments for patients who have low risk and will enable medical staff to focus on treating patients who have high risk of rapid disease progression. Resource-constrained environments can utilize these functions to enhance screening program effectiveness while preserving clinical safety standards. The need for computational models exists because static image analysis solutions fail to deliver forward-looking diabetic retinopathy progression assessments that meet patient needs.

A. Technical Gap

The existing automated DR detection and grading systems have achieved high accuracy through deep learning techniques however their current methods for processing data show considerable limitations. The models process each fundus image as a distinct sample which leads to predictions based on retinal characteristics observed during one medical appointment. The model fails to understand disease progression because it uses fixed assumptions which prevent it from tracking how medical conditions develop through different stages.

Researchers have conducted multiple studies to develop methods that use temporal data through the implementation of sequential learning models which include recurrent neural networks together with shallow temporal aggregation methods. The methods face difficulties because they cannot maintain extended temporal dependencies which emerge during extended disease progression time periods. Recurrent models demonstrate performance vulnerabilities which occur because their output depends on the length of input sequences while actual screening conditions that occur in clinical settings create unpredictable patterns of patient visits.

The technical challenge related to clinical metadata integration represents another obstacle. DR progression depends on systemic risk factors which medical experts use to assess patient risk but researchers currently use these factors only for basic feature merging methods and for later integration stages. Clinical variables fail to affect retinal feature processing because they spent time with patients whose retinal patterns changed which established an inflexible method to process their retinal data.

Longitudinal ophthalmic research relies on diverse imaging conditions together with various devices which result in different levels of image quality between distinct imaging sessions. Temporal models that do not explicitly account for this variability risk learning spurious temporal signals driven by imaging artifacts rather than true pathological change. Progression modeling needs to establish authentic disease development separation from data acquisition artifacts together with clinical operation exercises.

The current deep learning methods face technical limitations which prevent them from achieving accurate diabetic retinopathy progression predictions that match actual clinical practice requirements.

B. Research Gap

The field of medical imaging has seen increased interest in temporal and multimodal learning yet researchers still lack effective methods for predicting diabetic retinopathy progression. The existing research studies treat temporal modeling and multimodal integration as distinct challenges instead of recognizing their interlinked role in disease progression.

The temporal models depend on retinal image sequences to track progression yet they assume retinal images provide complete information about progression. The multimodal methods treat clinical metadata as secondary data that does not influence the learning process for temporal features. The clinical separation of risk factors tracks progression probabilities yet it also determines how retinal changes should be understood across various time periods.

Most previous studies investigate short-term predictions which include upcoming disease severity assessments for the next clinical appointment. The existing framework provides valuable insights yet it does not represent DR management which requires clinicians to monitor progression patterns during multiple upcoming appointments. Researchers have conducted limited studies on multi-horizon forecasting which includes assessing how prediction accuracy changes as the forecast period extends.

Researchers have yet to create comprehensive systems that can handle irregularly timed longitudinal data without requiring them to maintain regular temporal intervals. The existing models do not transfer successfully to actual clinical environments.

The implemented system requires a comprehensive modeling framework which should treat diabetic retinopathy progression as a conditional sequence learning challenge by combining two elements. The system should apply temporal retinal data together with personalized patient clinical information.

C. Contributions

This research addresses existing gaps through the following contributions.

- Longitudinal Formulation of DR Progression: Diabetic retinopathy progression is formulated as a patient-specific longitudinal learning task, in which future disease severity is predicted based on sequences of fundus images and corresponding visit-level clinical metadata.
- ChronoDR-Net Framework: The research introduces a new multimodal temporal architecture known as ChronoDR-Net. The framework combines Vision Transformer-based spatial feature extraction with temporal self-attention to model disease evolution across irregularly spaced clinical visits.
- Structured Cross-Modal Interaction: The system uses a structured cross-modal attention mechanism to integrate clinical metadata, which allows systemic risk factors to activate temporal retinal feature representations and creates patient-specific progression models.
- Multi-Horizon Progression Forecasting: The proposed approach allows disease severity predictions for several upcoming visits, which helps researchers study short-term and long-term disease progression patterns using one unified system.
- Comprehensive Evaluation: The research conducted extensive experiments, which included ablation studies and external validation, to show how temporal modeling and multimodal interaction effectively capture different diabetic retinopathy progression patterns.

II. LITERATURE REVIEW

A. Deep Learning for Retinal Disease Analysis

Regarding deep learning applications in ophthalmic imaging, initial works were devoted to automating the detection and grading of diabetic retinopathy using single fundus photographs. CNN-based architectures demonstrated strong performance, particularly in detecting lesion-level features such as microaneurysms, hemorrhages, and exudates, thereby establishing deep learning as a scalable screening solution for diabetic retinopathy. Several comprehensive reviews and systematic studies documented the evolution of these approaches, highlighting the effectiveness of architectural refinements, attention mechanisms, and data augmentation strategies in improving model performance [1], [5], [12].

Despite these advancements, a significant research gap remains, as most detection-oriented systems adopt a static paradigm where each fundus image is analyzed independently. Approaches employing depthwise separable convolutions and multi-view attention mechanisms have shown improved disease severity classification, especially in early stages, yet they remain fundamentally cross-sectional in nature [15]. Vision transformer-based models and hybrid transformer–capsule network architectures have further enhanced contextual feature

representation, achieving robust performance across large imaging variations [13]. However, these methods primarily address diagnostic accuracy and fail to model longitudinal disease progression.

B. Longitudinal Modeling and Disease Progression Prediction

The chronic and progressive nature of diabetic retinopathy has motivated increasing interest in longitudinal modeling approaches that leverage multi-visit retinal data. Unlike cross-sectional diagnostic systems, longitudinal frameworks aim to predict future disease states by incorporating both current retinal appearance and historical progression patterns. Empirical evidence confirms the clinical relevance of temporal modeling, as deep learning systems trained on sequential fundus images have demonstrated superior progression prediction accuracy compared to single-image models [10]. Large-scale clinical investigations further emphasize the importance of long-term modeling. Studies involving patient cohorts followed over multiple years have shown that deep learning systems can estimate not only the likelihood of disease progression but also the expected time to progression with clinically meaningful precision [4],[14]. These findings underscore that disease progression prediction is inherently a temporal inference problem rather than a static classification task.

To address this challenge, transformer-based architectures have emerged as a dominant modeling paradigm due to their capacity to capture long-range temporal dependencies. Temporal self-attention enables direct interaction between non-adjacent clinical visits, mitigating information compression and memory degradation issues commonly associated with recurrent neural networks [19]. Recent methods extend this capability by incorporating continuous-time representations and severity-aware encoding schemes, allowing robust performance under irregular follow-up intervals while preserving disease stage transitions [19].

In parallel, generative modeling techniques have been explored for forecasting future retinal appearances. Generative adversarial networks capable of synthesizing realistic future fundus images offer an alternative perspective on progression prediction and uncertainty modeling [9]. However, despite their methodological appeal, such approaches remain insufficiently validated in clinical workflows, and their role in decision support systems is yet to be firmly established.

C. Multimodal Learning and Clinical Context Integration

The progression of diabetic retinopathy is influenced not only by retinal pathology but also by broader systemic and demographic factors. Consequently, multimodal learning frameworks that integrate imaging data with clinical metadata have gained considerable attention. Population-scale studies have demonstrated improved progression prediction by combining longitudinal retinal images with multiple risk factors [11], [8].

Advanced multimodal systems incorporating heterogeneous imaging modalities such as fundus photography and optical coherence tomography have shown enhanced disease characterization for complex retinal conditions [18]. Moreover, telemedicine-oriented investigations indicate that multimodal AI systems can significantly improve screening and monitoring capabilities, particularly in resource-limited settings [7]. Nevertheless, most existing multimodal frameworks rely on simplistic early or late fusion strategies, limiting the depth of interaction between clinical context and temporal disease features.

D. Federated Learning, Explainability, and Clinical Deployment

Real-world deployment of artificial intelligence systems in ophthalmology requires both institutional robustness and interpretability of predictions. Federated learning enables collaborative model training across multiple screening centers while preserving data privacy and has demonstrated effectiveness for diabetic retinopathy detection using convolutional neural networks and vision transformers [2], [6]. For clinical adoption, explainability remains a critical requirement [3]. Recent research emphasizes that predictive models must align attention mechanisms with clinically relevant retinal structures and provide transparent, medically interpretable reasoning [16]. Such explainable frameworks foster clinician trust and facilitate the integration of AI-based disease progression systems into routine practice. Additionally, transformer-based natural language processing techniques have been applied to analyze unstructured clinical records, enabling multimodal reasoning across imaging and textual data for diabetic retinopathy management [17].

E. Summary of Research Gaps

The reviewed literature highlights several unresolved challenges. First, most high-performance deep learning systems remain limited to static diagnosis or short-term prediction, lacking robust mechanisms for modeling long-term disease progression. Second, existing multimodal approaches underutilize clinical information due to shallow fusion strategies that fail to capture complex temporal interactions. Third, explainability and robustness are often treated as post-hoc additions rather than integral components of model design. These limitations

underscore the need for unified frameworks that jointly model longitudinal retinal sequences and clinical data while providing transparent and trustworthy progression predictions suitable for real-world screening environments.

III. METHODOLOGY

A. Longitudinal Dataset Construction

The modeling of diabetic retinopathy progression requires temporal coherence at the patient level, which is not usually provided in most large-scale datasets of fundus imaging. Public datasets like EyePACS are predominantly structured for cross-sectional grading, thus images are considered as independent samples. Consequently, a well-structured reconstruction of patient-wise temporal sequences is necessary to allow for longitudinal forecasting.

Patient-Level Sequence Formation

All fundus images of a patient were put together by unique patient identifiers and then ordered according to the timestamps of acquisition. Suppose $\{I_t\}_{t=1}^{T}$ represents the ordered sequence of retinal images for one patient, where T is the total number of clinical visits available. The visit-level disease severity annotations and clinical metadata were then correspondingly aligned to that image in the sequence. The patients with only two visits separated by time were the only ones left, thus making it possible to model the disease's progression based on observable time change rather than inferring from one visit.

This step of filtering indicates one of the main assumptions in modeling: that progression cannot be inferred without longitudinal evidence. Allowing for the presence of single-visit patients would limit the task to that of static classification and muddle the learning objectives linked to the temporal aspect.

Handling Irregular Visit Intervals

The intervals of follow-up screening in real-life clinical setups are always irregular and different patients experience different intervals because of accessibility, compliance, and clinical recommendations. Instead of temporally regularizing the sequences artificially, they were allowed to keep their original order without using temporal interpolation. Such a design decision does not bring about any synthetic progression patterns and instead grants the model the ability to capture the temporal relationships from the data that was given for observation. The temporal index t thus becomes an ordering variable and not a crude representation of uniform time spacing. This premise is, in fact, aligned with the clinical situation in which the movement of the disease is considered fully dependent on the total pathological change and the risk exposure to the system rather than on the time that has elapsed in absolute terms only.

Multi-Horizon Target Construction

In order to facilitate the prediction looking ahead, each patient's sequence was converted into many input–target pairs that were related to different forecasting horizons. For example, observations made up to visit T, after that, the future severity labels at visits $T+k$, where $k\geq1$, were taken as the prediction targets. This formulation makes it possible to assess both the short-term and the long-term propagation of the dynamics in a single framework. What is more, in no case were the future labels resorted to for sequence construction or normalization, so temporal causality was strict. All the targets were based only on the visits which took place after the final input observation, thereby preventing any information leakage across the horizons.

Quality Control and Sequence Consistency

In order to minimize noise that was coming from imaging artifacts, the sequences that included images with severe illumination problems or lacking retinal coverage were left out. When there were several images for the same visit, the best one was selected based on the image quality assessment, thus maintaining consistency across the temporal sequences. The clinical metadata that was linked to each visit was standardized to the patient level. This was done to maintain relative trends across time and also to avoid normalization schemes that would secretly encode future information.

Implications for Modeling and Evaluation

This strategy of constructing longitudinal datasets directly backs the proposed modeling framework in three aspects. Patient-level sequence integrity, first, allows us to use temporal self-attention mechanisms to learn long-range dependencies across visits. Secondly, the preservation of irregular visit intervals puts the model to the test of being robust under realistic clinical conditions. Lastly, multi-horizon target generation enables the performance degradation assessment over the forecasting distance, which is very important for the clinical utility

evaluation. These design choices ensure that model performance is true temporal reasoning and not the result of dataset preprocessing or sequence alignment artifacts.

B. Image Preprocessing and Quality Control

The proper modeling of diabetic retinopathy progression highly depends on the consistency in retinal image quality throughout longitudinal studies. The inconsistencies can be caused by the differences in illumination, contrast, field-of-view, and acquisition devices, which may lead to the misinterpretation of noise reduction by the temporal models as a pathological change. Hence, a standardized preprocessing and quality control pipeline was implemented before the feature extraction step.

The fundus images were resized to a fixed spatial resolution to first ensure the same output dimensions for all visits and patients. The pixel intensity values were normalized to minimize the variation between images that resulted from the differences in the camera settings and acquisition conditions. Contrast enhancement, using Contrast Limited Adaptive Histogram Equalization (CLAHE), was performed in order to make the minute retinal structures, especially microaneurysms and vascular abnormalities, more visible. The process increases local contrast but controls noise amplification at the same time, making it vital for the detection of subtle progression-related features. The image quality was evaluated to single out the instances that might have a negative impact on the longitudinal consistency. The excluded images were those with serious illumination artifacts, motion blur, or incomplete retinal coverage. This filtering was done at the sequence level, meaning that visits with unusable images were removed to prevent the introduction of spurious temporal discontinuities. When there were several fundus photographs taken during the same clinical visit, one representative photo was picked according to image quality standards. This way of selecting guarantees that one retinal observation is registered for each visit only and consequently, the imaging data, clinical metadata, and disease severity labels are kept aligned over time. In order to maintain the temporal coherence, preprocessing operations were uniformly applied in a patient sequence across all visits.

Data augmentation was only applied to the transformations that do not affect the disease-related structures and also applied equally across visits to prevent the introduction of artificial temporal variation. The longitudinal sequences created as a result of these preprocessing and quality control measures represent real retinal changes associated with disease progression rather than artifacts caused by imaging variability. This is an important step for the temporal self-attention mechanisms to learn clinically meaningful progression patterns.

C. Feature Extraction using Vision Transformer

The modeling of diabetic retinopathy seen through the eyes of the patient during the course of his disease necessitates the use of spatial representations that would be good enough to not only show the localized lesions but also to connect with the whole image by depicting the long-range dependencies of the fat tissue. In this context, one will find that the fixed receptive field-based convolutional architectures will then limit their ability to generate the pathological patterns dispersed across the entire image by modeling them in a global manner. Using a Vision Transformer (ViT), which has been developed specifically for this purpose, one can extract the retinal features of diabetic patients.

For a fundus picture that has undergone preprocessing $I_t \in \mathbb{R}^{H\times W\times C}$ which was taken during visit t, this image is then divided into P patches that do not overlap. The linear projection of each patch into a d-dimensional embedding takes place and the positional encoding preserves the spatial structure which is added at this stage. The patch embeddings that are obtained undergo processing through self-attention layers stacked on top of each other, where attention is computed by the following method:

$$Attention(Q,K,V) = softmax(\frac{QK^{\top}}{\sqrt{d_k}})V. \qquad (1)$$

In this way, after the processing of the ViT encoder, the result is a collection of the visual tokens each contextualized $Z_t = \{Z_t{}^{(1)}, Z_t{}^{(2)}, \dots, Z_t{}^{(P)}\}$, $Z_t{}^{(P)} \in \mathbb{R}^d$. Afterwards, the whole tokens are kept so that the modules of temporal and cross-modal attention downstream can highlight the spatial areas that are most relevant for the longitudinal disease progression.

D. Temporal Vision Transformer for Longitudinal Modeling

The researchers used a Temporal Vision Transformer (TViT) to track how diseases developed during multiple clinical examinations. The TViT system processes all visit data at once to establish direct connections between distant time periods which distinguishes it from recurrent systems that handle visit data in a sequential pattern but lose important information during extended periods.

Let $\{X_t\}_{t=1}^{T}$ denote the spatial embeddings extracted from fundus images at T visits, where each $X_t \in \mathbb{R}^{N_p \times d}$ represents N_p patch tokens of dimension d. The multiple temporal elements which create the unified temporal sequence

$$Z_{temp} = [X_1, X_2, \dots, X_t] \in \mathbb{R}^{(T.N_p) \times d}. \quad (2)$$

Self-attention layers use temporal positional embeddings to track visit sequences through time while maintaining the ability to identify relationships between visits that occur at both near and far distances. The model uses this method to move initial disease indicators through time which helps it identify long-term disease patterns better than systems that use recurrent methods.

E. Metadata Embedding and Multimodal Integration

The clinical context gets established through the implementation of structured metadata elements. The shared latent space receives each clinical variable $m_{t,j}$ from visit t through a linear embedding which projects it as:

$$v_{t,j} = W_m \cdot m_{t,j} + b_m, W_m \in \mathbb{R}^{d \times 1}, b_m \in \mathbb{R}^d \quad (3)$$

The system creates a unified representation by combining two visual token streams which produce metadata tokens

$$H_t = [X_t; V_t] \in \mathbb{R}^{(N_p + k) \times d}. \quad (4)$$

where k denotes the number of metadata variables. The design enables clinical factors to shape feature representation through direct clinical factor impact instead of using them as extra input sources. The overall architecture of the proposed framework is depicted in Fig. 1.

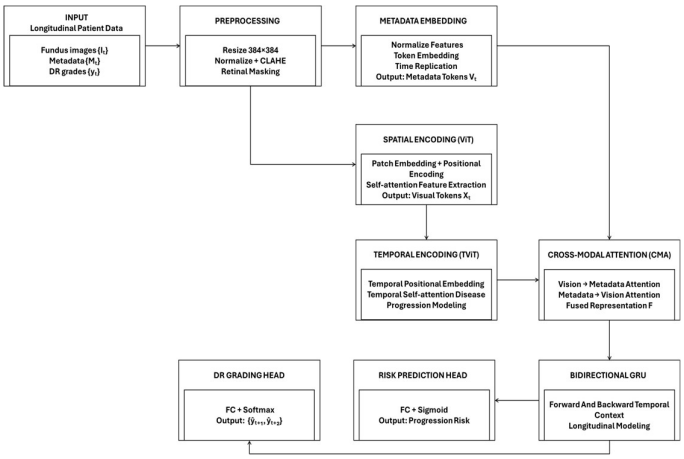

Figure 1. Architecture of ChronoDR-Net Model

F. Cross-Modal Attention Fusion

The mechanism of bidirectional cross-modal attention establishes connection between retinal structures and systemic risk factors. The attention computation for visual tokens (v) and metadata tokens $H(m)$ uses the following equations:

$$A_{vm} = \text{softmax}(\frac{Q_v K_m^T}{\sqrt{d}})V_m, A_{mv} = \text{softmax}(\frac{Q_m K_v^T}{\sqrt{d}})V_v. \quad (5)$$

The unified representation $F = [Avm; Amv]$ establishes conditional relationships between retinal characteristics and medical data which enables interpretable understanding through contextual information that resembles clinical decision-making.

G. Multi-Horizon Forecasting

The model conducts multi-horizon forecasting to produce predictions that have clinical value. The model makes disease severity predictions for future time periods:

$$\hat{y}_{t+k} = \text{softmax}(W_o F_T + b_o), k \in \{1,2\}, \quad (6)$$

using the final observed visit data F_T. The model enables organizations to evaluate risk for both short-term and mid-term periods which aligns with their normal practice of conducting longitudinal assessments and their subsequent monitoring activities.

H. Training Objective and Optimization

The loss function comes as a whole and consists of several objectives:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{ord} + \lambda_2 \mathcal{L}_{temp} + \lambda_3 \mathcal{L}_{align}. \quad (7)$$

in which, $\mathcal{L}_{CE}$: denotes the loss for classification, $\mathcal{L}_{ord}$ acts to preserve ordinal constraints, $\mathcal{L}_{temp}$ is for the purpose of maintaining temporal smoothness, and $\mathcal{L}_{align}$ is assigned to the task of bringing visual and clinical embeddings together. The optimization is implemented through the AdamW optimizer with a cosine learning rate schedule. The stopping criterion for early stopping is the validation performance assessed by the Quadratic Weighted Kappa measure.

I. Evaluation Protocol and Interpretability

INPUT: Multi-visit fundus images $\{I_t\}_{t=1}^{T}$ clinical metadata $\{M_t\}_{t=1}^{T}$, DR grades $\{y_t\}_{t=1}^{T}$
OUTPUT: Predicted DR grades $\{\hat{y}_{t+1}, \hat{y}_{t+2}\}$ for future visits
1: Preprocessing:
2: FOR each visit $t = 1$ to T DO
3: Resize I_t to 384 × 384
4: Normalize intensities to zero mean, unit variance
5: Apply CLAHE (Contrast-Limited Adaptive Histogram Equalization) and retinal masking
6: Optionally enhance vessels
7: END FOR
8: Apply sequence-level augmentation if needed
9: Spatial Feature Extraction (ViT):
10: FOR each visit $t = 1$ to T DO
11: Partition I_t into patches p_i
12: Compute patch embeddings $z_i = W_p p_i + b_p$
13: Add positional embedding $E_{\{pos\}}$
14: Apply ViT self-attention layers to obtain Visual Tokens X_t
15: END FOR
16: Temporal Modeling (TViT):
17: Construct temporal sequence $Z_{\{temp\}} = [X_1, X_2, \dots, X_T]$
18: Add temporal positional embeddings $E_{\{temp\}}$
19: Apply temporal self-attention to model disease progression over $Z_{\{temp\}}$
20: Metadata Integration & Embedding:
21: FOR each visit $t = 1\ to\ T$ DO
22: Normalize clinical features M_t
23: Embed metadata: $v_{\{t,j\}} = W_m\, m_{\{t,j\}} + b_m$ (creates Metadata Tokens V_t)
24: Concatenate with spatial features: $H_t = [X_t;\ V_t]$
25: END FOR
26: Dynamic Cross-Modal Attention (CMA) Fusion:
27: Compute Q_v, K_v, V_v from $H^{\{(v)\}}$ (Vision)
28: Compute Q_m, K_m, V_m from $H^{\{(m)\}}$ (Metadata)
29: $A_{vm} = \text{softmax}\left(\frac{Q_v K_m^T}{\sqrt{d}}\right) V_m$ (Metadata-to-Vision Attention)
30: $A_{mv} = softmax\left(\frac{Q_m K_v^T}{\sqrt{d}}\right) V_v$ (Vision-to-Metadata Attention)
31: Fused representation $F = \left[A_{\{vm\}};\ A_{\{mv\}}\right]$
32: Multi-Horizon Prediction:
33: Pass F through MLP (Multi-Layer Perceptron) to obtain $\hat{y}_{t+1}, \hat{y}_{t+2}$
34: Training:
35: Compute total loss: $\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{ord} + \lambda_2 \mathcal{L}_{temp} + \lambda_3 \mathcal{L}_{align}$
36: Optimize using AdamW with learning rate scheduling
37: Early stopping based on validation QWK (Quadratic Weighted Kappa)
38: Evaluation: Assess Accuracy, F1-score, QWK, calibration, and temporal stability

Figure 2. Algorithm of ChronoDR-Net: Longitudinal DR Progression Prediction

The evaluation is done considering the patient level, and this is to prevent the mixing of information across different visits. The accuracy, F1 score, QWK, calibration error, and temporal consistency metrics are the ones used to measure performance. External data sets are utilized to evaluate the robustness of the results in the presence of domain shift.

For interpretability support, attention maps are drawn for every time period to show which areas and visits had the most significant impact on the predictions. Temporal attribution scores also play their part in pointing out which past visits are the most critical ones in the decision-making process of the model. These instruments are indeed the ones that help to establish the connection between predictive performance and clinical trust. The complete training and inference process can be seen in Fig. 2.

IV. Results and Evaluation

The section presents an extensive assessment of the ChronoDR-Net system through five different evaluation methods which assess its performance metrics and ability to maintain time accuracy and its component testing and system comprehension capabilities.

The evaluation tests model performance through three main elements which include disease progression tracking and clinical data integration and system stability testing across different database types. The TimeDR-Net evaluation protocol establishes a base framework which researchers use to study temporal and cross-modal connection patterns at an advanced level.

A. Evaluation Metrics

The evaluation metrics were selected to reflect the progressive and ordinal nature of diabetic retinopathy (DR). The primary performance metric for this study adopted Quadratic Weighted Kappa (QWK) because diabetic retinopathy (DR) severity progresses through gradual changes instead of sudden changes. QWK shows improved performance for progression modeling because it assigns different classification errors according to their clinical importance. QWK is mathematically expressed through the following formula:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}. \qquad (8)$$

The observed agreement and expected agreement matrices are represented by the O_{ij} and E_{ij} symbols and the w_{ij} variable shows the squared distance measurement between class labels.

The researchers presented macro-averaged F1-score results to address the class imbalance that exists in diabetic retinopathy datasets because mild and moderate cases make up the majority of those cases. The research team included overall accuracy results to provide a complete assessment but they did not use it as their main measurement method.

The Horizon-2 Consistency Score tested temporal reliability by determining whether predictions made two visits ahead maintained consistency with the predictions made during the two intermediate visits. The researchers used the Ordinal Consistency Index (OCI) to confirm that their predicted severity sequences followed the required clinical progression patterns which should develop in a steady manner. The researchers assessed domain generalization through their examination of an external multi-center dataset which contained distinct camera systems and lighting conditions and different patient characteristics. The researchers used the performance gap between internal and external datasets as their method to measure system robustness.

B. Baseline Models for Comparison

The research team selected multiple baseline models which represent the main modeling approaches used in diabetic retinopathy research because they needed to conduct a thorough comparison study. A ResNet50-based static CNN trained on single fundus images served as the reference baseline, quantifying the limitations of cross-sectional diagnosis. A CNN–GRU hybrid was included to represent early attempts at progression modeling using recurrent architectures, which are known to struggle with long-range dependencies under irregular follow-up intervals.

The researchers tested the Temporal CNN–Transformer model without metadata integration to study how temporal self-attention impacted the results. The TimeDR-Net architecture with Patient-Aware Temporal Encoding (PATE) functioned as a strong baseline which uses clinical variables to determine visual features through basic connections. ChronoDR-Net uses a Vision Transformer backbone and a Temporal Vision Transformer and a Cross-Modal Attention (CMA) module which enables clinical variables to function as tokens that engage in two-way attention through multiple time periods.

TABLE I: SUMMARY OF BASELINE MODELS USED FOR COMPARISON

Baseline Model	Components	Temporal Modeling	Metadata Usage	Purpose in Comparison
Static CNN (ResNet50)	CNN encoder	None	None	Measures performance of single-image DR classification without sequence modeling.
CNN–GRU Hybrid	CNN + GRU	Recurrent (short range)	None	Assesses limitations of RNN-based progression modeling over longer timelines.
Temporal CNN–Transformer	CNN encoder + Temporal Transformer	Attention-based (visual only)	None	Isolates the contribution of temporal self-attention without metadata influence.
TimeDR-Net (PATE)	CNN encoder + GRU + PATE module	Recurrent (with conditioning)	Shallow conditioning	Provides strongest baseline with partial personalization via clinical metadata.
ChronoDR-Net (Proposed)	ViT backbone + TViT + CMA module	Full temporal self-attention	Deep cross-modal fusion	Evaluates impact of treating metadata as active tokens to guide temporal attention.

C. Quantitative Outcomes

The results of the study are presented in Table II and Figure 2 which show that ChronoDR-Net outperforms all baseline models in every evaluation metric. The internal dataset testing showed that ChronoDR-Net achieved a QWK score of 0.915 which exceeded the best baseline performance of TimeDR-Net whose QWK score reached 0.842. The model achieved 95.8 percent accuracy together with a macro F1 score of 0.902 which showed it performed evenly across different severity levels. The proposed framework maintains its core strength through temporal stability. The Horizon-2 Consistency Score of 0.86 shows a substantial advantage over competing models which suffered from severe accuracy losses during long-term forecasting. The result demonstrates that ChronoDR-Net tracks continuous progression paths instead of using transient connections. The external validation process confirmed the model's strength because ChronoDR-Net attained a QWK score of 0.892 despite facing major distribution changes. The model exhibits only slight performance decline because it acquires knowledge about actual disease-related patterns instead of studying hospital-specific image features. The error analysis process showed that the largest improvements occurred when distinguishing between moderate and severe non-proliferative DR which represents a clinically difficult boundary point. Table II shows the quantitative assessment results which compare ChronoDR-Net against baseline models and Fig. 3. displays the quantitative assessment results which compare ChronoDR-Net against baseline models.

TABLE II: QUANTITATIVE COMPARISON OF CHRONODR-NET AND BASELINES

Model	QWK ↑	Accuracy (%) ↑	F1-Score ↑	Horizon-2 Consistency ↑
Static CNN (ResNet50)	0.721	81.4	0.784	0.66
CNN–GRU Hybrid	0.793	86.2	0.812	0.71
Temporal CNN–Transformer	0.827	89.1	0.836	0.74
TimeDR-Net (PATE)	0.842	92.3	0.861	0.78
ChronoDR-Net (Proposed)	0.915	95.8	0.902	0.86

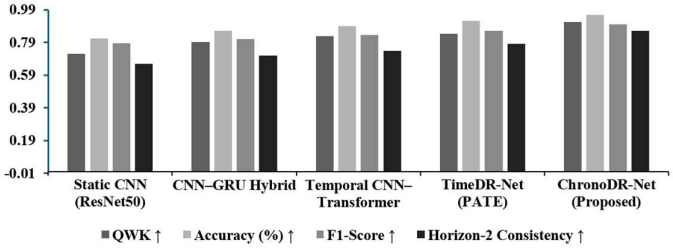

Figure 3. Quantitative Comparison of ChronoDR-Net and Baselines

D. Ablation Study

The research used ablation experiments to measure how different architectural components affect system performance according to the results shown in Table III and Fig. 4. The system showed decreased performance after the cross-modal attention module was removed which led to a QWK score of 0.861 because the system failed to use metadata as auxiliary inputs for proper longitudinal reasoning. The system showed better results through basic metadata merging which reached a QWK score of 0.874 yet deep cross-modal fusion technology

performed better than that method. The system performance decreased when the Temporal Vision Transformer got replaced with a GRU which produced a QWK score of 0.832 because recurrent architectures cannot handle irregular follow-up periods. The system showed its most severe performance drop when all temporal modeling components got removed which resulted in a QWK score of 0.801 because the system depended on static images to predict DR progression. Table III shows the ablation study results which test different components of ChronoDR-Net and Figure 3 shows the results of the ablation study which tests different components of ChronoDR-Net.

TABLE III: ABLATION STUDY OF CHRONODR-NET COMPONENTS

Model Variant	QWK ↑	Accuracy (%) ↑	Interpretation
Full Model (CMA + TViT)	0.915	95.8	Optimal Configuration
Without CMA	0.861	92.1	Metadata Not Fully Utilized
Metadata Concatenation Only	0.874	93.2	Limited Cross-Modal Interaction
TViT Replaced with GRU	0.832	90.6	Poor Long-Range Modeling
Without Temporal Encoding	0.801	87.4	Sequence-Level Reasoning Lost

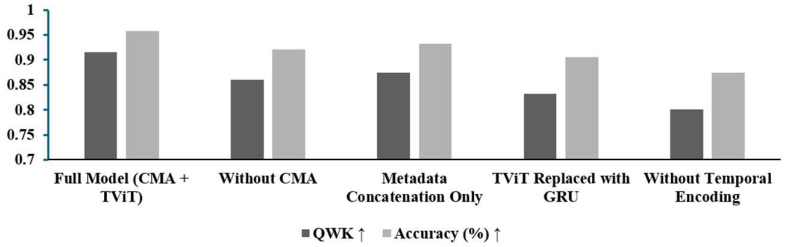

Figure 4. Ablation Study of ChronoDR-Net Components

E. Visualization and Temporal Interpretability

The analysis of interpretability showed that model attention patterns matched established clinical patterns. Saliency maps which depended on metadata showed that higher HbA1c levels created a focus on microaneurysms and hemorrhagic regions while extended disease duration showed the importance of viewing all vascular changes. The mapping of temporal attention showed that early visits made the highest predictions about patients who developed symptoms quickly while later visits created more accurate predictions about patients who developed symptoms at a slower pace. The distribution of patient attention showed different patterns which depended on their personal health progress instead of following a standard method to reach decisions. The acquired representations show clinical credibility because they create understanding of medical knowledge.

F. Discussion and Clinical Implications

The experimental results demonstrate that ChronoDR-Net effectively captures diabetic retinopathy progression as a temporally conditioned and patient-specific process, rather than as a sequence of independent grading decisions. The consistent improvements observed across internal and external evaluations (Table II), together with the ablation findings (Table III), indicate that performance gains arise not from increased model capacity alone, but from the explicit modeling of long-range temporal dependencies and structured cross-modal interactions.

The superior Horizon-2 consistency achieved by ChronoDR-Net shows its capacity to forecast multiple future visits while keeping trajectory-level plausibility intact. The first behavior shows distinct results because recurrent models together with their shallowly conditioned counterparts show declining accuracy results when their forecasting stretch extends. The clinical results show that active clinical metadata functions as tokens within the attention system which allows dynamic risk factors to change temporal feature importance thus helping better identification of disease stages which are clinically similar. The mechanism functions as an essential tool during borderline situations which occur at moderate to severe transitional points because static visual cues do not provide sufficient information.

The ability to predict disease severity which extends beyond the immediate upcoming visit has benefits for both risk-based screening and customized patient follow-up appointment planning. The medical team will focus their monitoring on patients who show high-risk progression while patients who show stability will undergo testing at longer time intervals. The interpretability analysis in Figure 1 shows that the model's attention patterns follow established clinical reasoning which connects systemic factors including glycemic control to specific times and

locations of retinal pathology examination. The model needs this transparent system because it helps clinicians trust the system which functions as a decision-support system instead of a black-box prediction tool.
ChronoDR-Net needs extra processing power because it uses transformer-based systems for processing time data and cross-modal attention but the resulting benefits in temporal stability and general dataset performance and clinical interpretability make this decision worth it for studies that involve longitudinal screening. ChronoDR-Net functions as a methodologically and clinically grounded framework which monitors diabetic retinopathy over time while connecting progression prediction with direct clinical applicability.

V. Conclusion

The research introduced ChronoDR-Net which acts as an all-encompassing framework for multimodal temporal learning that models diabetic retinopathy progression through specific patient-based tracking of their condition. The system establishes a novel architecture, which implements spatial retinal mapping, clinical visit-based temporal self-attention and structured cross-modal interaction with systemic metadata. The Vision Transformer backbone together with the Temporal Vision Transformer and metadata-based attention system enables ChronoDR-Net to track the complete disease evolution from retinal pathology to long-term progression through irregular patient monitoring intervals.
ChronoDR-Net demonstrates superior performance compared to static models, recurrent models, and temporal baselines without metadata, which include performance metrics for ordinal agreement and temporal consistency and external generalization assessment. The ablation experiments demonstrate that the model achieves performance improvements through its capacity for handling temporal elements and conducting deep cross-modal fusion not through enhanced model complexity. The interpretability results show that the learned attention patterns match established clinical knowledge, which connects systemic risk factors to specific retinal alterations that occur over time in individual patients.
The research validates ChronoDR-Net as a scientific framework that provides effective methods for monitoring diabetic retinopathy through extended observation periods. The proposed system enables multi-horizon forecasting which uses clear decision-making processes to transform disease grading from a reactive model into a proactive personalized risk assessment system for ongoing patient evaluation.

Future Enhancement

The clinical usefulness and modeling abilities of ChronoDR-Net for diabetic retinopathy prediction need more than its current features because the system requires additional components which should be implemented. First, the incorporation of richer longitudinal clinical context, including medication history, laboratory trajectories, and lifestyle-related indicators, could enable more comprehensive modeling of systemic disease evolution and its interaction with retinal pathology. The framework shows capability to analyze various causal pathways which lead to development of progress patterns through its extensions.
Second, future work may explore multimodal imaging integration, incorporating optical coherence tomography (OCT) and ultra-widefield fundus imaging alongside color fundus photographs. The development of cross-modal attention mechanisms will enable researchers to study multiple imaging types which include both microvascular and structural retinal alterations.
Third, although ChronoDR-Net implicitly handles irregular visit spacing through attention-based temporal modeling, explicit continuous-time or adaptive temporal representations may further improve robustness under highly variable follow-up schedules commonly encountered in real-world screening programs. Research into model compression, lightweight attention mechanisms, and efficient inference strategies will help decrease computational requirements which will still maintain predictive accuracy from a deployment viewpoint.
The development of ChronoDR-Net into regular medical use requires its clinical verification process which must include integration into ophthalmic decision-support systems as its next stage of progress. The system provides stable progression predictions which healthcare providers can easily understand and which focus on patient needs.

References

[1] Bali, A., & Mansotra, V. (2023). Analysis of deep learning techniques for prediction of eye diseases: A systematic review. Archives of Computational Methods in Engineering. https://doi.org/10.1007/s11831-023-09989-8.
[2] Chetoui, M., & Akhloufi, M. A. (2023). Federated learning for diabetic retinopathy detection using vision transformers. BioMedInformatics, 3(4), 948–961. https://doi.org/10.3390/biomedinformatics3040058.

[3] Choudhary, A., Gopalakrishnan, N., Joshi, A., Balakrishnan, D., Chhablani, J., Yadav, N. K., Reddy, N. G., Rani, P. K., Gandhi, P., Shetty, R., Roy, R., Bavaskar, S., Prabhu, V., & Venkatesh, R. (2024). Recommendations for diabetic macular edema management by retina specialists and large language model–based artificial intelligence platforms. International Journal of Retina and Vitreous, 10(1). https://doi.org/10.1186/s40942-024-00544-6.

[4] Dai, L., Sheng, B., Chen, T., Wu, Q., Liu, R., Cai, C., Wu, L., Yang, D., Hamzah, H., Liu, Y., Wang, X., Guan, Z., Yu, S., Li, T., Tang, Z., Ran, A., Che, H., Chen, H., Zheng, Y., & Shu, J. (2024). A deep learning system for predicting time to progression of diabetic retinopathy. Nature Medicine, 30(2), 584–594. https://doi.org/10.1038/s41591-023-02702-z.

[5] Driban, M., Yan, A., Selvam, A., Ong, J., Vupparaboina, K. K., & Chhablani, J. (2024). Artificial intelligence in chorioretinal pathology through fundoscopy: A comprehensive review. International Journal of Retina and Vitreous, 10(1). https://doi.org/10.1186/s40942-024-00554-4.

[6] Matta, S., Ben Hassine, M., Lecat, C., Borderie, L., Le Guilcher, A., Massin, P., Cochener, B., Lamard, M., & Quellec, G. (2023). Federated learning for diabetic retinopathy detection in a multi-center fundus screening network. Proceedings of the IEEE EMBS Conference. https://doi.org/10.1109/EMBC40787.2023.10340772.

[7] Nakayama, L. F., Ribeiro, L. Z., Miyawaki, F., Miyawaki, A. E., Valadares, A. J., Malerbi, F. K., Regatieri, C., Celi, L. A., & Silva, P. S. (2023). Artificial intelligence for telemedicine diabetic retinopathy screening: A review. Annals of Medicine, 55(2). https://doi.org/10.1080/07853890.2023.2258149.

[8] Nderitu, P., Nunez do Rio, J. M., Webster, L., Mann, S. S., Hopkins, D., Bergeles, C., & Jackson, T. L. (2025). Augmenting a prognostic deep learning system for referable diabetic retinopathy and maculopathy with synthetic retinal images. Communications Medicine. https://doi.org/10.1038/s43856-025-01316-5.

[9] Qiao, H., Tang, F., Zhou, H., Cai, Y., Guo, K., Wang, J., Ma, T., Ju, L., Feng, W., Ma, Z., Chen, J., Luo, Y., Wang, B., Ge, Z., & Yang, Q. (2025). Forecasting the diabetic retinopathy progression using generative adversarial networks. Communications Medicine, 5(1). https://doi.org/10.1038/s43856-025-01092-2.

[10] Rabhi, S., Blanchard, F., Diallo, A. M., Zeghlache, D., Lukas, C., Berot, A., Delemer, B., & Barraud, S. (2022). Temporal deep learning framework for retinopathy prediction in patients with type 1 diabetes. Artificial Intelligence in Medicine, 133, 102408. https://doi.org/10.1016/j.artmed.2022.102408.

[11] Romero-Aroca, P., Verges, R., Pascual-Fontanilles, J., Valls, A., Franch-Nadal, J., Mundet, X., Moreno, A., Basora, J., Garcia-Curto, E., & Baget-Bernaldiz, M. (2024). Referable diabetic retinopathy prediction algorithm applied to a population of 120,389 type 2 diabetics over 11 years follow-up. Diagnostics, 14(8), 833. https://doi.org/10.3390/diagnostics14080833.

[12] Senapati, A., Tripathy, H. K., Sharma, V., & Gandomi, A. H. (2024). Artificial intelligence for diabetic retinopathy detection: A systematic review. Informatics in Medicine Unlocked, 101445. https://doi.org/10.1016/j.imu.2024.101445.

[13] Sharma, V., Rishu, Kukreja, V., Dogra, A., & Goyal, B. (2025). Transforming retinal diagnostics: Advanced detection of diabetic retinopathy using vision transformers and capsule networks. Journal of Computer Science, 21(2), 304–321. https://doi.org/10.3844/jcssp.2025.304.321.

[14] Wang, V. Y., Lo, M.-T., Chen, T.-C., Huang, C.-H., Huang, A., & Wang, P.-C. (2024). A deep learning-based ADRPPA algorithm for the prediction of diabetic retinopathy progression. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-82884-9.

[15] Yang, Q., Wei, Y., Liu, F., & Wu, Z. (2025). An accurate and efficient diabetic retinopathy diagnosis method via depthwise separable convolution and multi-view attention mechanism. Applied Sciences, 15(17), 9298. https://doi.org/10.3390/app15179298.

[16] Yao, J., Lim, J., Yong, G., Chiat, J., Ke, Y., Tan, T. F., Tan, T.-E., Vujosevic, S., & Shu, D. (2024). Novel artificial intelligence algorithms for diabetic retinopathy and diabetic macular edema. Eye and Vision, 11(1). https://doi.org/10.1186/s40662-024-00389-y.

[17] Yu, Z., Yang, X., Sweeting, G. L., Ma, Y., Stolte, S. E., Fang, R., & Wu, Y. (2022). Identify diabetic retinopathy–related clinical concepts and their attributes using transformer-based natural language processing methods. BMC Medical Informatics and Decision Making, 22(Suppl 3). https://doi.org/10.1186/s12911-022-01996-2.

[18] Zedadra, A., Salah-Salah, M. Y., Zedadra, O., & Guerrieri, A. (2025). Multi-modal AI for multi-label retinal disease prediction using OCT and fundus images: A hybrid approach. Sensors, 25(14), 4492. https://doi.org/10.3390/s25144492.

[19] Zeghlache, R., Conze, P.-H., Mostafa, D., Li, Y., Rezaei, A., Boité, H. L., Tadayoni, R., Massin, P., Cochener, B., Brahim, I., Quellec, G., & Lamard, M. (2024). L-MAE: Longitudinal masked auto-encoder with time and severity-aware encoding for diabetic retinopathy progression prediction. arXiv preprint. https://arxiv.org/abs/2403.16272.