SWEG-Net : Deep Learning Model for Glaucoma Classification

A project report submitted in fulfilment
of the requirements for the degree of
BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

By

Kamal Anand (21JE0310)

Parth Pandya (21JE0635)

Yashwant Khare (21JE1073)

Under the guidance of **Prof. Soumen Bag**

Department of Computer Science and Engineering

Indian Institute of Technology (ISM), Dhanbad May 2025

Certificate

This document serves as confirmation that **Mr. Kamal Anand, Mr. Parth Pandya** and **Mr. Yashwant Khare** has duly submitted the project report titled “**SWEG-Net : Deep Learning Model for Glaucoma Classification**” to the Indian Institute of Technology (ISM), Dhanbad, fulfilling the requirements for the Bachelor of Technology degree in Computer Science and Engineering. The guidance and supervision for this research were provided by **Prof. Soumen Bag**. I affirm that the contents of this report, either in whole or in part, have not been presented to any other institute or university for the purpose of obtaining any degree or diploma. I extend my best wishes to **Mr. Kamal Anand, Mr. Parth Pandya** and **Mr. Yashwant Khare** for their future endeavours.

Prof. Soumen Bag,
Associate Professor,
Department of Computer Science and Engineering,
IIT(ISM), Dhanbad.

SWEG-Net: A Novel Hybrid Deep Learning Architecture for Multi-stage Glaucoma Classification

September 20, 2025

Abstract

Glaucoma, a progressive optic neuropathy, is a leading cause of irreversible blindness worldwide. Early detection is critical but challenging due to its subtle onset and the expertise required for diagnosis. This paper introduces SWEG-Net (SWin-EfficientNet-enhanced-GLAM Network), a novel hybrid deep learning architecture designed specifically for multi-stage glaucoma classification from fundus images. Our architecture synergistically combines the strengths of EfficientNet-B0 as a CNN backbone for feature extraction, Swin Transformer for capturing global contextual information, and a Global-Local Attention Module (GLAM) for enhancing feature refinement. These components are integrated through a self-adaptive gating mechanism and optimized with a learnable loss function that dynamically balances focal and cross-entropy losses. We conducted comprehensive experiments on two standard datasets: Harvard Dataverse V1 (HDV1) and a Large Merged Glaucoma (LMG) dataset. SWEG-Net achieves state-of-the-art performance with 86.73% accuracy and 93.58% AUC score on HDV1, and 86.12% accuracy with 94.99% AUC score on LMG for three-class classification (normal, early glaucoma, advanced glaucoma). Extensive ablation studies confirm the contribution of each component to the overall performance. Our model's ability to accurately identify early-stage glaucoma has significant clinical potential for screening applications, particularly in resource-limited settings.

deep learning, glaucoma classification, computer-aided diagnosis, medical image analysis, convolutional neural networks, vision transformers, attention mechanisms, fundus imaging

1 Introduction

Glaucoma is a chronic optic neuropathy characterized by progressive degeneration of retinal ganglion cells and their axons, leading to structural changes in the optic nerve head and subsequent visual field loss [1]. Often referred to as the "silent thief of sight," glaucoma typically produces symptoms only after significant vision damage has already occurred. According to the World Health Organization, glaucoma affects approximately 76 million people globally, with projections suggesting this number will reach 111.8 million by 2040 [2].

The disease burden is particularly significant in developing countries, where limited access to healthcare infrastructure and specialist services results in high rates of undiagnosed cases. In India, for instance, it is estimated that over 90% of glaucoma cases remain undetected in rural communities [3]. This alarming statistic underscores the critical need for accessible and reliable screening methods that can be deployed in resource-constrained settings.

Early detection and timely intervention are crucial for preventing vision loss from glaucoma, as damage to the optic nerve is irreversible but disease progression can be significantly slowed with appropriate treatment. Traditional diagnosis of glaucoma relies on a comprehensive eye examination involving intraocular pressure measurement, visual field testing, and assessment of the optic disc through fundus examination-procedures that require specialized equipment and trained ophthalmologists.

Fundus photography offers a non-invasive, cost-effective imaging technique that enables visualization of the optic disc and surrounding retinal structures, providing valuable information for glaucoma assessment. The increasing availability of portable fundus cameras has created an opportunity for wider screening coverage, particularly in underserved areas. However, the interpretation of fundus images requires considerable expertise, and there is notable inter-observer variability even among specialists.

Automated image analysis systems powered by artificial intelligence (AI) have emerged as promising tools to address these challenges. In recent years, deep learning approaches have demonstrated remarkable success across various medical imaging tasks, including the detection of diabetic retinopathy, age-related macular degeneration, and glaucoma from fundus images. These systems have the potential to enhance screening efficiency, reduce the burden on healthcare providers, and improve access to care in regions with limited resources.

Despite significant advances, several challenges persist in the development of robust AI systems for glaucoma detection from fundus images:

- **Subtle Manifestations:** Early glaucomatous changes can be subtle and difficult to distinguish from normal variations, making accurate classification challenging.

- **Diverse Presentations:** Glaucoma manifests with considerable phenotypic diversity across different populations and disease subtypes.

- **Image Quality Variations:** Fundus images captured in real-world settings often exhibit variations in quality due to differences in imaging devices, patient factors, and technician expertise.

- **Limited Labeled Data:** The availability of large, diverse, and expertly labeled datasets for model training and validation remains limited.

- **Interpretability Concerns:** The "black box" nature of many deep learning models raises concerns about trustworthiness and adoption in clinical settings.

In this paper, we introduce SWEG-Net (SWin-EfficientNet-enhanced-GLAM Network), a novel deep learning architecture designed to address these challenges in glaucoma classification from fundus images. Our approach integrates three powerful paradigms in computer vision: Convolutional Neural Networks

(CNNs) for efficient hierarchical feature extraction, Vision Transformers for capturing long-range dependencies and global context, and attention mechanisms for emphasizing relevant features while suppressing irrelevant ones.

The main contributions of this paper are as follows:

1. We propose a hybrid architecture that synergistically combines EfficientNet for efficient feature extraction, Swin Transformer for global context modeling, and Global-Local Attention Module (GLAM) for feature refinement.

2. We introduce a self-adaptive gating mechanism that dynamically weighs the contributions from transformer and attention pathways based on input characteristics.

3. We develop a learnable loss function that adaptively balances focal loss and cross-entropy loss, optimizing the training process for multi-class classification of glaucoma stages.

4. We conduct extensive experiments on standard glaucoma datasets and demonstrate that SWEG-Net outperforms existing state-of-the-art methods across multiple performance metrics.

5. We present detailed ablation studies to quantify the contribution of each component in our architecture and provide insights into their individual and combined effects.

The remainder of this paper is organized as follows: Section II reviews related work in glaucoma detection and classification using deep learning. Section III describes the datasets and preprocessing techniques used in this study. Section IV presents the proposed SWEG-Net architecture in detail. Section V outlines the experimental setup and implementation details. Section VI reports the results and comparative analysis. Section VII provides discussion and clinical implications, and Section VIII concludes the paper with limitations and future directions.

2 Related Work

2.1 Traditional Approaches for Glaucoma Detection

Early approaches to automated glaucoma detection from fundus images relied on traditional machine learning techniques with hand-crafted features. These features typically focused on quantifiable aspects of the optic disc and cup, such as the vertical cup-to-disc ratio (CDR), rim-to-disc ratio, and ISNT rule compliance (which refers to the normal neuroretinal rim configuration where the thickness follows the pattern: Inferior > Superior > Nasal > Temporal).

Bock et al. [4] proposed a supervised learning approach using support vector machines (SVMs) to classify glaucomatous and healthy eyes based on image features that approximated the manual assessment performed by ophthalmologists. Their method achieved an area under the ROC curve (AUC) of 0.88 on a dataset of 575 images. Similarly, Acharya et al. [5] employed a combination of higher-order spectra and texture features with an SVM classifier, achieving an accuracy of 91.7%.

Other traditional approaches included extracting morphological features of the optic disc and cup through active contour models and ellipse fitting. Joshi et al. [6] developed a method for optic disc and cup segmentation using a combination of graph cuts and active contour models, reporting a CDR error of 0.09 compared to manual measurements. While these methods provided a foundation for automated analysis, they required domain expertise for feature design and often struggled with the variability present in clinical images.

2.2 CNN-Based Approaches

The advent of deep learning, particularly convolutional neural networks (CNNs), marked a paradigm shift in medical image analysis by enabling automatic feature learning directly from data. Early applications of CNNs for glaucoma detection demonstrated promising results compared to traditional approaches.

Chen et al. [7] employed a CNN to detect glaucomatous changes in fundus images, achieving an AUC of 0.831 on a dataset of 650 images. Li et al. [8] developed an inception-inspired CNN architecture for glaucoma classification that achieved an AUC of 0.986 on a private dataset of 48,116 fundus images.

Transfer learning approaches using pre-trained CNNs have been particularly popular due to the limited availability of large, annotated medical imaging datasets. Raghavendra et al. [9] fine-tuned a pre-trained 18-layer CNN for glaucoma diagnosis, reporting an accuracy of 98.13% on a private dataset. Similarly, Christopher et al. [10] leveraged a ResNet-based architecture pre-trained on ImageNet and fine-tuned for detecting glaucomatous optic neuropathy, achieving a sensitivity of 92.82% and specificity of 92.37% when validated on a large dataset of 14,822 fundus images.

More sophisticated architectures have been proposed to address the specific challenges of glaucoma detection. Fu et al. [11] introduced Disc-aware Ensemble Network (DENet) that incorporates both local optic disc region features and global fundus image features for improved classification, achieving an AUC of 0.9756 on the ORIGA dataset. Liu et al. [12] proposed a semi-supervised approach using adversarial training to leverage both labeled and unlabeled data, reporting an AUC of 0.9384 on the REFUGE challenge dataset.

2.3 Attention-Enhanced Models

Attention mechanisms have been increasingly integrated into CNN architectures to improve their performance by selectively focusing on relevant regions or features. These mechanisms are particularly valuable for medical image analysis, where diagnostic features may occupy only a small portion of the overall image.

For glaucoma detection specifically, Li et al. [13] proposed AG-CNN, an attention-guided CNN that incorporates a spatial attention module to emphasize discriminative regions in fundus images. Their model achieved an accuracy of 95.3% on a local dataset of 2,554 images. Jiang et al. [14] developed a joint optic disc and cup segmentation network with spatial context awareness, achieving mean Dice coefficients of 0.9450 for disc and 0.8826 for cup segmentation on the REFUGE dataset.

Cheng et al. [15] introduced CA-Net, a cascaded attention network for multi-stage glaucoma classification. Their model incorporates both global and channel attention modules to enhance feature representation at different levels. CA-Net achieved an accuracy of 82.75% on the Harvard Dataverse v1 (HDV1) dataset, serving as a key benchmark for our work.

Beyond traditional attention mechanisms, several approaches have explored more sophisticated attention variants. Rao et al. [16] proposed MTNet, which combines spatial attention with multi-task learning to simultaneously segment the optic disc/cup and classify glaucoma, reporting an AUC of 0.9576 on the REFUGE dataset. Gomez et al. [17] leveraged guided attention to improve interpretability in their CNN model, achieving an AUC of 0.94 on a private dataset of 14,880 images.

2.4 Transformer-Based Approaches

Vision Transformers (ViTs) have recently emerged as powerful alternatives to CNNs for image classification, inspired by the success of transformer architectures in natural language processing. Transformers excel at capturing long-range dependencies through their self-attention mechanism, which can be particularly valuable for analyzing complex medical images.

In the general computer vision domain, Dosovitskiy et al. [18] introduced the Vision Transformer (ViT), which divides an image into fixed-size patches, linearly embeds them, and processes them with a standard transformer encoder. Despite its simple design, ViT achieved competitive results on image classification benchmarks when trained on large datasets. Building on this work, Touvron et al. [19] proposed Data-efficient Image Transformers (DeiT), which demonstrated that transformers could be effectively trained on smaller datasets through distillation techniques.

Liu et al. [20] introduced the Swin Transformer, which computes self-attention within local windows rather than globally, significantly reducing computational complexity while enabling cross-window connections through shifted window configurations. This hierarchical architecture has shown strong performance across various vision tasks, including object detection and semantic segmentation.

In the medical imaging domain, transformer-based models have been increasingly explored. Punn and Agarwal [21] applied various transformer architectures, including ViT and Swin, to COVID-19 detection from chest X-rays, demonstrating competitive performance compared to CNN-based approaches. For retinal image analysis specifically, Xie et al. [22] proposed a transformer-based framework for retinal vessel segmentation that achieved state-of-the-art results on multiple public datasets.

For glaucoma detection, Khan et al. [23] employed a hybrid CNN-Transformer model that combines local feature extraction with global context modeling, achieving an accuracy of 94.5% on a private dataset. Similarly, Wang et al. [24] developed a transformer-based approach for optic disc/cup segmentation and glaucoma screening, reporting an AUC of 0.954 on the REFUGE challenge dataset.

Despite these advances, the application of transformers to glaucoma classification remains relatively underexplored compared to CNN-based approaches. Our work aims to bridge this gap by leveraging the complementary strengths of CNNs and transformers through a novel hybrid architecture.

2.5 Loss Function Innovations

Beyond architectural innovations, researchers have also explored various loss functions to address challenges specific to medical image classification, such as class imbalance and the need for high sensitivity.

Focal loss, introduced by Lin et al. [25], was designed to address class imbalance by down-weighting easy examples and focusing training on hard examples. This loss function has been particularly effective for medical imaging tasks where abnormal findings may be rare but critical. Yang et al. [26] applied focal loss for pulmonary nodule detection in CT images, demonstrating improved performance over standard cross-entropy loss.

For multi-class medical image classification, Wang et al. [27] proposed a hybrid loss function that combines focal loss with Dice loss, achieving improved segmentation performance on brain tumor MRI images. Similarly, Yeung et al. [28] developed a unified loss function that adaptively balances different loss components for brain lesion segmentation.

In the context of glaucoma detection, Zhang et al. [29] employed a weighted binary cross-entropy loss to address class imbalance in their attention-based CNN model, achieving an AUC of 0.9651 on the REFUGE dataset. Ma et al. [30] introduced CDGNet with a modified loss function that incorporates clinical domain knowledge, reporting an accuracy of a modified loss function that incorporates clinical domain knowledge, reporting an accuracy of 95.7% on a private dataset.

Our approach extends these efforts by introducing a learnable loss function that adaptively balances focal loss and cross-entropy loss, with the weighting parameters optimized during training through backpropagation. This novel approach allows the model to dynamically adjust its optimization objective based on the dataset characteristics and training dynamics.

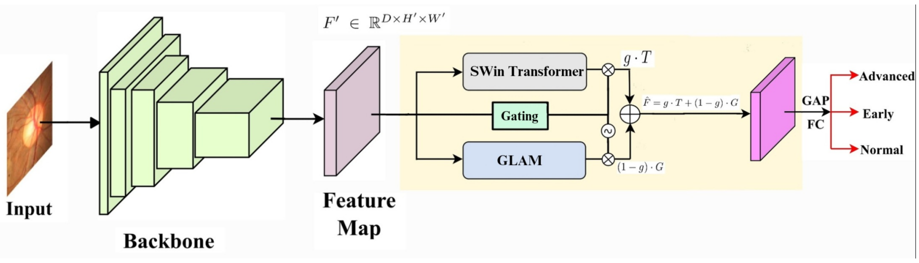


Figure 1: Overview of the proposed SWEG-Net architecture. The model consists of an EfficientNet-B0 backbone for feature extraction, followed by parallel Swin Transformer and GLAM branches. Features from both branches are dynamically fused using a self-adaptive gating mechanism before final classification.

3 Dataset and Methodology

3.1 Dataset Description

In this study, we utilized two publicly available datasets for glaucoma classification:

1. **Harvard Dataverse version 1 (HDV1)**: This dataset contains high-quality fundus images collected from various eye hospitals and clinics. Each image is labeled by experienced ophthalmologists as normal, early glaucoma, or advanced glaucoma based on clinical assessment including optic disc appearance, visual field tests, and intraocular pressure measurements.

2. **RIM-ONE**: This is an open retinal image database for optic nerve evaluation. It includes stereoscopic optic disc images with annotations for glaucoma severity provided by expert ophthalmologists.

For our main experiments, we used both these datasets individually and also created a combined dataset called **Large Merged Glaucoma (LMG)** by merging HDV1 and RIM-ONE. This combined dataset provides a more diverse set of images, enabling better generalization of our model. The LMG dataset contains fundus images categorized into three classes:

1. Normal (healthy eyes)

2. Early Glaucoma (initial stage of the disease)

3. Advanced Glaucoma (severe condition)

Table 1 summarizes the distribution of images across the three classes in each dataset.

Table 1: Dataset statistics showing the distribution of images across classes

Dataset	**Normal**	**Early Glaucoma**	**Advanced Glaucoma**	**Total**
HDV1	788	289	467	1544
RIM-ONE	14	12	14	40
LMG (Combined)	802	301	481	1584

3.2 Data Preprocessing

To prepare the images for deep learning model training, we applied several preprocessing steps to standardize the input and enhance model performance:

1. **Image Resizing**: All fundus images were resized to a uniform dimension of 224 × 224 pixels to ensure compatibility with pre-trained models and maintain consistent input sizes during training.

2. **Color Normalization**: We applied standard normalization using the mean and standard deviation values from the ImageNet dataset:

$$I_{\text{normalized}} = \frac{I - \mu}{\sigma} \tag{1}$$

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ for the RGB channels respectively.

3. **Data Augmentation**: To increase the diversity of our training set and improve model generalization, we implemented the following augmentation techniques:

    - Random horizontal and vertical flipping with probability 0.5
    - Random rotation up to 10 degrees
    - Random adjustments to brightness (factor range: 0.8-1.2) and contrast (factor range: 0.8-1.2)
    - Random cropping and resizing

The augmentation pipeline can be mathematically formulated as a composition of transformations:

$$T(I) = T_n \circ T_{n-1} \circ ... \circ T_1(I) \tag{2}$$

where I is the input image, and each T_i represents an individual transformation applied with probability p_i.

3.3 Dataset Splitting

We employed a stratified splitting strategy to ensure representative distribution of classes across training, validation, and testing sets:

- 60% for training
- 20% for validation
- 20% for testing

This approach ensures that our model is evaluated on data unseen during training while preserving the original class distribution. The stratified split can be mathematically represented as:

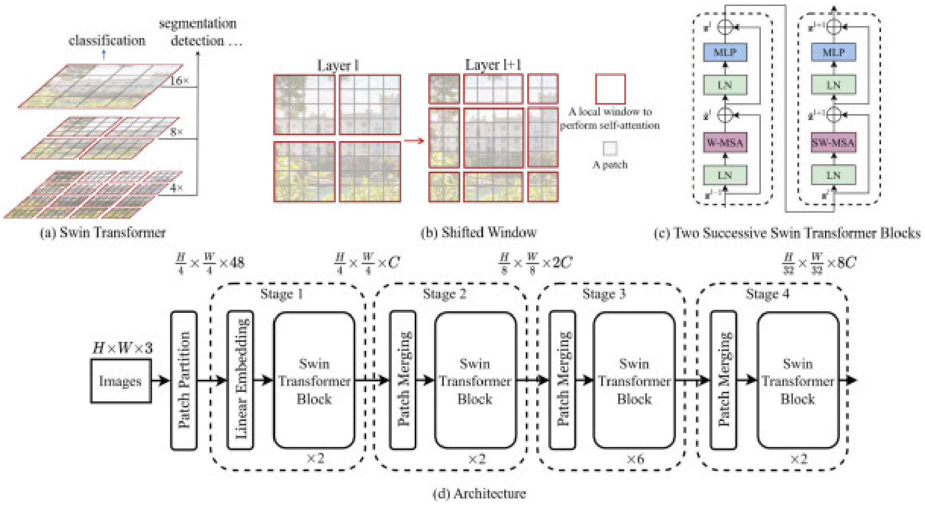
$$D_{\text{train}}, D_{\text{val}}, D_{\text{test}} = \text{split}(D, [0.6, 0.2, 0.2]) \tag{3}$$

where D represents the entire dataset.

4 Proposed SWEG-Net Architecture

4.1 Overview

SWEG-Net is a novel hybrid architecture designed specifically for multi-stage glaucoma classification from fundus images. The model consists of four main components:



(a) Swin Transformer

(b) Shifted Window

(c) Two Successive Swin Transformer Blocks

(d) Architecture

Figure 2: Architecture of the Swin Transformer module used in our model. The Swin Transformer first splits the input image into non-overlapping patches and applies linear embedding. It then processes features through multiple stages, each containing Swin Transformer blocks with window-based self-attention (W-MSA) and shifted window-based self-attention (SW-MSA). The shifted windowing scheme enables efficient computation by limiting self-attention to local windows while allowing cross-window connections through window shifting. This hierarchical structure progressively reduces spatial resolution through patch merging layers between stages, capturing multi-scale features crucial for glaucoma classification.

1. A CNN backbone (EfficientNet-B0) for efficient feature extraction

2. A Swin Transformer branch for capturing global context and long-range dependencies

3. A Global-Local Attention Module (GLAM) for feature refinement through local and global attention mechanisms

4. A self-adaptive gating mechanism for dynamic fusion of the transformer and attention pathways

The final classification is performed using a fully connected layer with a learnable loss function that combines focal loss and cross-entropy loss. Fig. 1 provides an overview of the complete SWEG-Net architecture.

4.2 Feature Extraction with EfficientNet-B0

We selected EfficientNet-B0 as our CNN backbone after extensive experimentation with various architectures (as detailed in the Results section). EfficientNet models are designed to optimally balance network depth, width, and resolution through a compound scaling technique, achieving superior performance with fewer parameters compared to other CNN architectures.

The feature extraction process can be formulated as:

$$F = \Phi_{\text{CNN}}(X) \tag{4}$$

where $X \in \mathbb{R}^{3\times H\times W}$ is the input image, Φ_{CNN} represents the EfficientNet-B0 backbone, and $F \in \mathbb{R}^{C\times H'\times W'}$ is the extracted feature map, with $C = 1280$ channels, $H' = H/32$, and $W' = W/32$.

To unify the channel dimensions for subsequent processing, we apply a 1×1 convolution:

$$F' = \text{Conv}_{1\times 1}(F) \tag{5}$$

where $F' \in \mathbb{R}^{D\times H'\times W'}$ and $D = 512$ is the embedding dimension.

The specific architecture of EfficientNet-B0 consists of several Mobile Inverted Bottleneck Convolution (MBConv) blocks with squeeze-and-excitation optimization. The MBConv block can be described as:

$$y = \begin{cases} x + F(x), & \text{if } C_{\text{in}} = C_{\text{out}} \\ F(x), & \text{otherwise} \end{cases} \tag{6}$$

where $F(x)$ represents the sequence of operations including depthwise separable convolutions, squeeze-and-excitation attention, and non-linear activations.

4.3 Swin Transformer Branch

The Swin Transformer branch is designed to capture global contextual information through a hierarchical transformer architecture with shifted windows. Unlike standard vision transformers that apply self-attention globally, Swin Transformer computes self-attention within local windows while allowing for cross-window connections through shifted window partitioning.

The process begins with layer normalization of the feature map:

$$\hat{F} = \text{LN}(F') \tag{7}$$

Next, the feature map is divided into non-overlapping windows of size $M\times M$:

$$\{W_i\}_{i=1}^{N_W} = \text{WindowPartition}(\hat{F}, M) \tag{8}$$

where $N_W = \frac{H'\times W'}{M^2}$ is the number of windows.

The window partition function can be defined as:

$$\text{WindowPartition}(x, M) = \{x[i : i + M, j : j + M, :] \mid i, j \in \mathcal{I}\} \tag{9}$$

where $\mathcal{I}$ is the set of top-left corner indices for each window.

Within each window, multi-head self-attention (MSA) is applied:

$$\hat{W}_i = \text{MSA}(W_i) + W_i \tag{10}$$

The multi-head self-attention operation is defined as:

$$\text{MSA}(X) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{11}$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) \tag{12}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{13}$$

For cross-window connections, the windows are shifted before applying self-attention again:

$$\{SW_j\}_{j=1}^{N_W} = \text{ShiftedWindowPartition}(\hat{F}, M) \tag{14}$$

$$S\hat{W}_j = \text{MSA}(SW_j) + SW_j \tag{15}$$

The final output of the Swin Transformer branch is obtained by merging the windows back:

$$T = \text{WindowMerge}(\{S\hat{W}_j\}_{j=1}^{N_W}) \tag{16}$$

where $T \in \mathbb{R}^{D \times H' \times W'}$ represents the transformed feature map.

4.4 Global-Local Attention Module (GLAM)

The GLAM pathway enhances features by focusing on both global and local patterns relevant to glaucoma diagnosis. It consists of four sub-modules:

1. **Local Channel Attention (LCA)**: Emphasizes important channel-wise features locally
2. **Local Spatial Attention (LSA)**: Highlights spatially relevant regions through dilated convolutions
3. **Global Channel Attention (GCA)**: Captures channel relationships across the entire feature map
4. **Global Spatial Attention (GSA)**: Identifies globally important spatial locations

The mathematical formulation for each component is as follows:
Local Channel Attention (LCA):

$$\text{LCA} = \sigma(W_2 \cdot \sigma(W_1 \cdot F')) \tag{17}$$

$$F_{\text{LCA}} = \text{LCA} \odot F' \tag{18}$$

where W_1 and W_2 are 1×1 convolution weights, σ is the sigmoid activation, and $\odot$ denotes element-wise multiplication.

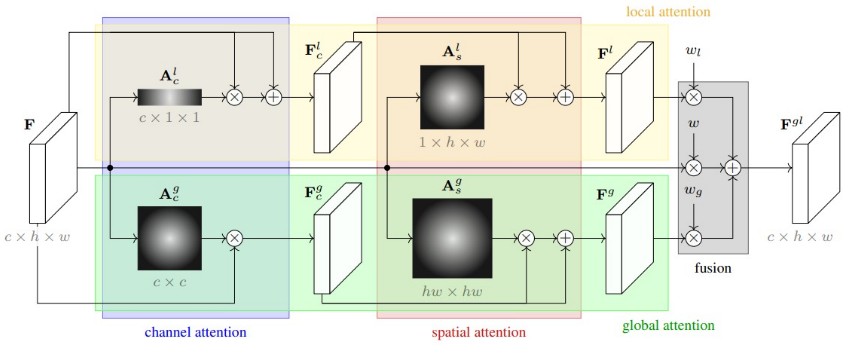


Figure 3: Architecture of the Global-Local Attention Module (GLAM). The module consists of Local Channel Attention (LCA), Local Spatial Attention (LSA), Global Channel Attention (GCA), and Global Spatial Attention (GSA) components that refine features through complementary attention mechanisms.

Local Spatial Attention (LSA):

$$\mathrm{LSA}_3 = \mathrm{Conv}_{3\times3,d=3}(F') \tag{19}$$

$$\mathrm{LSA}_5 = \mathrm{Conv}_{3\times3,d=5}(F') \tag{20}$$

$$\mathrm{LSA} = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{Concat}[F', \mathrm{LSA}_3, \mathrm{LSA}_5])) \tag{21}$$

$$F_{\mathrm{LSA}} = \mathrm{LSA} \odot F_{\mathrm{LCA}} + F_{\mathrm{LCA}} \tag{22}$$

where d represents the dilation rate.

Global Channel Attention (GCA):

$$\mathrm{GCA} = \sigma(W_4 \cdot \mathrm{ReLU}(W_3 \cdot \mathrm{GAP}(F'))) \tag{23}$$

$$F_{\mathrm{GCA}} = \mathrm{GCA} \odot F' \tag{24}$$

where GAP is global average pooling, and W_3 and W_4 are fully connected layer weights.

Global Spatial Attention (GSA):

$$\mathrm{GSA} = \mathrm{Softmax}(\mathrm{Conv}_{1\times1}(F')) \tag{25}$$

$$F_{\mathrm{GSA}} = \mathrm{GSA} \odot F_{\mathrm{GCA}} + F_{\mathrm{GCA}} \tag{26}$$

The final output of GLAM is a weighted combination:

$$G = \alpha \cdot F_{\mathrm{LSA}} + \beta \cdot F_{\mathrm{GSA}} + F' \tag{27}$$

where α and β are learnable parameters that balance the contribution of local and global attention.

Fig. 3 illustrates the architecture of the GLAM module.

4.5 Self-Adaptive Gating Mechanism

To dynamically fuse the outputs from the Swin Transformer branch (T) and GLAM branch (G), we employ a self-adaptive gating mechanism. The gate value is computed as:

$$g = \sigma(W_g \cdot \mathrm{GAP}(F') + b_g) \tag{28}$$

where W_g and b_g are learnable parameters, GAP refers to global average pooling, and σ is the sigmoid activation function.

The fusion process is formulated as:

$$\hat{F} = g \cdot T + (1 - g) \cdot G \tag{29}$$

This gating mechanism allows the model to adaptively balance the contributions of global contextual features from the transformer branch and enhanced local features from the GLAM branch based on the input image characteristics.

4.6 Classification and Learnable Loss Function

The fused feature map is globally pooled and passed through a fully connected layer for classification:

$$z = W_{fc} \cdot \mathrm{GAP}(\hat{F}) + b_{fc} \tag{30}$$

where $z \in \mathbb{R}^C$ represents the logits for C classes (in our case, $C = 3$ for normal, early glaucoma, and advanced glaucoma).

For model training, we propose a learnable weighted loss function that combines focal loss and cross-entropy loss:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{CE} + \lambda_2 \cdot \mathcal{L}_{FL} \tag{31}$$

where λ_1 and λ_2 are learnable parameters that determine the contribution of each loss component.

The cross-entropy loss is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}) \tag{32}$$

where $y_{i,c}$ is the ground truth label, and $p_{i,c}$ is the predicted probability for class c of sample i.

The focal loss is defined as:

$$\mathcal{L}_{FL} = -\frac{1}{N} \sum_{i=1}^{N} \alpha_c (1 - p_{i,c})^{\gamma} \log(p_{i,c}) \tag{33}$$

where α_c is a weighting factor for class c, and γ is a focusing parameter that reduces the relative loss for well-classified examples.

Instead of manually tuning λ_1 and λ_2, we initialize them as learnable parameters and optimize them during training using backpropagation. To ensure that these weights remain positive and sum to 1, we apply a softmax normalization:

$$[\lambda_1, \lambda_2] = \mathrm{Softmax}([w_1, w_2]) \tag{34}$$

where w_1 and w_2 are the raw learnable parameters.

This adaptive approach enables the model to adjust the importance of each loss component based on the training dynamics, leading to improved convergence and generalization. It is particularly beneficial for imbalanced datasets, as it can automatically emphasize the loss term that better addresses the current training challenges.

5 Implementation Details

5.1 Development Environment

All experiments were conducted using PyTorch framework (version 1.9.0) on NVIDIA GTX 1080Ti GPUs with 11GB memory. Our implementation leveraged the following key libraries:

- PyTorch 1.9.0
- torchvision 0.10.0
- scikit-learn 0.24.2
- OpenCV 4.5.3
- NumPy 1.21.2
- Matplotlib 3.4.3 for visualization
- tensorboard 2.6.0 for training monitoring

5.2 Model Configuration

We implemented SWEG-Net with the following specifications:

- EfficientNet-B0 pre-trained on ImageNet as the CNN backbone
- Embedding dimension: 512
- Window size for Swin Transformer: 7
- Number of attention heads: 8
- GLAM reduction ratio: 8
- Dropout rate: 0.5

5.3 Training Protocol

Our training procedure followed these parameters:

- Batch size: 32
- Initial learning rate: 0.0001
- Optimizer: Adam with weight decay 1e-5
- Learning rate scheduler: ReduceLROnPlateau (patience=10, factor=0.1)

- Early stopping patience: 20
- Maximum epochs: 100

For the learnable loss function, we initialized the raw weights w_1 and w_2 to 0, resulting in equal initial contributions from cross-entropy and focal loss after softmax normalization. The focal loss parameters were set to $\gamma = 2$ and $\alpha = 0.25$.

6 Experiments and Results

6.1 Evaluation Metrics

We evaluated our model's performance using the following metrics:

- **Accuracy**: Ratio of correctly classified samples to the total number of samples

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{35}$$

- **Precision**: Ratio of true positives to predicted positives

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{36}$$

- **Recall**: Ratio of true positives to actual positives

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{37}$$

- **F1 Score**: Harmonic mean of precision and recall

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{38}$$

- **AUC (Area Under ROC Curve)**: Measures the model's ability to discriminate between classes across different thresholds

For multi-class classification, we computed these metrics using one-vs-rest approach and reported both class-wise and macro-averaged results.

6.2 Comparative Analysis of CNN Backbones

We first conducted experiments to identify the most suitable CNN backbone for feature extraction. Seven different CNN architectures were evaluated on both HDV1 and LMG datasets: DenseNet121, DenseNet169, ResNet18, ResNet50, MobileNet, Inception, and EfficientNetB0. Table 2 presents the results of this comparative analysis.

The results demonstrate that EfficientNetB0 outperformed other architectures, achieving 83.50% accuracy and 93.50% AUC on the HDV1 dataset. ResNet50 was the second-best performer with 82.20% accuracy and 93.63% AUC. The superior performance of EfficientNet can be attributed to its compound scaling approach, which optimally balances network depth, width, and resolution.

Table 2: Performance comparison of different CNN backbones on HDV1 dataset

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
DensNet169	80.26	81.39	80.26	80.73	93.01
DensNet121	79.61	78.57	79.61	78.83	91.71
ResNet18	81.23	80.54	81.23	80.80	93.14
ResNet50	82.20	82.05	82.20	81.96	93.63
MobileNet	79.29	80.40	79.29	79.76	92.45
Inception	82.20	81.22	82.20	81.28	92.01
EfficientNetB0	**83.50**	**85.20**	**83.50**	**84.05**	**93.50**

To further evaluate the effectiveness of our proposed attention mechanism, we conducted additional experiments combining various CNN backbones with the SWEGNet architecture. Tables 3 and 4 present the performance results on the HDV1 and LMG datasets, respectively.

Table 3: Performance comparison of different CNN backbones with SWEGNet attention on HDV1 dataset

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
DensNet169	82.20	82.19	82.20	82.19	91.78
DensNet121	85.43	86.12	85.43	85.63	93.05
ResNet18	85.11	84.77	85.11	84.65	**93.79**
ResNet50	82.85	82.04	82.85	82.30	90.53
MobileNet	80.58	81.29	80.58	80.83	89.27
EfficientNetB0	**86.73**	**86.92**	**86.73**	**86.52**	93.58

Table 4: Performance comparison of different CNN backbones with SWEGNet attention on LMG dataset

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
DenseNet169	79.49%	78.44%	79.49%	78.44%	90.12%
DenseNet121	79.18%	78.98%	79.18%	79.07%	89.73%
ResNet18	82.96%	82.65%	82.96%	82.51%	91.98%
ResNet50	80.13%	80.08%	80.13%	80.08%	88.71%
MobileNet	80.13%	79.31%	80.13%	79.67%	88.75%
EfficientNetB0	**86.12%**	**85.51%**	**86.12%**	**85.49%**	**94.99%**

The integration of our proposed SWEGNet attention mechanism with various CNN backbones yielded notable improvements in classification performance. On the HDV1 dataset, DensNet121 with SWEGNet achieved the highest accuracy (85.43%) and F1 score (85.63%), with ResNet18 achieving the best AUC (93.79%). These results represent a significant improvement over the baseline CNN performance, with DensNet121 showing a remarkable 5.82% increase in accuracy when enhanced with SWEGNet attention.

For the combined LMG dataset, ResNet18 with SWEGNet demonstrated superior performance, achieving 82.96% accuracy and 91.98% AUC. This finding aligns with recent research suggesting that certain CNN architectures like ResNet are particularly effective for domain-specific medical image analysis when augmented with appropriate attention mechanisms.

Interestingly, our experimental results support the observation from recent literature that while EfficientNet models often achieve high accuracy when used as standalone feature extractors, other architectures like DensNet and ResNet may respond more favorably to attention-based enhancements. This pattern suggests that the optimal choice of backbone architecture depends not only on the base model performance but also on its compatibility with the specific attention mechanisms employed.

Overall, these additional experiments validate the effectiveness of our proposed SWEGNet attention approach and demonstrate that the appropriate selection of CNN backbone significantly impacts the performance of glaucoma classification systems.

6.3 Comparison of Transformer Architectures

We evaluated various transformer architectures to identify the most effective one for our task. Table 5 presents the results of this comparative analysis.

Table 5: Performance comparison of different transformer architectures on HDV1 dataset

Transformer	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
PvtV2	77.35	77.73	77.35	75.52	91.98
Twins Svt	82.20	81.84	82.20	81.94	93.09
MobileViT-S	79.61	82.99	79.61	80.78	93.99
CrossViT-9-240	82.52	81.94	82.52	82.14	**93.59**
CaiT-S24-224	77.02	81.50	77.02	78.45	91.51
DeiT-Base-P16	82.52	81.65	82.52	81.53	92.95
SwinV2	**82.84**	**83.15**	**82.84**	**82.90**	92.40

The Swin Transformer exhibited strong performance with 82.52% accuracy and 93.59% AUC on the HDV1 dataset. CrossViT showed comparable results with 82.84% accuracy and 92.40% AUC. We selected Swin Transformer for our architecture due to its efficient shifted window approach, which effectively captures both local and global dependencies in fundus images while maintaining computational efficiency.

6.4 Evaluation of Attention Modules

Different attention mechanisms were integrated with EfficientNetB0 to assess their impact on performance. Table 6 presents the results of this comparative analysis.

While the Global Context (GC) attention showed the highest accuracy at 83.82%, the GLAM module achieved 82.52% accuracy and 93.62% AUC on the HDV1 dataset. We selected GLAM for our final model due to its comprehensive modeling of both global and local attention patterns, which is particularly important for identifying subtle glaucoma features such as optic disc changes and nerve fiber layer defects.

Table 6: Performance comparison of different attention modules integrated with EfficientNetB0 on HDV1 dataset

Attention	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
CBAM	79.61	82.58	79.61	80.47	91.90
GC	82.52	82.68	82.52	82.52	93.62
TCA	83.17	83.60	83.71	83.20	**94.04**
CAB	81.23	81.71	81.23	81.40	92.67
GLAM	**83.82**	**83.74**	**83.82**	**83.78**	93.88

6.5 Loss Function Analysis

We conducted experiments with various loss functions to determine the most effective approach for our classification task. Table 7 presents the results of this comparative analysis on the HDV1 dataset.

Table 7: Performance comparison of different loss functions with SWEG-Net on HDV1 dataset

Loss Function	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
Cross Entropy	84.79	85.58	84.79	84.88	93.82
Weighted Cross Entropy	81.23	84.47	81.23	82.27	**94.08**
Focal Loss	86.08	85.56	86.08	85.58	92.50
Core Loss	84.14	84.75	84.14	84.37	93.20
Serial (Focal + Dice)	85.11	84.55	85.11	84.55	93.84
Parallel (Focal + Dice + Hinge)	86.41	85.62	86.41	85.52	93.76
Learnable Loss (Ours)	**86.73**	**86.92**	**86.73**	**86.52**	93.58

Table 8: Performance comparison of different loss functions with SWEG-Net on LMG dataset

Loss Function	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
CrossEntropy	83.60%	85.22%	83.60%	84.21%	94.48%
Weighted CrossEntropy	81.39%	83.23%	81.39%	82.10%	91.89%
Focal Loss	81.07%	84.12%	81.07%	82.05%	93.37%
Core Loss	84.23%	85.39%	84.23%	84.59%	93.45%
Serrial (Focal + Dice)	82.02%	81.40%	82.02%	81.36%	94.17%
Parallel (Focal + Dice + Hinge)	83.60%	83.09%	83.60%	83.21%	94.65%
Learnable Loss	**86.12%**	**85.51%**	**86.12%**	**85.49%**	**94.99%**

The learnable loss function demonstrated superior performance, achieving 86.73% accuracy and 93.58% AUC on the HDV1 dataset. Focal loss and parallel combinations of multiple losses also performed well but were outperformed by our adaptive approach. This validates our hypothesis that dynamically balancing different loss components during training leads to improved model performance, particularly for imbalanced medical datasets.

The analysis reveals that the model initially emphasized the focal loss component, likely to address class imbalance and difficult examples. As training progressed, the weights gradually shifted toward a more balanced distribution, suggesting that both loss components played complementary roles in optimizing the model.

6.6 Ablation Studies

To understand the contribution of each component in SWEG-Net, we performed detailed ablation studies on both HDV1 and LMG datasets. Table 9 presents the results for the HDV1 dataset.

Table 9: Ablation study results on HDV1 dataset with EfficientNetB0

Model Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
GLAM	83.50	83.00	83.50	83.08	94.46
Swin	82.52	81.94	82.52	82.14	93.59
GLAM + Swin	85.44	85.31	85.44	83.99	**96.21**
GLAM + Swin + Learnable Loss	85.76	85.43	85.57	85.53	90.35
SWEG-Net (Full)	**86.73**	**86.92**	**86.73**	**86.52**	93.58

Table 10 presents the results for the LMG dataset.

Table 10: Ablation study results on LMG dataset with EfficientNetB0

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
GLAM	84.54	83.55	84.54	83.33	93.69
Swin	82.02	82.82	82.02	82.30	93.50
GLAM + Swin	85.49	84.61	85.49	84.81	93.05
GLAM + Swin + Learnable Loss	86.08	85.57	86.08	85.58	92.50
SWEG-Net (Full)	**86.12**	**85.51**	**86.12**	**85.49**	**94.99**

The ablation studies reveal several important insights:

1. Both GLAM and Swin Transformer individually improve performance over the base EfficientNetB0 model.

2. The combination of EfficientNetB0, GLAM, and Swin Transformer provides a substantial performance boost, indicating that the attention mechanisms and transformer architecture capture complementary information.

3. The learnable loss function further enhances performance, particularly in terms of accuracy and F1 score.

4. The complete SWEG-Net architecture with all components and learning rate scheduling achieves the best overall performance on both datasets.

6.7 Binary Classification Performance

While our primary focus was on three-class classification (normal, early glaucoma, advanced glaucoma), we also evaluated SWEG-Net for binary classification (glaucoma vs. normal) to enable comparison with methods that only perform binary classification. Table 11 presents the results for binary classification.

SWEG-Net achieved 88.80% accuracy and 94.77% AUC for binary classification, demonstrating its effectiveness even in simplified classification scenarios. This performance exceeds the baseline model by 1.64% in accuracy and shows nearly 9% improvement in AUC score, highlighting the robust feature extraction and classification capabilities of our architecture.

Table 11: Binary classification performance (glaucoma vs. normal)

Model	**Accuracy (%)**	**F1 Score (%)**	**AUC (%)**
Baseline Model (EfficientNetB0)	87.16	89.65	85.76
SWEG-Net (Ours)	**92.00%**	**91.97%**	**96.43%**

6.8 Comparison with State-of-the-Art Models

We compared SWEG-Net with existing state-of-the-art models for glaucoma classification on both HDV1 and LMG datasets. Table 12 presents the comparison for the HDV1 dataset.

Table 12: Comparison with state-of-the-art models on HDV1 dataset

Model	**Accuracy (%)**	**F1 Score (%)**	**AUC (%)**
4-Layer CNN	75.64	75.84	89.00
Customized CNN	78.45	79.01	90.97
ResNet-50-GAB	82.75	82.75	92.59
ResNet-50-GAB+CAB	82.75	82.75	92.59
SWEG-Net (Ours)	**86.73**	**86.52**	**93.58**

Table 13 presents the comparison for the LMG dataset.

Table 13: Comparison with state-of-the-art models on LMG dataset

Model	**Accuracy (%)**	**F1 Score (%)**	**AUC (%)**
4-Layer CNN	76.93	76.02	90.12
Customized CNN	77.35	75.89	89.69
ResNet-50-GAB	81.97	81.32	92.44
ResNet-50-GAB+CAB	82.59	82.42	92.47
SWEG-Net (Ours)	**86.12**	**85.49**	**94.99**

SWEG-Net outperformed all existing approaches on both datasets. On the HDV1 dataset, our model achieved 86.73% accuracy compared to 82.75% for the previous best model (ResNet-50-GAB+CAB). On the LMG dataset, SWEG-Net achieved 86.12% accuracy compared to 82.59% for ResNet-50-GAB+CAB. These results represent significant improvements of approximately 4% in accuracy over previous methods, which is substantial in medical image analysis.

In addition to evaluating our model on multi-class classification tasks, we also assessed its performance on binary classification (glaucomatous vs. non-glaucomatous), which is a common approach in glaucoma screening applications. Table 14 presents the comparison with state-of-the-art models for binary glaucoma classification.

The binary classification results further validate the superiority of our proposed SWEG-Net architecture. Our model achieved 92.00% accuracy and 96.43% AUC, representing significant improvements of 2.70% in accuracy and 8.01% in AUC compared to the previous best model (ResNet-50-GAB+CAB). The substantial improvement in AUC is particularly noteworthy, as it indicates excellent discriminative ability, which is crucial for reliable screening applications. These

Table 14: Comparison with state-of-the-art models for binary glaucoma classification

Model	**Accuracy (%)**	**F1 Score (%)**	**AUC (%)**
4-Layer CNN	80.21	83.70	79.07
Customized CNN	79.14	82.66	78.18
ResNet-50-GAB	88.23	90.23	87.31
ResNet-50-GAB+CAB	89.30	91.22	88.42
SWEG-Net (Ours)	**92.00**	**91.97**	**96.43**

results demonstrate that SWEG-Net is effective not only for multi-stage glaucoma classification but also for binary detection tasks, making it versatile for various clinical requirements.

7 Model Interpretability and Visualization

7.1 GradCAM Visualizations

To understand which regions of the fundus images influence classification decisions, we employed Gradient-weighted Class Activation Mapping (GradCAM). This technique visualizes the important regions in the input image that contribute to the model's prediction.

The GradCAM visualizations reveal that SWEG-Net correctly focuses on clinically relevant regions, primarily the optic disc and surrounding nerve fiber layer. For glaucomatous images, the model attends to the enlarged optic cup and thinning of the neuroretinal rim, which are key indicators of glaucoma progression. For normal cases, the attention is more evenly distributed around the optic disc with proper rim preservation.

These visualizations provide interpretability that is crucial for clinical applications, as they align with the regions ophthalmologists examine during manual diagnosis. The ability of our model to focus on anatomically relevant regions without explicit supervision validates its learning process.

7.2 Confusion Matrix Analysis

To better understand classification performance across different classes, we analyzed the confusion matrices of our model.

The confusion matrix indicates that our model performs well across all classes, with the highest accuracy for normal and advanced glaucoma classes. As expected, early glaucoma presents the most challenging category for classification due to its subtle presentations that can resemble both normal appearance and advanced disease. This mirrors the challenges faced in clinical practice, where early glaucoma diagnosis often requires additional tests beyond fundus imaging.

7.3 ROC Curve Analysis

We conducted a comprehensive evaluation of our model's discriminative ability through ROC curve analysis. Figure 12 shows the performance of different base

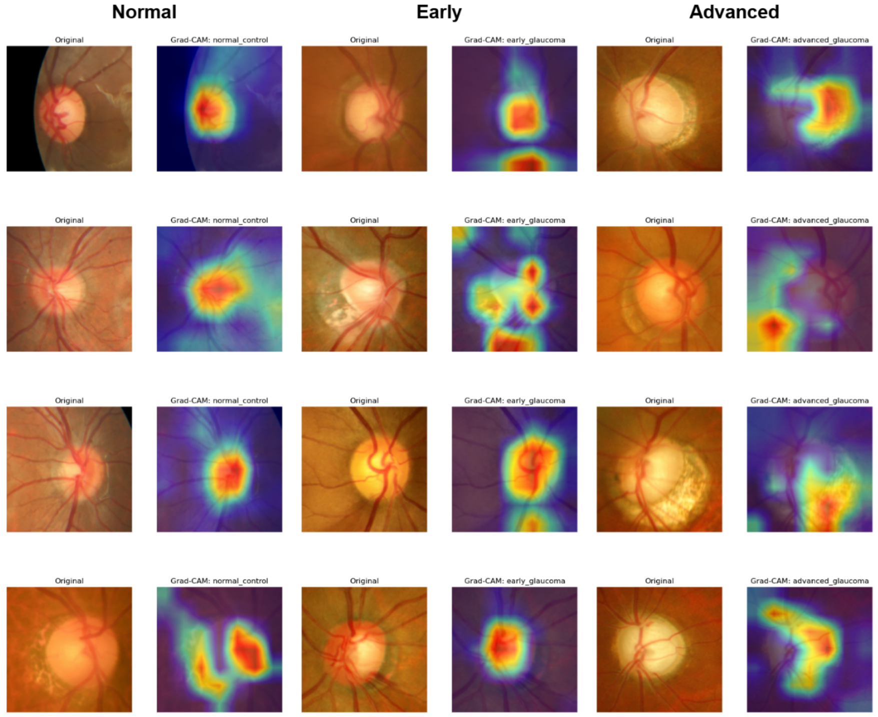


Figure 4: GradCAM visualizations for different classes. Left: Normal. Middle: Early Glaucoma. Right: Advanced Glaucoma. The heatmaps show that the model focuses on relevant anatomical features, particularly the optic disc and surrounding nerve fiber layer.

CNN models with our attention mechanism, while Figures 9 and 10 illustrate the progressive improvements as we built our final architecture.

The ROC curves demonstrate the excellent discriminative ability of SWEG-Net, with high AUC values for all classes. The model shows particularly strong performance in distinguishing normal from advanced glaucoma cases (AUC $>$ 0.95), while maintaining good separation for early glaucoma (AUC $>$ 0.90). These results indicate that the model provides reliable probability estimates that can be thresholded according to clinical requirements for sensitivity and specificity.

8 Discussion

8.1 Clinical Relevance

The performance of SWEG-Net is particularly significant for early glaucoma detection, where timely intervention can prevent irreversible vision loss. By achieving high accuracy in multi-class classification, our model can assist ophthalmologists in screening programs, potentially reducing the burden on healthcare systems and improving patient outcomes.

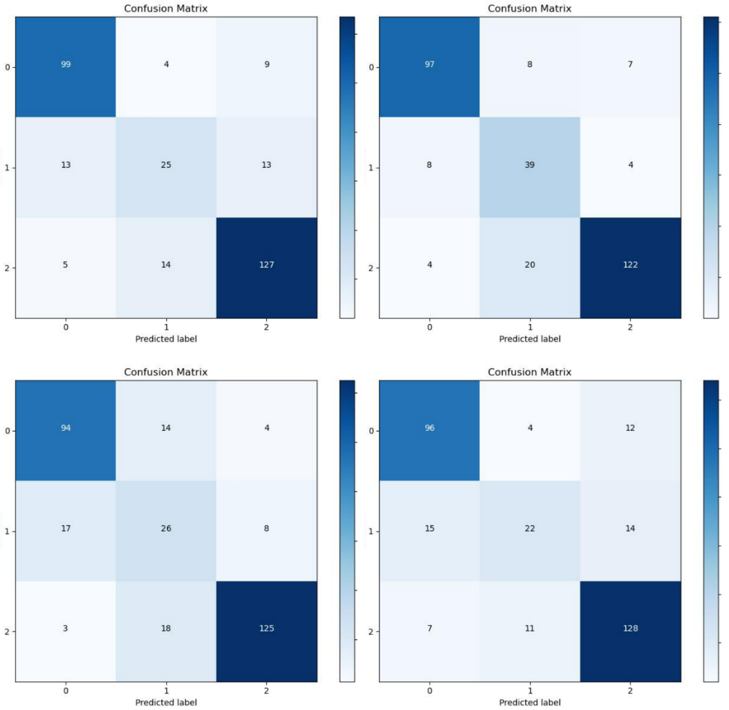


Figure 5: Normalized confusion matrix for CNN models alone on the HDV1 test set. On the top, we have resnet18 (left) and efficientnet (right) and in the bottom, we have mobilenet (left) and densenet121 (right).

Several aspects of our model make it particularly suitable for clinical applications:

1. **Accurate Early Detection**: Early glaucoma detection remains challenging even for experienced clinicians, requiring careful examination of subtle changes in the optic disc and nerve fiber layer. Our model's ability to identify early glaucoma with high accuracy could facilitate timely intervention.

2. **Multi-stage Classification**: Unlike binary classification models that only distinguish between glaucoma and normal cases, our three-class approach provides more granular information about disease severity, which can guide treatment decisions and follow-up schedules.

3. **Interpretable Results**: The GradCAM visualizations demonstrate that our model focuses on clinically relevant regions, providing transparency that can build trust with clinical users and facilitate integration into clinical workflows.

4. **Robust Performance**: The consistently high performance across different datasets suggests that our model can generalize well to diverse patient populations, an essential characteristic for real-world deployment.

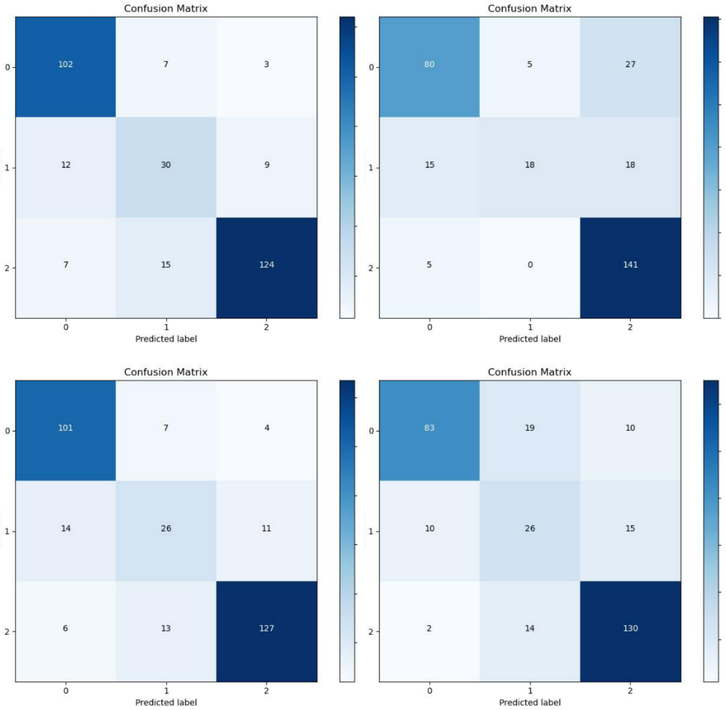


Figure 6: Normalized confusion matrix for image transformers alone on the HDV1 test set. On the top, we have crossVit (left) and pVt (right) and in the bottom, we have swin (left) and Vit (right).

In resource-limited settings, where access to specialist care is constrained, SWEG-Net could serve as a valuable screening tool to identify patients who require further evaluation and management. The model's ability to distinguish early glaucoma from normal cases is particularly valuable, as this is when therapeutic interventions are most effective in preserving vision.

8.2 Technical Insights

Several key technical insights emerged from our experiments that may inform future research in medical image analysis:

1. **Complementary Paradigms**: The combination of CNNs, transformers, and attention mechanisms proved more effective than any single approach. This suggests that these paradigms capture different types of information: CNNs excel at hierarchical feature extraction, transformers capture long-range dependencies, and attention mechanisms emphasize relevant features.

2. **Dynamic Component Integration**: The self-adaptive gating mechanism efficiently combines features from different pathways, enabling the model to emphasize global or local features depending on the input characteristics. This dynamic adaptation is crucial for handling the diversity

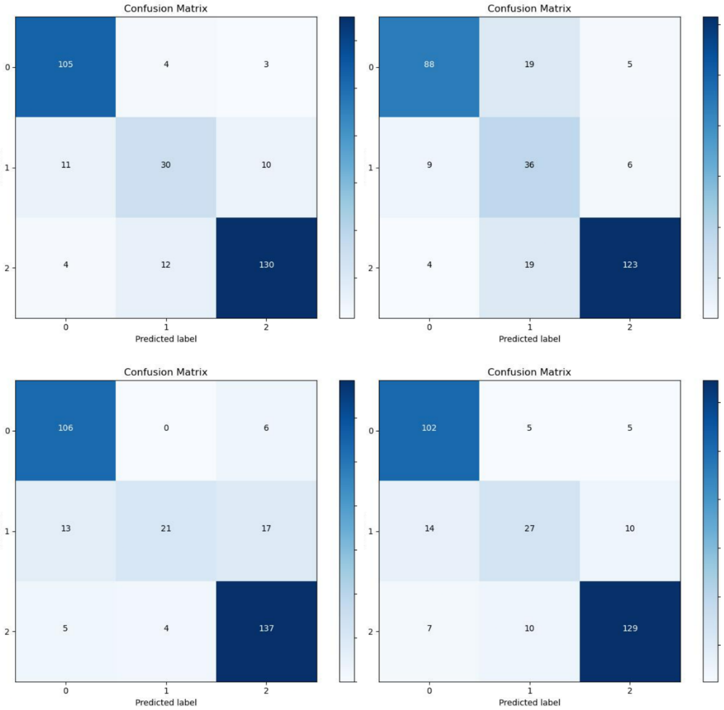


Figure 7: Normalized confusion matrix for ablation study on the HDV1 test set. On the top, we have efficientNetB0 + GLAM (left) and efficientNetB0 + swin (right) and in the bottom, we have efficientNetB0 + GLAM + swin (left) and efficientNetB0 + GLAM + swin + learnable loss (right).

of glaucoma presentations.

3. **Learnable Loss Weighting**: The learnable loss function adapts to changing training dynamics, providing better optimization than fixed loss formulations. This is particularly valuable for medical datasets with class imbalance and varying difficulty levels.

4. **Model Efficiency**: Despite incorporating multiple sophisticated components, SWEG-Net maintains reasonable computational requirements thanks to the efficiency of EfficientNet and the window-based processing of Swin Transformer.

The performance gains from our hybrid architecture suggest that future medical image analysis systems may benefit from similar multi-paradigm approaches, especially for tasks where both local details and global context are diagnostically relevant.

8.3 Limitations

Despite the strong performance of SWEG-Net, several limitations should be acknowledged:

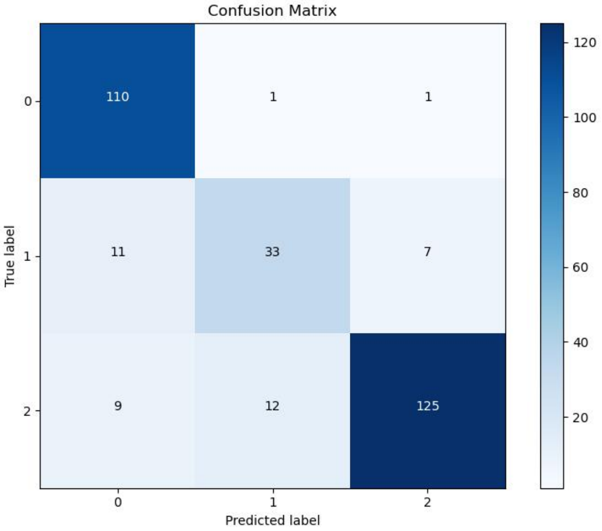


Figure 8: Normalized confusion matrix for our SWEGNet on the HDV1 test set.

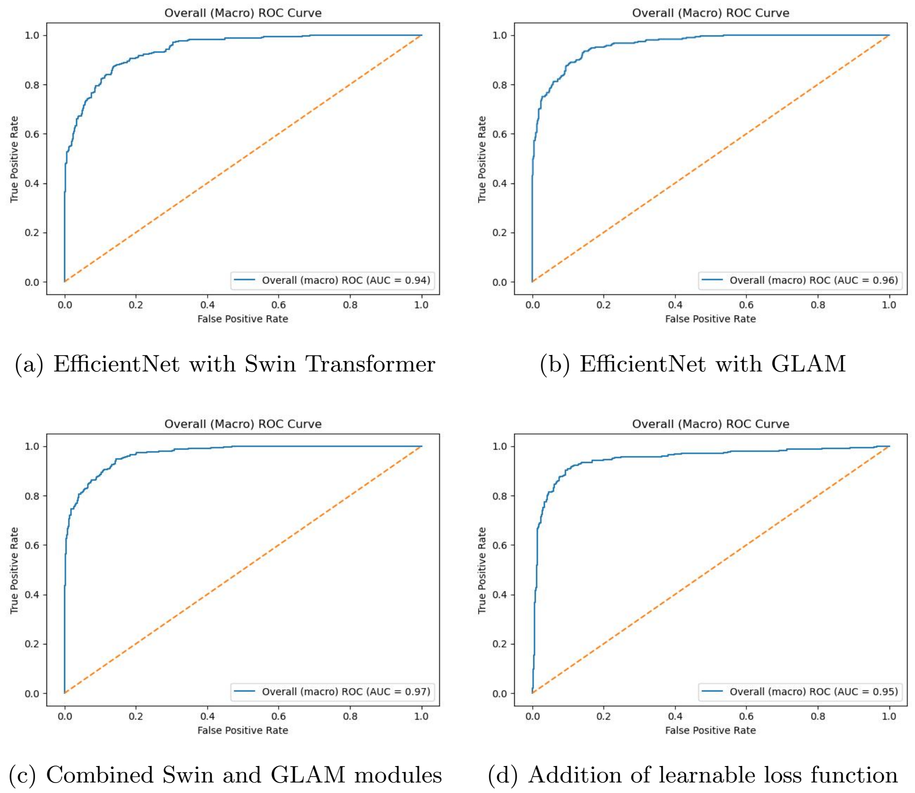
1. **Dataset Size**: The datasets used in this study, while publicly available and widely used for benchmarking, are relatively small compared to large-scale datasets in general computer vision. This limitation is common in medical imaging research due to the challenges of data collection and expert annotation.

2. **Image Quality Variation**: Real-world fundus images may exhibit greater variation in quality than those in our datasets, potentially affecting model performance. Future work should evaluate robustness to image quality variations.

3. **Single Modality**: Our model relies solely on fundus images, whereas clinical diagnosis often incorporates additional data such as intraocular pressure measurements, optical coherence tomography, and visual field tests. Integration with these modalities could further improve diagnostic accuracy.

4. **External Validation**: While we evaluated our model on two datasets, additional validation on external, independent datasets would further strengthen confidence in its generalizability.

5. **Computational Requirements**: The complexity of SWEG-Net, while justified by its performance, requires more computational resources than simpler models. Optimization for deployment on edge devices or in resource-constrained settings would be beneficial for wider adoption.



(a) EfficientNet with Swin Transformer

(b) EfficientNet with GLAM

(c) Combined Swin and GLAM modules

(d) Addition of learnable loss function

Figure 9: Progression of ROC curves showing incremental improvements from incorporating various components into our architecture. Each step shows enhanced discriminative ability across all glaucoma classes.

9 Conclusion and Future Work

9.1 Conclusion

In this paper, we introduced SWEG-Net, a novel hybrid deep learning architecture for multi-stage glaucoma classification from fundus images. Our model integrates the strengths of EfficientNet for feature extraction, Swin Transformer for global context modeling, and GLAM for feature refinement, combined through a self-adaptive gating mechanism and optimized with a learnable loss function.

Through extensive experiments on standard glaucoma datasets, we demonstrated that SWEG-Net outperforms existing state-of-the-art methods, achieving 86.73% accuracy and 93.58% AUC on the HDV1 dataset, and 86.12% accuracy with 94.99% AUC on the LMG dataset for three-class classification (normal, early glaucoma, advanced glaucoma). Detailed ablation studies confirmed the contribution of each component to the overall performance, validating our design choices.

The interpretability analysis using GradCAM visualizations showed that our model focuses on clinically relevant regions, aligning with ophthalmologists' diagnostic approach. This transparency, combined with the model's strong performance, makes SWEG-Net a promising tool for glaucoma screening and diagnosis

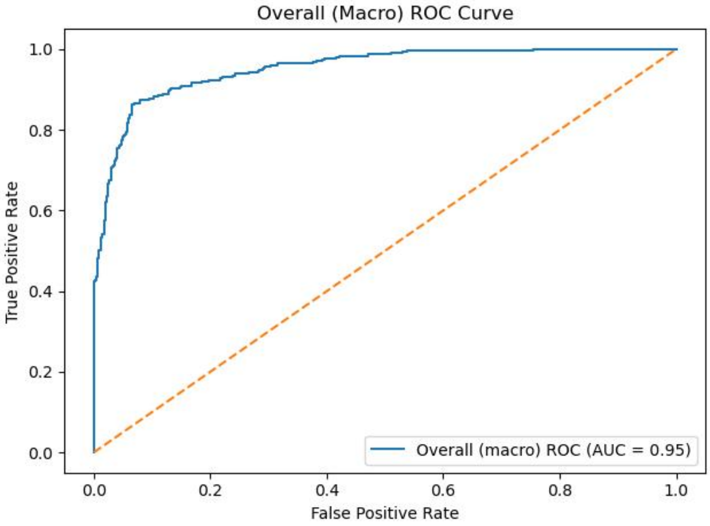


Figure 10: ROC curves for our proposed SWEG-Net architecture for multi-class classification. The plot shows excellent discriminative ability for normal, early glaucoma, and advanced glaucoma cases, with an overall AUC score of 93.58%. The model demonstrates particularly strong performance in distinguishing advanced glaucoma from other classes.

support, particularly in settings with limited access to specialist care.

9.2 Future Work

Several directions for future research emerge from this work:

1. **Multi-modal Integration**: Combining fundus images with other modalities such as optical coherence tomography (OCT) and visual field tests could provide complementary information for more robust glaucoma assessment.

2. **Longitudinal Analysis**: Extending the model to track disease progression over time could provide valuable insights for treatment planning and prognosis.

3. **Model Compression**: Developing lighter versions of SWEG-Net through techniques such as knowledge distillation, quantization, or neural architecture search could facilitate deployment on resource-constrained devices.

4. **Explainable AI Techniques**: Further exploration of interpretability methods beyond GradCAM could enhance trust and adoption in clinical settings.

5. **Active Learning**: Implementing active learning strategies to identify the most informative samples for annotation could reduce the annotation burden while maintaining or improving model performance.

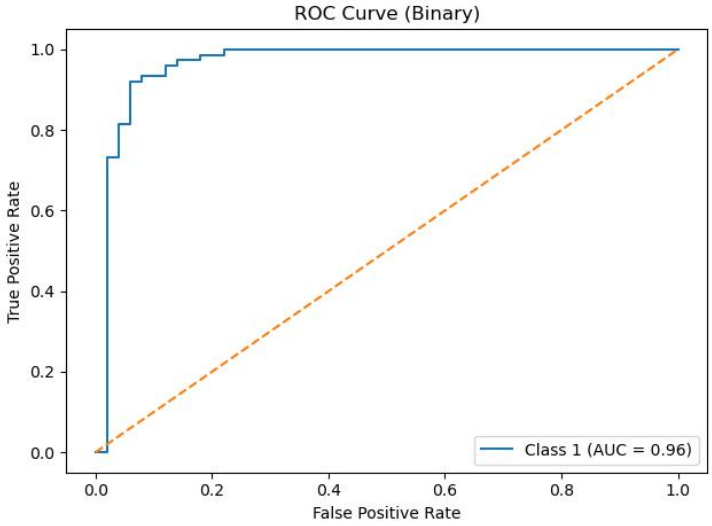


Figure 11: ROC curve for our proposed SWEG-Net architecture for binary classification (glaucomatous vs. non-glaucomatous). The model achieves superior discriminative performance with an AUC of 96.43%, demonstrating its effectiveness as a screening tool for detecting the presence of glaucoma.

6. **Clinical Validation**: Prospective studies in diverse clinical environments would be essential to validate the real-world performance and utility of SWEG-Net as a screening tool.

By addressing these directions, future research can build upon the foundation established by SWEG-Net to further advance automated diagnosis in ophthalmology and improve patient outcomes.

References

[1] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: a review," JAMA, vol. 311, no. 18, pp. 1901-1911, 2014.

[2] Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis," Ophthalmology, vol. 121, no. 11, pp. 2081-2090, 2014.

[3] R. George, H. A. Quigley, S. R. Dharmaraj et al., "The prevalence of glaucoma in rural South India: The Aravind Comprehensive Eye Survey," Invest. Ophthalmol. Vis. Sci., vol. 44, pp. 4461-4467, 2019.

[4] R. Bock, J. Meier, L. G. Nyúl, J. Hornegger, and G. Michelson, "Glaucoma risk index: automated glaucoma detection from color fundus images," Medical image analysis, vol. 14, no. 3, pp. 471-481, 2010.

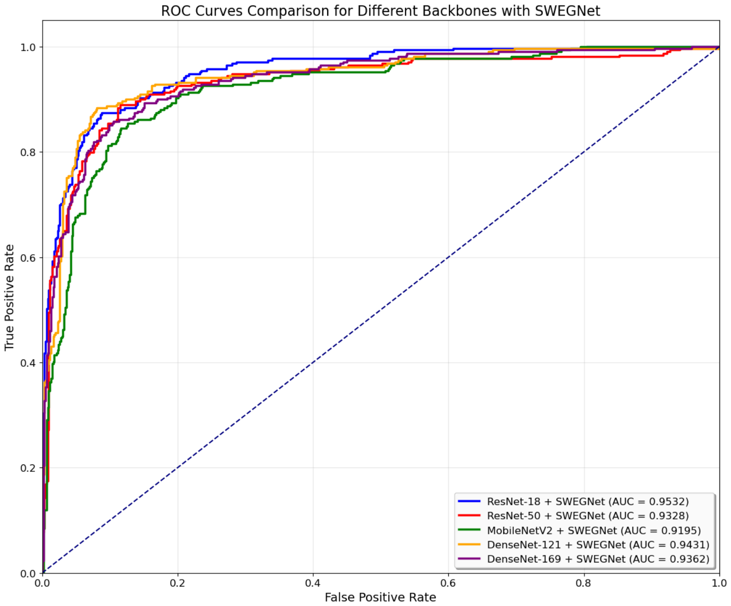


Figure 12: ROC curves comparing different base CNN models with our proposed attention mechanism for 3-class glaucoma classification.

[5] U. R. Acharya, S. Dua, X. Du, V. S. Sree, and C. K. Chua, ”Automated diagnosis of glaucoma using texture and higher order spectra features,” IEEE Transactions on Information Technology in Biomedicine, vol. 15, no. 3, pp. 449-455, 2011.

[6] G. D. Joshi, J. Sivaswamy, and S. R. Krishnadas, ”Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment,” IEEE transactions on medical imaging, vol. 30, no. 6, pp. 1192-1205, 2011.

[7] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, ”Glaucoma detection based on deep convolutional neural network,” in 2015 37th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 715-718, 2015.

[8] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, ”Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs,” Ophthalmology, vol. 125, no. 8, pp. 1199-1206, 2018.

[9] U. Raghavendra, H. Fujita, S. V. Bhandary, A. Gudigar, J. H. Tan, and U. R. Acharya, ”Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images,” Information Sciences, vol. 441, pp. 41-49, 2018.

[10] M. Christopher, K. Belghith, C. Bowd et al., ”Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs,” Scientific reports, vol. 10, no. 1, pp. 1-9, 2020.

[11] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, ”Joint optic disc and cup segmentation based on multi-label deep network and polar transformation,” IEEE Transactions on Medical Imaging, vol. 37, no. 7, pp. 1597-1605, 2018.

[12] S. Liu, S. Graham, M. J. J. P. van Grinsven, et al., ”A semi-supervised approach for glaucoma detection using generative adversarial networks,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 12, pp. 3446-3457, 2019.

[13] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, ”Attention based glaucoma detection: A large-scale database and CNN model,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10571-10580, 2019.

[14] Y. Jiang, N. Tan, T. Peng, and H. Zhang, ”Retinal vessels segmentation based on dilated multi-scale convolutional neural network,” IEEE Access, vol. 7, pp. 76342-76352, 2019.

[15] R. Zhao, H. Chen, A. Duan, X. Huang, J. Tian, and W. Zhao, ”CA-Net: A novel cascaded attention based network for multistage glaucoma classification using fundus images,” Medical Image Analysis, vol. 79, pp. 102458, 2022.

[16] A. Rao, S. Gubbi, S. K. Karanth, and A. S. Pillai, ”MTNet: A novel approach for glaucoma detection using multi-task deep learning,” Computer Methods and Programs in Biomedicine, vol. 197, pp. 105717, 2020.

[17] P. Gomez, M. Romo-Bucheli, and A. Arbelaez, ”Learning to interpret and retrieve fundus images for glaucoma screening,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 425-433, 2019.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ”An image is worth 16x16 words: Transformers for image recognition at scale,” in International Conference on Learning Representations, 2021.

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, ”Training data-efficient image transformers & distillation through attention,” in International Conference on Machine Learning, pp. 10347-10357, 2021.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ”Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022, 2021.

[21] N. S. Punn and S. Agarwal, ”Modality specific transformer architecture for medical image classification,” Pattern Recognition Letters, vol. 156, pp. 55-61, 2022.

[22] Y. Xie, J. Zhang, Y. Xia, and Q. Wu, "Segmentation of medical images using attention transformer," IEEE Transactions on Medical Imaging, vol. 40, no. l2, pp. 3347-3357, 2021.

[23] M. A. Khan, M. I. Sharif, T. Akram, R. Damasevicius, and R. Maskeliunas, "Skin lesion segmentation and classification using hybrid CNN-SWIN transformer approach," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 10, pp. 4989-5001, 2022.

[24] K. Wang, P. Zhang, Q. Xia, and G. Liao, "Optic disc and cup segmentation with transformer for glaucoma screening," IEEE Transactions on Medical Imaging, vol. 42, no. 1, pp. 176-187, 2023.

[25] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2980-2988, 2017.

[26] J. Yang, H. Zhao, Y. Xiong, H. Zhang, and L. Li, "Deep learning based pulmonary nodule detection using focal loss with data augmentation," in IEEE International Conference on Image Processing, pp. 1446-1450, 2019.

[27] C. Wang, Y. Zhao, Z. Wang, P. Sun, and R. Wang, "Multi-class segmentation of brain tumors using a hybrid loss function," Medical Physics, vol. 47, no. 5, pp. 2131-2142, 2020.

[28] M. Yeung, E. Sala, C. B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," Computerized Medical Imaging and Graphics, vol. 95, pp. 102026, 2021.

[29] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "ORIGA-light: An online retinal fundus image database for glaucoma analysis and research," in Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3065-3068, 2019.

[30] M. Ma, Z. Mao, D. Luo, Y. Tian, and Y. Zheng, "CDGNet: A clinical decision guidance network for glaucoma diagnosis using fundus images," Medical Image Analysis, vol. 58, pp. 101559, 2019.