A FRAMEWORK FOR TEMPORAL-BASED PREDICTION OF EYE DISEASES

By

Saumya Singh Jaiswal

May, 2025

Director of Thesis: Dr. David Marvin Hart, PhD

Major Department: Computer Science

ABSTRACT

Diabetic retinopathy (DR) is a leading cause of preventable blindness, and deep learning has shown promise in automating its diagnosis. However, most models treat retinal images as static inputs, overlooking the temporal nature of disease progression. In this work, we propose a **Temporal Vision Recurrent Transformer (TVRT)**: a hybrid architecture combining a fine-tuned ViT-Tiny backbone with a bidirectional LSTM, to capture both spatial features and temporal evolution from fundus image sequences.

To address the lack of temporal data in the APTOS 2019 dataset, we introduce two synthetic sequence generation methods: (1) *stage-based augmentation* using contrast and geometric transformations to mimic progressive DR stages, and (2) *neural style transfer* to simulate intra-stage variability using higher-stage fundus images as style references.

Experimental results show that while ViT and ResNet perform well on static classification, TVRT significantly outperforms them on progression modeling, achieving an F1-score of 0.86 on synthetic sequences with 5+ timesteps. Furthermore, soft attention maps derived from the ViT encoder provide interpretable visualizations that highlight clinically relevant features like hemorrhages and exudates. Our findings suggest that temporal modeling not only enhances predictive accuracy but also improves interpretability, offering a promising direction for intelligent, progression-aware eye care systems.

A FRAMEWORK FOR TEMPORAL-BASED PREDICTION OF EYE DISEASES

A Thesis

Presented to The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Computer Science

By

Saumya Singh Jaiswal

May, 2025

Director of Thesis: Dr. David Marvin Hart, PhD

Thesis Committee Members:

Dr. Nic Herndon, PhD

Dr. Moritz Dannhauer, PhD

DEDICATION

To the millions of individuals living with diabetic retinopathy and other vision-threatening conditions - this work is dedicated to you. May it contribute, in however small a way, to the development of intelligent and compassionate eye care solutions that protect sight and restore hope.

To my parents, whose endless encouragement and unconditional support have been my greatest strength.

To my mentors and teachers who have guided me with knowledge, patience, and inspiration throughout this journey.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor, **Dr. David Hart**, for his continuous guidance, encouragement, and technical mentorship throughout the course of this research. His insightful feedback and expertise in computer vision and applied deep learning have been instrumental in shaping the direction and outcome of this work.

I would also like to sincerely thank my committee members, **Dr. Nic Herndon** and **Dr. Moritz Dannhauer**, for their valuable time, constructive suggestions, and support during my thesis journey. Their thoughtful perspectives and academic insights significantly contributed to refining the scope and rigor of this study.

Additionally, I am grateful to East Carolina University for providing access to computational resources and a stimulating academic environment that enabled me to pursue this research.

Finally, I extend heartfelt thanks to my family and friends for their unwavering support, patience, and encouragement throughout this process.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The increasing prevalence of diabetic retinopathy (DR) and other retinal diseases necessitates the development of automated, intelligent, and proactive eye care solutions. Early detection and prediction of disease progression are crucial to prevent irreversible vision loss and enable timely medical intervention. Diabetic retinopathy, a complication of diabetes, affects the retina and is one of the leading causes of blindness globally.

Fundus imaging is a standard retinal imaging technique used to capture the interior surface of the eye, including the retina, optic disc, macula, and blood vessels. Figure 1.1 shows an example of a healthy fundus image and one affected by diabetic retinopathy. The disease is typically characterized by microaneurysms, hemorrhages, and neovascularization, which are visible in such images.

With the advancement of deep learning, especially transformer-based architectures such as Vision Transformers (ViT), there has been significant progress in medical image classification. However, most existing approaches focus solely on static image classification and do not model disease evolution over time, a factor particularly important in chronic and progressive conditions like diabetic retinopathy (DR). While temporal reasoning has shown promise in domains such as video analysis and language modeling, its potential remains largely unexplored in medical imaging for disease progression modeling. There is a growing hypothesis that capturing the temporal trajectory of anatomical changes may improve predictive accuracy and support earlier diagnosis, but few studies have systematically evaluated this. Our work is one of the first to investigate the role of temporal modeling in retinal

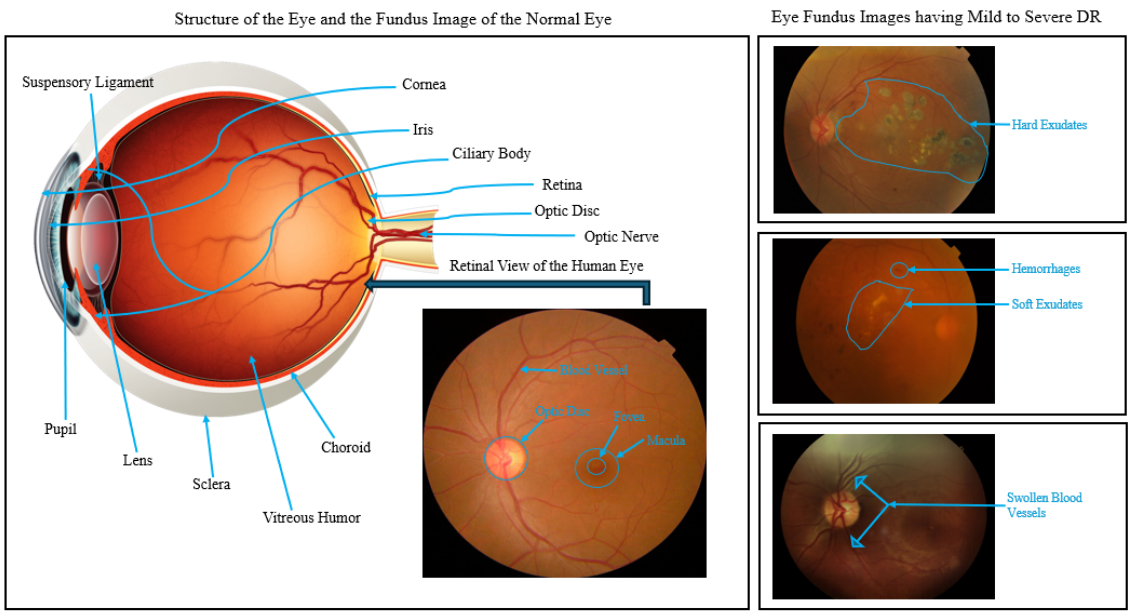

Figure 1.1: **Comparison of a Normal Fundus Image and a Fundus Image with Diabetic Retinopathy (DR).** Left: Normal fundus image. Right: Fundus image showing signs of diabetic retinopathy - DR (Hard and Soft Exudates, Swollen Blood Vessels, hemorrhages).

disease analysis, aiming to assess whether learning from synthetic temporal sequences can enable models to anticipate disease onset and severity more effectively than static baselines.

In this study, we propose a comprehensive pipeline that combines spatial encoding with temporal modeling for disease progression understanding in retinal images. Our work begins with the use of the APTOS 2019 Blindness Detection dataset, which contains 3662 high-resolution fundus images labeled across five DR stages Figure 1.2.

Although the dataset does not contain actual temporal sequences, we argue that modeling disease progression temporally is vital because DR evolves in stages over time. By synthetically creating temporal sequences and using models like Bi-LSTM on top of vision encoders, we simulate and learn from progression patterns, enabling early prediction and improving generalizability.

To address the need for temporal data, we introduce a novel synthetic data generation technique that simulates disease progression. This method generates progressive stages (e.g.,

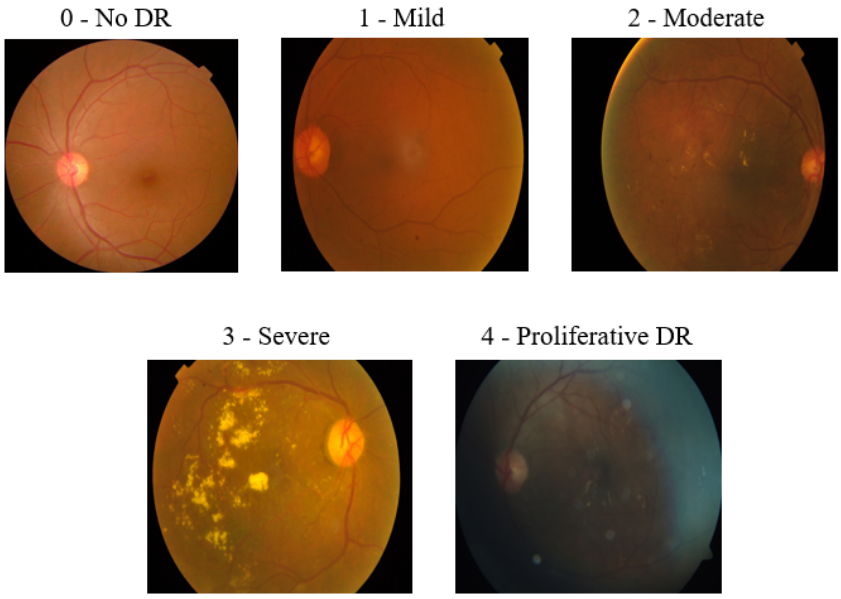

Figure 1.2: **Diabetic Retinopathy Stage Classification of Eye Fundus Images on a Scale of 0–4: A Glimpse of the APTOS Dataset**. The stages include No DR (0), Mild (1), Moderate (2), Severe (3), and Proliferative DR (4), demonstrating progressive pathological changes in retinal fundus images.

Stage 1 to Stage 2 to Stage 3) from a single fundus image by applying controlled image transformations such as contrast enhancement and geometric distortions. In addition to conventional augmentation, we employ neural style transfer as a secondary synthetic data generation technique. Style transfer enables domain adaptation by transforming fundus images with varying texture and lighting styles while preserving disease-specific anatomical structures. These stylized images are incorporated as intermediate temporal frames, further enriching the synthetic disease trajectory data. This synthetic sequence forms the temporal input required for training the Temporal Vision Recurrent Transformer (TVRT) model.

We first train a baseline Vision Transformer (ViT-Tiny) and ResNet-50 model on real APTOS images for static classification and evaluate them on the synthetic sequences to assess generalization. While the models perform reasonably well on original images, their performance degrades on progression sequences, indicating a need for temporal modeling.

To design our own model, we fine-tune the ViT model and feed the temporal sequences into TVRT, which combines a ViT backbone for spatial feature extraction with temporal

attention layers to model disease progression. This allows TVRT to effectively attend to temporal patterns and subtle morphological changes across the progression stages.

Our approach introduces three key innovations: (1) A synthetic disease progression pipeline that generates realistic temporal sequences, (2) Integration of neural style transfer to enrich temporal diversity, and (3) Deployment of a hybrid spatial-temporal model (TVRT) for end-to-end progression modeling. Through extensive experiments, we show that TVRT outperforms static models in predicting future disease stages, achieving up to 86% F1-score accuracy, especially in both short-term (3 images) and long-term (5+ images) synthetic sequences. This paves the way for intelligent systems capable of not only detecting disease but also anticipating its progression - a critical requirement for proactive healthcare solutions in ophthalmology.

Chapter 2

Related Work

This work builds on fundamental pieces of deep learning, temporal methods, style transfer, and explainable AI. The relevant pieces related to this work are presented.

2.1 Transformer Models in Vision

Recent years have witnessed the rise of transformer-based architectures in computer vision. Dosovitskiy et al. [8] introduced the Vision Transformer (ViT), which replaces convolutional layers with pure transformers applied to sequences of image patches. Their work demonstrated that, with large-scale pretraining, ViTs could rival or outperform CNNs in image recognition tasks. Following this, He et al. [12] proposed Masked Autoencoders (MAE), a self-supervised training approach where a transformer reconstructs masked regions of an image, learning robust feature representations with high pretraining efficiency. Extending this transformer paradigm to the spatiotemporal domain, Bertasius et al. [3] introduced TimeSformer, a transformer architecture for video understanding that efficiently captures spatial and temporal relationships using factorized attention mechanisms.

2.2 Transformers and Deep Learning in Retinal Imaging

Deep learning methods have seen extensive application in eye disease diagnosis, particularly in analyzing fundus and OCT images for diabetic retinopathy (DR). Wang et al. [26] proposed a self-supervised ViT approach that accurately detects four major eye diseases using

unlabeled fundus images. Several studies have leveraged CNNs [23, 6, 24] and ViTs [2, 20] to classify DR severity from fundus images. Notably, Varanasi et al. [23] explored multiple CNN-based architectures and transfer learning approaches for DR detection, while Das et al. [6] evaluated 26 different deep learning networks, identifying EfficientNetB4 as a high-performing model. Atwany et al. [2] and Nazih [20] proposed ViT-based pipelines specifically targeting DR stage classification, demonstrating competitive performance.

Studies by Tomy [22] and Praneeth [21] emphasized the importance of pre-processing and dataset selection in model accuracy, while Dasari [7] and Kumar [16] further highlighted the role of transfer learning and architectural optimization. Meanwhile, Manoj et al. [11] achieved a weighted kappa score of 0.92546 in the APTOS 2019 challenge using transfer learning with CNNs and segmentation via U-Net, underlining the utility of hybrid architectures. Yang et al. [28] extended ViT training with MAE to large retinal datasets, showing domain-specific pretraining significantly improves diabetic retinopathy classification, with their VMLRI model reaching 93.42% accuracy on the APTOS dataset.

2.3 Hybrid Architectures and Interpretability in Retinal Diagnosis

Several studies combine transformer and CNN backbones to enhance interpretability and performance. Bhuiyan et al. [4] proposed a hybrid model integrating VGG-16 and Swin Transformer for OCT-based retinal diagnosis. The model achieved 98.8% classification accuracy and used Grad-CAM visualizations to enhance explainability. Other reviews [25, 15] and comparative studies emphasize the role of explainable AI (XAI) in retinal disease screening, with visualization techniques such as Grad-CAM++ and LIME used to localize critical image features, making these models more transparent and clinically viable.

2.4 Data Scarcity and Synthetic Data Generation

To combat the scarcity of labeled data, especially in high DR severity levels, multiple studies have explored data augmentation and synthesis. Araujo et al. [1] proposed inserting semi-

synthetic neovessel structures into fundus images to balance the dataset and improve PDR detection. Zhou et al. [30] introduced DR-GAN, a conditional GAN that generates high-resolution fundus images with controllable lesion information and grading severity. Their work improves DR classification by enriching training data in underrepresented classes. Similarly, Wang et al. [27] presented DeepMT-DR, a multi-task model that integrates lesion segmentation and image super-resolution with DR grading. Cutur et al. [5] explored ViT-based transfer learning for multi-class retinal disease classification and demonstrated that ViTs outperform CNNs when pretrained on domain-specific datasets.

Photorealistic style transfer has also become a viable method for synthetic data generation that strictly preserves content. While artistic style transfer aims to manipulate content in ways similar to painting and other mediums [9, 10, 13], photorealistic style transfer preserves the content in such a way that is strictly identifiable with the original scene, such as switching day to night in a landscape photo. Many works present photorealstic style transfer approaches [18, 29, 17]. In this work, we use the proposed method by Liu et al. [19] given its publicly available code and lightweight model.

2.5 Transfer Learning in Medical Imaging

A comprehensive review by Kim et al. [14] found transfer learning—particularly using ResNet and Inception models—as a widely adopted strategy for medical image classification. The study encouraged using pretrained networks as feature extractors to reduce training time and improve accuracy on limited medical datasets.

2.6 Summary and Positioning

Collectively, these works underscore the growing relevance of transformer architectures in retinal disease classification. Yet, few studies model disease *progression* or simulate temporal sequences from static datasets. Our work addresses this gap by introducing a Temporal Vision Recurrent Transformer (TVRT) trained on synthetic sequences generated from AP-

TOS fundus images. By combining spatial encoding via ViT, temporal modeling using BiLSTM, and attention-based visualization, we provide an end-to-end framework for not only detecting but also anticipating DR progression, paving the way for next-generation intelligent ophthalmic systems.

Chapter 3

Methodology

This study proposes a novel approach to modeling diabetic retinopathy (DR) progression using a Temporal Vision Recurrent Transformer (TVRT) architecture. Our methodology integrates three major components: synthetic data generation, vision transformer fine-tuning, and temporal sequence modeling via a bidirectional LSTM (Bi-LSTM). Below we present each component in detail.

3.1 Dataset Overview

We used the publicly available **APTOS 2019 Blindness Detection** dataset, which contains high-resolution fundus images labeled with five stages of diabetic retinopathy: 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe), and 4 (Proliferative DR) like stated in Figure 1.2. This dataset was chosen due to its balanced clinical annotations, diversity in pathology, and standard use in DR classification research. It forms the foundation for training and validating our proposed models.

Although the APTOS dataset contains only single-point observations (no temporal sequences), it is well suited for synthetic sequence generation due to the clear stage-wise labeling of disease severity. This allows us to simulate the progression of DR over time by augmenting static images in a structured manner.

3.2 Synthetic Data Generation

To bridge the gap between static and temporal learning, we generate sequences of synthetic fundus images that emulate how DR manifests and worsens over time. The key motivations underpin this component:

- *Medical Rationale*: DR typically progresses gradually, manifesting as structural and vascular changes in the retina.
- *Data Scarcity*: While clinical patients will regularly visit an optometrist that obtains fundus images over time, publicly available longitudinal fundus datasets are rare and expensive to obtain, especially with labeled disease stages.

Thus, we present two possible algorithms for synthetically generating future disease stages given a single reference image.

3.2.1 Stage-Based Synthetic Augmentation

Since temporal fundus datasets are scarce, we introduced controlled contrast enhancements and geometric transformations to emulate progressive retinal deterioration. These transformations, such as increased brightness and lesion enhancement, are medically motivated - reflecting real visual cues seen in advancing diabetic retinopathy, such as growing hemorrhages, exudate buildup, and vessel distortion [11].

We simulate the temporal worsening of DR by applying progressive image transformations, including increased contrast and minor rotations, which approximate changes seen in worsening retinal pathology. A similar approach as shown to be an effective estimation various stages [1, 30].

For each image I, we generate additional images $\{I_2, I_3\}$ such that:

$$I_k = R_{\theta_k}(C_{\alpha_k}(I_1)), \quad \text{for } k = 2, 3$$

where C_α represents contrast adjustment with parameter $\alpha = 1 + 0.1 \cdot k$, and R_θ represents rotation by a small random angle $\theta_k \in [-3°, 3°]$. These are saved as JPEGs at a controlled

compression level to manage dataset size while maintaining visual fidelity. Each synthetic image figure 3.1 is annotated with an increased diagnosis label, mimicking temporal disease progression:

$$y_k = \min(y_1 + k - 1, 4), \quad k = 2, 3$$

Figure 3.1: **Illustration of Stage-Based Synthetic Augmentation**: Starting from the original fundus image (Stage 1), we apply contrast enhancement and slight rotation to simulate intermediate disease progression (Stage 2 and Stage 3). These transformations mimic the morphological changes seen in worsening diabetic retinopathy while preserving anatomical structure. The augmented images serve as temporal inputs to train TVRT on disease trajectory modeling.

The APTOS 2019 Blindness Detection dataset, used in this study, contains 3,662 high-resolution fundus images, each labeled with a diabetic retinopathy (DR) stage from 0 (No DR) to 4 (Proliferative DR). Each image entry consists of a unique id_code and a corresponding diagnosis, enabling a direct association between the image and its disease severity level.

To facilitate temporal modeling of disease progression, we introduced a stage-based synthetic augmentation pipeline. In this approach, each original image was treated as Stage 1, and two additional stages (Stage 2 and Stage 3) were generated using controlled image transformations—namely, contrast enhancement and small rotational distortions—to simulate the visual characteristics of worsening DR. This process produced a 3-stage sequence for each original image, resulting in a total of 10,986 synthetic images (3 images per case $\times$ 3,662 original images).

To support extended temporal modeling, we further expanded these sequences up to a

sequence length of 7 by applying progressive transformations to previously generated images, simulating long-term disease evolution. All synthetic images were labeled accordingly to reflect increasing severity levels.

Finally, the dataset was split into 80% training and 20% validation, preserving sequence consistency to ensure robust model evaluation. This enriched synthetic dataset provided sufficient volume and progression depth for training temporal models like our proposed Temporal Vision Recurrent Transformer (TVRT).

3.2.2 Style Transfer for Disease Progression Simulation

As an alternative approach, photorealistic style transfer can be used to generate synthetic data without distorting the original content [29, 19].Photorealstic style transfer can simulate intra-class variation and domain shift, enabling us to generate new image variants with preserved anatomical content but altered visual attributes, mimicking variability observed across patients, devices, and clinical stages.

Fundus images at higher DR stages were used as style targets to generate new versions of lower-stage images while preserving original anatomical content. Let I_s and I_c denote the style and content images, respectively. The transformed image $\hat{I}$ is optimized to minimize the total loss:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{content}}(I_c, \hat{I}) + \beta \mathcal{L}_{\text{style}}(I_s, \hat{I})$$

where α and β are balancing hyperparameters. This results in a synthetic sequence that reflects plausible temporal changes in the retina's appearance across DR stages Figure 3.2. In this work, we use the photorealistic style transfer method proposed by Liu et al. [19] An example of this style transfer is given in Figure 3.2.

Figure 3.2: **Demonstration of Style Transfer-Based Disease Simulation**: The original Stage 1 fundus image (left) is combined with a higher-stage test image as the style reference (middle) using neural style transfer. The resulting synthetic image (right) exhibits enhanced pathological textures and color tones resembling Stage 2, while preserving the anatomical structure of the original image. This technique is used to simulate plausible disease progression for training temporal models like TVRT.

To evaluate the practical utility of style transfer in disease modeling, we applied it across the entire APTOS training dataset. For each Stage 1 training image, we generated Stage 2 and Stage 3 versions using test set fundus images of higher severity as style references. This yielded a synthetic sequence of three temporal stages per training instance, effectively tripling the dataset volume. Each stylized output preserved anatomical consistency while embedding stage-specific textures, color shifts, and lesion-like structures. After generation, the resulting sequences were organized into time steps and split into 80/20 partitions for model training and validation. These sequences were later fed into the TVRT model to capture disease progression in a temporal learning framework.

While previous works have applied style transfer for synthetic data generation, none have applied it in the domain of fundus photography. This provides a unique and new approach for synthetically generating such images.

3.3 Vision Transformer Fine-Tuning

The Vision Transformer (ViT-Tiny) is chosen as our base spatial model due to its capacity for long-range spatial dependency modeling via self-attention, which is advantageous for identifying subtle patterns like microaneurysms and hemorrhages. ViT divides each image into

fixed-size non-overlapping patches, embeds them into vectors, and adds learnable position encodings. These are then processed by a standard transformer encoder architecture.

Formally, for an input image $I \in \mathbb{R}^{H \times W \times C}$, it is divided into N patches of size $P \times P$. More specifically,

$$N = \frac{HW}{P^2}$$

Each patch is linearly projected and added with positional encodings as presented in [8]:

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{\text{pos}}$$

where x_p^i is the i-th image patch, E is the embedding matrix, and E_{pos} is the positional embedding.

We experimented with various ViT parameter configurations, and we use ViT-Tiny because it offers an optimal trade-off between computational efficiency and performance. Larger ViT variants such as ViT-Base and ViT-Large exhibited marginal performance gains but required significantly more GPU memory and training time, which limited their scalability for sequence-based modeling in our pipeline. ViT-Tiny, with its reduced number of parameters, was more suitable for training on synthetic sequences, especially when combined with a temporal model like TVRT. Furthermore, ViT-Tiny showed faster convergence during fine-tuning and sufficient capacity to capture pathological retinal features, making it a practical choice for deployment and integration with temporal modules.

ViT-Tiny is pretrained on ImageNet and fine-tuned on APTOS to recognize DR patterns. Fine-tuning helps adapt the model to medical data by shifting its attention to pathological regions such as the macula, blood vessels, and optic disc.

3.4 Temporal Vision Recurrent Transformer (TVRT)

While the ViT captures spatial structure in static images, it lacks the ability to model temporal dynamics. To address this, we introduce the **Temporal Vision Recurrent Transformer (TVRT)**, a framework designed to model DR progression as a time-series sequence. TVRT accepts a series of image embeddings $\{z_1, z_2, \ldots, z_T\}$ from ViT-Tiny and feeds them into a bidirectional LSTM. An overview of the TVRT networks is shown in Figure 3.3.

The LSTM can be bidirectional because DR progression may depend on patterns observed both in the past and future image frames. In a sequence of retinal images, some progression cues (such as vessel dilation or hemorrhage expansion) are more easily understood when the network has access to the entire sequence context. A bidirectional LSTM enables the model to consider both forward and backward dependencies, thus improving prediction robustness, especially when the order of progression is synthetically generated. This is particularly helpful in cases where disease features may not appear linearly across frames.

$$h_t = \text{BiLSTM}(z_t, h_{t-1})$$

In this formulation, z_t represents the spatial feature embedding of the t^{th} input image in the sequence, as produced by the ViT-Tiny encoder. The BiLSTM module processes this input by incorporating both the past context (via a forward LSTM) and future context (via a backward LSTM). The hidden state h_t is a concatenation of the outputs from the forward and backward passes, enabling the model to capture bidirectional temporal dependencies.

This design is particularly useful in synthetic progression sequences, where intermediate images may exhibit subtle or non-linear changes that depend on earlier and later visual cues. By using BiLSTM, the model does not rely solely on a one-directional timeline, allowing it to better understand disease evolution patterns that unfold across the full image sequence.

The final hidden state h_T (corresponding to the last timestep) serves as a temporally-aware representation of the entire sequence. This embedding is passed to a linear classifica-

tion layer followed by softmax to predict the DR stage:

$$\hat{y}_T = \text{Softmax}(W h_T + b)$$

Here, W and b are learnable parameters that map the final BiLSTM hidden representation into a 5-class output space corresponding to DR stages. The use of BiLSTM allows TVRT to effectively model temporal transitions in retinal pathology, improving classification accuracy on synthetic sequences that simulate disease progression.

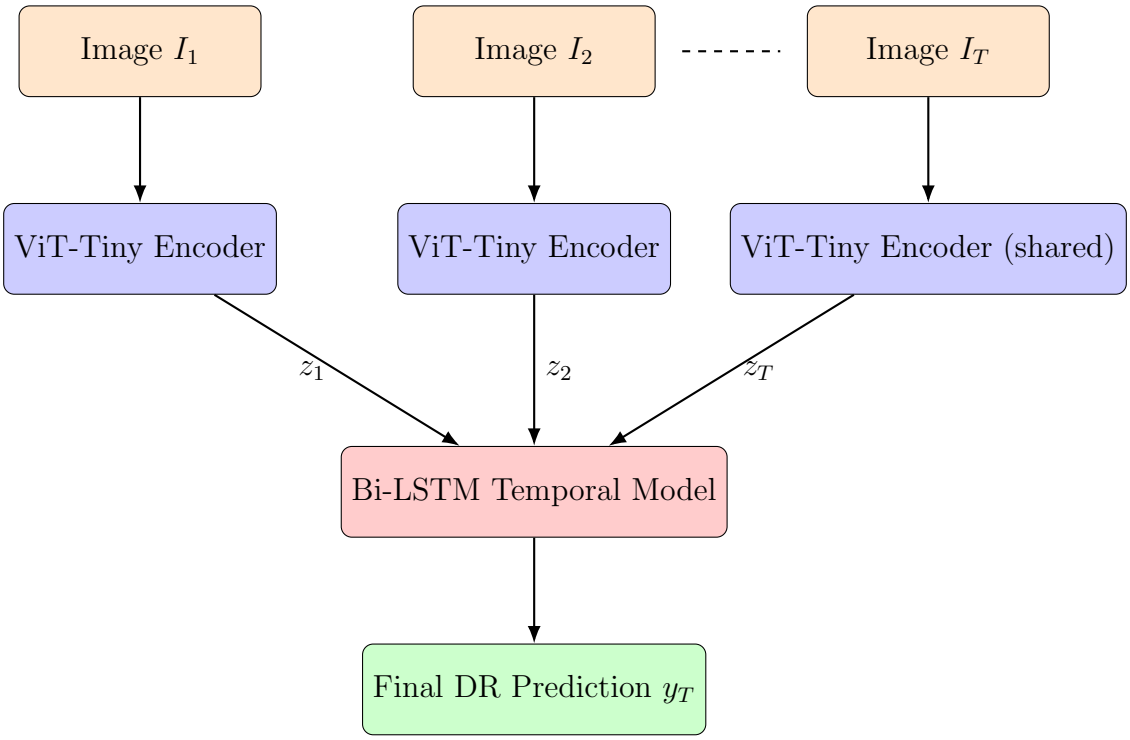

Figure 3.3: **Temporal Vision Recurrent Transformer (TVRT) Flow diagram**: Each input retinal image I_t is encoded by a shared ViT-Tiny encoder pretrained and fine-tuned on the APTOS dataset. The latent embeddings $\{z_1, z_2, \ldots, z_T\}$ are passed into a Bi-LSTM to capture temporal dependencies and progression patterns, leading to a final diabetic retinopathy stage prediction y_T.

3.4.1 Time-Step Handling

Unlike ViT which processes images independently without temporal awareness, TVRT requires explicit sequence modeling. We manually assign synthetic images into time-steps

based on their simulated progression levels:

$$I = \{I_1, I_2, \ldots, I_T\}, \quad T \in \{3, 5, 7\}$$

Each image I_t in a sequence corresponds to a plausible progression state of the same patient's retina. These sequences are used to train TVRT to predict the DR stage of the last image, reflecting either short-term (3 steps) or long-term (5–7 steps) disease development.

Since real temporal durations between disease stages are not available in the APTOS dataset, our synthetic sequences assume equal temporal spacing between stages. That is, we do not assign variable or random time intervals between I_1, I_2, and I_3, but instead treat them as uniformly progressing stages of diabetic retinopathy (e.g., each transition approximates an equal clinical interval). This deterministic setup simplifies training and ensures consistency across sequences while still preserving the progression trend necessary for temporal modeling.

3.5 Soft Attention Maps

We integrate soft attention visualization from the ViT's self-attention weights to enhance interpretability, which is critical in clinical applications. Given attention weights $A \in \mathbb{R}^{N \times N}$ at a given transformer layer, we derive a spatial attention map over input patches, and upscale it to image resolution. This allows clinicians to localize disease-relevant regions. These visualizations highlight which regions (e.g., hemorrhages, exudates) contributed to the model's final decision. This not only enhances trust but also validates model reasoning through clinical inspection.

The soft attention maps in our model are generated directly from the self-attention layers within the Vision Transformer (ViT) encoder. Specifically, at each transformer block, the model computes attention weights that indicate how much each patch attends to every other patch in the input. These weights form an attention matrix $A \in \mathbb{R}^{N \times N}$, where N is the number of input patches (including the [CLS] token).

For visualization, we extract the attention scores from the last layer of the ViT encoder, focusing on the attention weights from the [CLS] token to all other image patches. The [CLS] token is used as the global representation for classification, and its attention distribution provides insight into which image regions most influenced the final decision.

We select the attention vector associated with the [CLS] token, discard its self-attention component, and reshape the remaining values into a 2D grid (e.g., 14×14 for a 224×224 image with 16×16 patches). This 2D attention map is then interpolated to the original image resolution and overlaid as a heatmap.

Unlike methods such as Grad-CAM [4], which rely on backpropagated gradients, this soft attention mechanism is intrinsic to the ViT architecture and requires no additional computation or modification. It offers a more direct and interpretable way to understand model focus, making it particularly suitable for sensitive domains like medical diagnostics where transparency is critical.

3.6 Sequence Length Encoding

To bridge ViT (non-temporal) and TVRT (temporal) models, we added a learnable timestep encoding $E_{temp}(t)$ to each image's embedding:

$$\tilde{z}_t = z_t + E_{temp}(t)$$

This formulation is conceptually analogous to the positional encoding used in the original Vision Transformer, where spatial patch embeddings are augmented with positional information to preserve the 2D structure of the image. In our case, however, the position refers to the **temporal order** of the image within the sequence, not its spatial layout.

The addition of $E_{temp}(t)$ enables the model to differentiate between image embeddings based not just on spatial content, but also on their relative place in the disease timeline (e.g., Stage 1 vs. Stage 3). This is crucial because identical or similar spatial features could

represent different diagnostic implications depending on when they appear during disease progression.

Unlike fixed sinusoidal positional encodings used in early transformer models, we employ a *learnable* encoding so that the network can optimize how much importance to assign to the relative position in the sequence based on training data. This flexibility allows the model to discover temporal dynamics most predictive of disease evolution, thereby improving the interpretability and accuracy of the progression modeling.

By incorporating $E_{temp}(t)$, we effectively embed temporal awareness into each frame-level representation before it is passed to the BiLSTM. This ensures that the temporal model can learn not only how the visual features evolve, but also how their position in the sequence impacts the final diagnosis.

Chapter 4

Results and Discussion

In this chapter, we present results showcasing multiple aspects of the work. First, we present comparisons between variations on CNN and Vision Transformer models to find the best ViT setup. Second, we present the accuracy of the ViT models on the augmented and synthetically generated data, showcasing the realistic nature of our synthetic approach. Third, we present the performance of the TVRT model on sequential data, both for short and long term sequences. Finally, we present visualizations of the attention mapping and discuss the possible implications of the results for clinical settings.

4.1 Comparisons between Base Networks

To determine the best starting model for the spatial component of the network, we compare two standard architectures, ResNet50 and ViT. Table 4.1 summarize the F1-scores obtained across different models trained on the original APTOS dataset and its synthetic counterpart. These results are also visualized in the line chart in Figure 4.1, which highlights the training dynamics across epochs. On the original dataset, the ViT model achieved the highest F1-score of 0.85 using a stepwise learning rate strategy, outperforming ResNet-50 which reached 0.81 in 20 epochs. The accuracy trend shows consistent improvements at every third epoch as the learning rate was decreased progressively.

Table 4.1: **Model Performance on Original APTOS Dataset**. The models were trained using the Adam optimizer with both fixed and stepwise learning rate schedules. Fixed learning rates used were 1×10^{-4}, 1×10^{-5}, and 1×10^{-6}. In the stepwise learning rate strategy, the learning rate was adjusted every third epoch in the sequence $1 \times 10^{-4} \rightarrow 1 \times 10^{-5} \rightarrow 1 \times 10^{-6}$. F1 scores were used to evaluate model performance over 20 training epochs.

Model	**Learning Rate Schedule**	**Final F1 Score**
ViT (Baseline)	Fixed 1×10^{-4}	0.78
ViT (Baseline)	Fixed 1×10^{-5}	0.79
ViT (Baseline)	Fixed 1×10^{-6}	0.77
ViT (Baseline)	Step-wise (every 3 epochs)	**0.85**
ResNet50	Fixed 1×10^{-4}	0.74
ResNet50	Fixed 1×10^{-5}	0.76
ResNet50	Fixed 1×10^{-6}	0.75
ResNet50	Step-wise (every 3 epochs)	**0.81**

4.2 Spatial Network Performance on the Synthetic Data

To validate the fidelity and diagnostic utility of our synthetic dataset, we evaluate the performance of the pretrained Vision Transformer (ViT) models on the generated synthetic fundus images. Specifically, we assess both the original ViT (trained on real APTOS data) and a ViT-Tiny model fine-tuned on the synthetic data. This twofold evaluation strategy serves distinct purposes: First, by testing the real-data-trained ViT on synthetic images without any additional fine-tuning, we verify whether the synthetic augmentations—such as contrast adjustments, rotations, and style transfers—preserve the critical pathological features necessary for accurate classification. Second, the fine-tuned ViT-Tiny allows us to evaluate how well the synthetic data supports training a robust classifier from scratch. Successful classification performance in both cases indicates that our synthetic data not only mimics visual realism but also captures clinically relevant patterns. This provides a strong foundation for

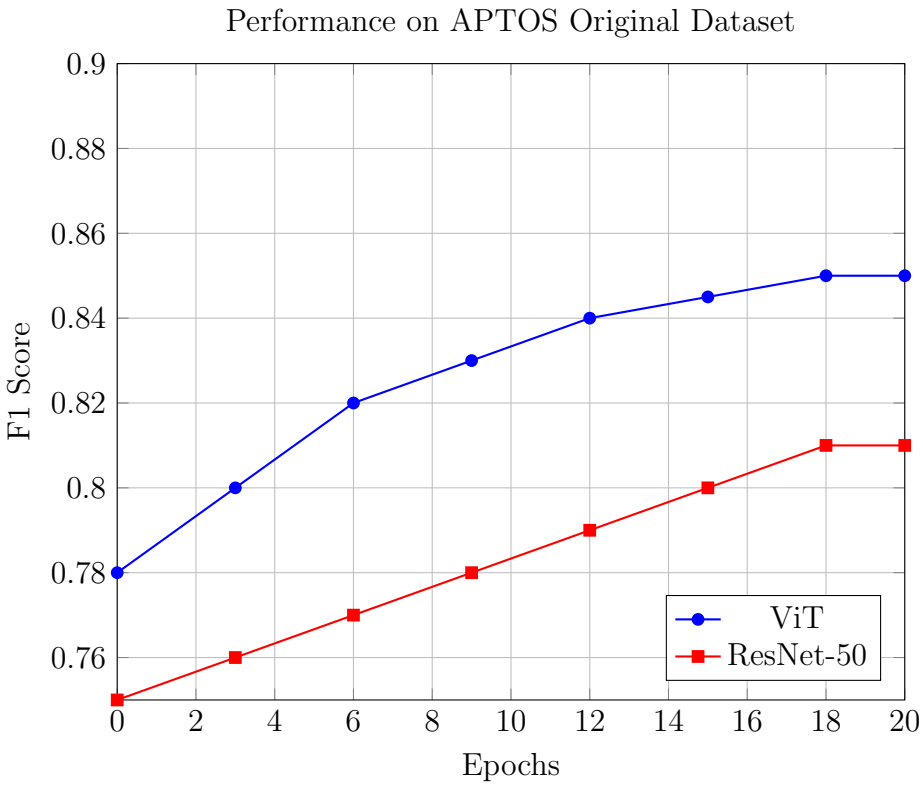

Figure 4.1: Model performance on original APTOS data with stepwise LR

using such data in temporal modeling tasks, such as in our proposed TVRT architecture.

Table 4.2 summarizes the F1-scores obtained across different models trained on the original APTOS dataset and its synthetic counterpart. On the original dataset, the ViT model achieved the highest F1-score of 0.85 using a stepwise learning rate strategy, outperforming ResNet-50 which reached 0.81 in 20 epochs. The accuracy trend shows consistent improvements at every third epoch as the learning rate was decreased progressively.

Table 4.2: **Model performance on synthetic data generated via stage-based augmentation and style transfer**. The table compares F1 scores of ViT and ViT-Tiny models on contrast-enhanced sequences and style-transferred images. The ViT-Tiny model demonstrates superior performance, particularly with contrast-augmented inputs.

Stage-Based Synthetic Augmentation		
Model	**Input Type**	**Final F1 Score**
ViT (Baseline)	Contrast-augmented Images	0.79
ViT-Tiny	Contrast-augmented Images	**0.86**
ViT-Tiny	Style-Transferred Images	0.83

4.3 Spatial vs Temporal Networks

Diseases like diabetic retinopathy progress over time, making it essential to evaluate whether a model can capture this progression. Spatial models like ViT operate on single images and lack temporal context, which limits their ability to understand disease evolution. In contrast, TVRT incorporates temporal modeling, allowing it to reason across sequential image inputs. By comparing these two approaches on synthetic progression data, we aim to quantify the added value of temporal awareness in improving diagnostic accuracy. This comparison is critical for justifying the use of more complex temporal architectures and for validating the design choices in TVRT.

Table 4.3 summarizes the F1-scores obtained across temporal data. The line charts in Figure 4.2 highlight the training dynamics across epochs. For the synthetic dataset, TVRT achieved the best performance with an F1-score of 0.86, followed by ViT-Tiny at 0.83, and ViT at 0.79. The style transfer variant of ViT-Tiny reached an F1-score of 0.80, indicating the utility of augmenting data using generative techniques.

We compared the performance of a spatial-only Vision Transformer (ViT) with the TVRT model using the same synthetic dataset, with the goal of evaluating the contribution of temporal modeling.

TVRT outperforms ViT by a margin of 7% in F1-score. This performance gain validates the importance of modeling temporal dependencies in disease progression, which static spatial models fail to capture. The temporal fusion layer in TVRT enables the model to utilize history effectively for final-stage predictions.

Overall, stepwise learning rate scheduling significantly improved model generalization. Synthetic data, particularly when combined with transformer-based architectures, proved beneficial and, in some cases, outperformed models trained on real data. This demonstrates the feasibility of leveraging synthetic augmentation to enhance retinal disease classification models.

4.3.1 Ablation Study: Temporal Modeling Length

Diabetic retinopathy progresses at different rates in individuals - some show rapid changes, while others develop symptoms slowly over time. This variability motivates the use of temporal models like TVRT, which are designed to capture both short- and long-term dependencies in disease progression.

To evaluate the effectiveness of temporal modeling, we analyzed the performance of the Temporal Vision Recurrent Transformer (TVRT) in predicting disease severity based on the synthetic temporal sequences shown in Table 4.3. The sequences simulate disease progression using 3 to 6 generated images per subject.

- *Stage-Based Synthetic Augmentation:* When evaluated in stage-based synthetic sequences, ViT-Baseline and ViT-Tiny achieved F1-scores of **0.79** and **0.83** respectively when provided with only the final image in the sequence, highlighting the limited temporal context. In contrast, TVRT achieved an F1-score of **0.84** with short-term sequences (3 time steps), and **0.86** with long-term sequences (5+ time steps), demonstrating that temporal modeling leads to more accurate predictions by leveraging progression-aware cues.
- *Style Transfer Synthetic Augmentation:* On style-transferred sequences, ViT-Tiny achieved an F1-score of **0.83** when predicting from a single frame. When using temporal sequences, TVRT reached an F1-score of **0.80** (3 time steps) and improved further to **0.82** (5+ time steps). These results indicate that even when stylistic domain shifts are introduced, temporal consistency enhances performance.

These findings emphasize that incorporating temporal sequences - whether derived from contrast-based progression or style-based transformations - enables better modeling of disease evolution and leads to more robust predictions at later stages of diabetic retinopathy.

Table 4.3: **Performance comparison of ViT, ViT-Tiny, and our proposed TVRT model on two types of synthetic augmentations: stage-based progression and style-transferred images**. Models are evaluated on either the last image in the sequence (spatial-only) or the full sequence (temporal modeling with TVRT). TVRT achieves the highest F1 score (0.86) when using full-length sequences (5+ time steps) from stage-based augmentation over 20 epochs

Stage-Based Synthetic Augmentation		
Model	**Input Type**	**Final F1 Score**
ViT (Baseline)	Last Image Only	0.79
ViT-Tiny	Last Image Only	0.83
TVRT (Ours)	Sequences (Time Steps = 3)	0.84
TVRT (Ours)	Sequences (Time Steps = 5+)	**0.86**
Style Transfer Synthetic Augmentation		
Model	**Input Type**	**Final F1 Score**
ViT-Tiny	Last Image Only	0.83
TVRT (Ours)	Sequences (Time Steps = 3)	0.80
TVRT (Ours)	Sequences (Time Steps = 5+)	0.82

4.4 Visualizing Soft Attention Map

The Temporal Vision Recurrent Transformer (TVRT) incorporates a soft attention mechanism via a modified Vision Transformer (ViT) backbone to enhance model interpretability. Specifically, the attention maps are derived from the self-attention layers within the ViT encoder. By applying a softmax operation to the attention scores, the model computes a probability distribution over the image patches, allowing it to focus more heavily on diagnostically relevant areas.

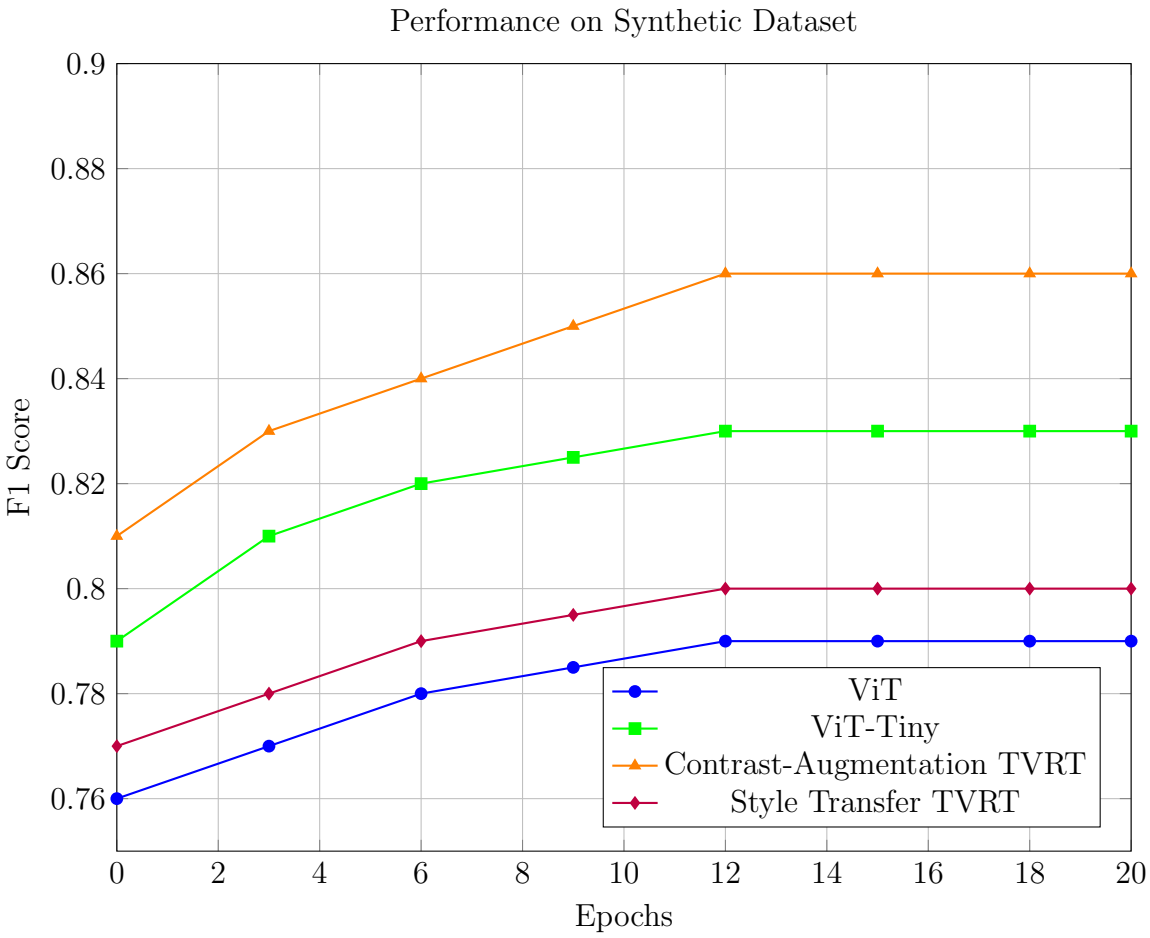

Figure 4.2: Model performance on synthetic and style transfer data

During inference, attention weights associated with the `[CLS]` token - representing global image understanding—are extracted from the final transformer layer. These weights are reshaped into a 2D grid and upsampled to match the image resolution. The resulting attention heatmap is overlaid on the original image to visualize which regions most influenced the model's decision.

Figure 4.3 presents an example of such a visualization: the left image shows a synthetic stage image of the DR fundus, and the right image highlights the model's attention, focusing on areas of pathology such as exudates and hemorrhages.

Notably, these attention maps display **color-coded intensity patterns**: regions shown in red correspond to areas with the highest attention, typically aligned with prominent patho-

logical features like exudates, hemorrhages, etc. Yellow, and green zones indicate regions with moderate attention, while **blue fringes at the periphery** reflect areas with minimal model focus. These blue fringes imply that the model has effectively learned to de-emphasize regions with low diagnostic value, reinforcing the clinical relevance of the highlighted zones. Interestingly, in several examples—such as the first image shown in Figure 4.3—the attention heatmap appears to follow the vascular structure of the retina, particularly the arcades of retinal veins. This anatomical alignment suggests that the model is learning to prioritize areas where disease markers such as hemorrhages or neovascularization typically emerge, offering clinically meaningful interpretability for ophthalmologists.

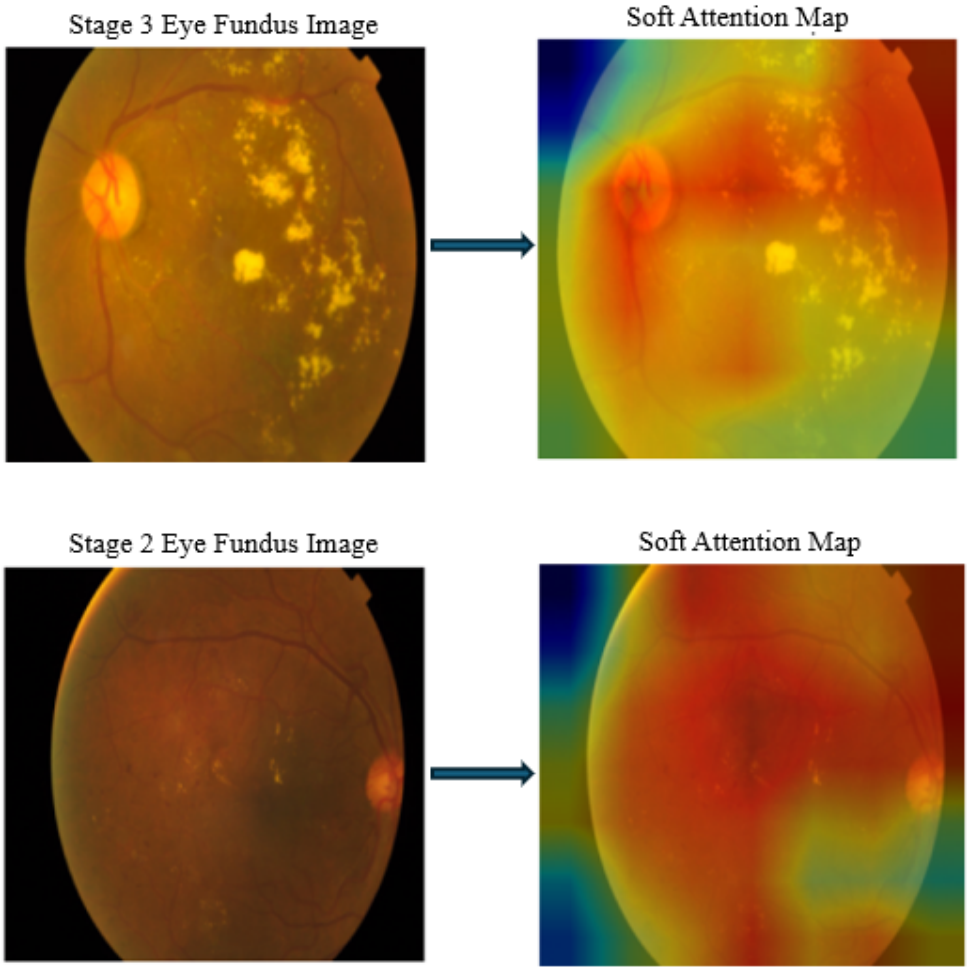

Figure 4.3: **Soft Attention Map Visualization:** The left image displays a synthetic Stages of diabetic retinopathy fundus image, and the right image shows the corresponding soft attention heatmap generated by TVRT, highlighting diagnostically significant regions.

4.5 Experimental Details

All experiments were conducted using PyTorch on a workstation equipped with an NVIDIA RTX 3090 GPU (24GB VRAM) and 128GB RAM. Training was performed over 20 epochs with a batch size of 32. The Adam optimizer was used across all experiments. A stepwise learning rate schedule was adopted: an initial learning rate of 1×10^{-4} was applied for epochs 1–3, followed by 1×10^{-5} for epochs 4–6, and 1×10^{-6} for epochs 7–20. This schedule allowed progressive fine-tuning and helped mitigate overfitting.

4.6 Discussion

The primary goal of this work was not only to achieve state-of-the-art performance on the APTOS dataset but to validate the feasibility of temporal modeling for progressive ocular diseases using transformer-based architectures. Our use of diabetic retinopathy (DR) and the APTOS data set served as a case study, a testbed to evaluate how effectively our proposed *Temporal Vision Recurrent Transformer (TVRT)* can learn progression-based representations from image sequences. The promising results demonstrate that this framework is robust to both stage-based and style-transferred synthetic sequences and highlight its generalizability for other chronic eye conditions that exhibit temporal evolution.

The superior performance of TVRT over spatial-only baselines suggests that progression modeling is indeed valuable and can capture visual patterns that evolve subtly but meaningfully over time. This has important implications: while DR was used here for experimentation, the architecture is designed to be condition-agnostic. We envision applying TVRT to other ophthalmic conditions such as age-related macular degeneration (AMD), glaucoma, and hypertensive retinopathy, where understanding the progression of the disease is critical for timely intervention.

In future iterations, we aim to expand our collaboration with clinical experts, specifically those actively studying the effects of hypertension on retinal health. Since hypertensive

retinopathy shares similar fundus-level manifestations (e.g., vessel narrowing, hemorrhages, and cotton wool spots), applying our pipeline to this domain could aid in early detection and personalized monitoring by identifying subtle vascular changes over time through temporal modeling.

An essential aspect of our approach is its interpretability. By integrating soft attention directly into the ViT backbone, our framework produces *native attention maps* that are spatially aligned with the disease-relevant regions of the input image. These soft attention maps have several advantages over post hoc explainability methods such as Grad-CAM [4], which rely on gradients and often suffer from spatial coarseness and lack of consistency across layers. In contrast, attention maps in TVRT:

- Are naturally produced as part of the forward pass of the model, enabling real-time interpretability.
- Offer high spatial resolution tied to specific patch embeddings.
- Provide consistency across samples, improving trustworthiness in clinical settings.

This makes the model not only accurate but also transparent, an essential feature for the adoption of AI in healthcare. As we continue to refine the pipeline, we also plan to explore incorporating expert annotations and conducting usability studies with ophthalmologists to assess the clinical alignment of these attention maps.

Ultimately, this study demonstrates that modeling temporal changes in retinal images is both feasible and beneficial. By bridging synthetic sequence generation, temporal learning, and built-in interpretability, our TVRT model lays a foundational framework for broader applications in proactive eye care.

Chapter 5

Conclusion and Future Work

This thesis presents a novel and unified approach for understanding and modeling the progression of diabetic retinopathy using both spatial and temporal learning techniques. Recognizing the limitations of traditional spatial-only models, we proposed the Temporal Vision Recurrent Transformer (TVRT), which combines the representational power of Vision Transformers (ViT) with the sequential modeling capacity of Bi-LSTMs. This allows the system to process temporal sequences of fundus images and make more informed predictions about disease severity.

To overcome the lack of real temporal data in publicly available datasets like APTOS 2019, we developed two synthetic data generation strategies. The first, Stage-Based Synthetic Augmentation, creates pseudo-temporal sequences by progressively applying transformations such as contrast enhancement and rotation. The second, Style Transfer-based Simulation, augments the diversity and realism of disease progression by applying visual style adaptations between fundus images while retaining anatomical content. Together, these synthetic pipelines simulate disease evolution over time, allowing temporal models like TVRT to be trained effectively.

Fine-tuning of ViT-Tiny on both original and synthetic data yielded strong spatial representations, which were then passed through the TVRT for temporal classification. Attention map visualizations further validated the interpretability of our model by highlighting diagnostically significant retinal regions.

Overall, this research demonstrates the feasibility and effectiveness of integrating syn-

thetic temporal sequences with transformer-based architectures. It establishes a path forward for models that do not merely classify diseases statically but also anticipate their progression, thus supporting early intervention strategies in real-world healthcare.

Although our synthetic approach proved highly effective, it is limited by its handcrafted and visually-driven nature. Future work could involve the use of generative models, such as GANs or diffusion models, trained to produce temporally consistent disease trajectories that better reflect clinical patterns. In addition, the inclusion of real longitudinal retinal datasets would help validate the biological accuracy of progression modeling.

Another promising direction involves multi-modal fusion, where fundus images are supplemented with structured patient data, such as age, blood sugar levels, or medical history, to create richer temporal contexts. This would improve robustness and personalize predictions.

To ensure clinical applicability, lightweight versions of TVRT should be explored for deployment on mobile or low-resource devices.

Finally, the generalizability of this framework should be tested across other chronic, progressive diseases beyond ophthalmology. By combining synthetic data generation, temporal modeling, and explainable attention mechanisms, this work sets a foundation for real-time, progression-aware AI systems in preventive healthcare.

BIBLIOGRAPHY

[1] Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A. M., and Campilho, A. Data augmentation for improving proliferative diabetic retinopathy detection in eye fundus images. *IEEE Access 8* (2020), 181587–181602.

[2] Atwany, M. Z., Sahyoun, A., and Yaqub, M. Deep learning techniques for diabetic retinopathy classification: A survey. *IEEE Access PP* (2022), 1–1.

[3] Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? *ArXiv abs/2102.05095* (2021).

[4] Bhuiyan, M. H., Haldar, S., Chowdhury, M. S., Bushra, N., and Jilan, T. Z. *An interpretable diagnosis of retinal diseases using vision transformer and Grad-CAM.* Thesis, Brac University, 2024.

[5] Cutur, E. S., and Inan, N. G. Multi-class classification of retinal eye diseases from ophthalmoscopy images using transfer learning-based vision transformers. *Journal of Imaging Informatics in Medicine* (2025).

[6] Das, D., Biswas, S. K., and Bandyopadhyay, S. Detection of diabetic retinopathy using convolutional neural networks for feature extraction and classification (drfec). *Multimedia tools and applications 81* (2022), 30823–30847.

[7] Dasari, B., and Rajput, D. Optical imaging for diabetic retinopathy diagnosis and detection using ensemble models. *PeerJ Computer Science 10* (2024), e1771.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2414–2423.

[10] Gatys, L. A., Ecker, A. S., Bethge, M., Hertzmann, A., and Shechtman, E. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 3985–3993.

[11] H, M. S., and Bosale, A. A. Detection and classification of diabetic retinopathy using deep learning algorithms for segmentation to facilitate referral recommendation for test and treatment prediction. *arXiv preprint arXiv:2401.02759* (2024).

[12] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 16000–16009.

[13] Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (2016), Springer, pp. 694–711.

[14] Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., and Ganslandt, T. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging 22*, 1 (2022), 69.

[15] Kinger, S., and Kulkarni, V. A review of explainable ai in medical imaging: implications and applications. *International Journal of Computers and Applications 46*, 11 (2024), 983–997.

[16] Kumar, S. A., Kumar, J. S., Neraniki, P., Yadav, K. A., and Basha, S. S. A review of deep learning techniques on fundus images for detecting diabetic retinopathy on public datasets. *Journal of Innovative Image Processing 1* (2022), 3–14.

[17] Li, M., Wang, G., Zhang, X., Liao, Q., and Xiao, C. D-lut: Photorealistic style transfer via diffusion process. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2025), pp. 9206–9214.

[18] Li, X., Liu, S., Kautz, J., and Yang, M.-H. Learning linear transformations for fast arbitrary style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition* (2019).

[19] Liu, R., Zhao, E., Liu, Z., Feng, A. W.-W., and Easley, S. J. Instant photorealistic style transfer: A lightweight and adaptive approach, 2023.

[20] Nazih, W., Aseeri, A. O., Atallah, O. Y., and El-Sappagh, S. Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. *IEEE Access 11* (2023), 2487–2498.

[21] Praneeth, D., and Kumar, N. S. Diabetic retinopathy detection and categorizing using a lightweight deep learning approach. *Journal of Engineering and Applied Sciences 9* (2024), 1–6.

[22] Tomy, M. C., and Malathy, C. Exploring deep learning methods for diabetic retinopathy detection and classification. In *5th International Conference for Emerging... 2024* (2024), IEEE.

[23] Varanasi, K., and Dasari, C. M. A novel deep learning framework for diabetic retinopathy detection. In *Conference Information and Communication...2022* (2022), IEEE.

[24] VinothiniRaghul, S., and Sivasakthy, S. Diabetic retinopathy classification using deep learning techniques. In *Second International Conference on Emerging... 2024* (2024), IEEE.

[25] Vipparthi, V., Rao, D. D. R., Mullu, S., and Patlolla, V. Diabetic retinopathy classification using deep learning techniques. In *International Conference Electronic Systems...2022* (2022), IEEE.

[26] Wang, J., Kang, M., Liu, Y., Zhang, C., Liu, Y., Li, S., Qi, Y., Xu, W., Tang, C., Occhipinti, E., Yusufu, M., Wang, N., Bai, W., Gao, S., and Occhipinti, L. G. Ssvt: Self-supervised vision transformer for eye disease diagnosis based on fundus images. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* (2024), 1–4.

[27] Wang, X., Xu, M., Zhang, J., Jiang, L., and Li, L. Deep multi-task learning for diabetic retinopathy grading in fundus images. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)* (2021), Association for the Advancement of Artificial Intelligence, pp. 2826–2833.

[28] Yang, Y., Cai, Z., Qiu, S., and Xu, P. Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. *PLoS ONE 19*, 3 (2024), e0299265.

[29] Yoo, J., Uh, Y., Chun, S., Kang, B., and Ha, J.-W. Photorealistic style transfer via wavelet transforms. In *International Conference on Computer Vision (ICCV)* (2019).

[30] Zhou, Y., Wang, B., He, X., Cui, S., and Shao, L. Dr-gan: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. In *IEEE Journal of Biomedical and Health Informatics* (2020), vol. 25, IEEE, pp. 1503–1514.