KEC Conference 2026

ISSN: 3021-940X (Print) DOI: 10.3126/injet.v3i2.95539

Glaucoma Disease Detection System Using Hybrid Deep Learning Architecture

Krishna Ojha[1*], Aaditya Kafle[2], Aryan Bhattarai[3], Gautam Gupta[4], Pralhad Chapagain[5]

[1]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, krishnaojhaa@outlook.com*

[2]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, akafle99@icloud.com*

[3]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, aryanbhattarai.dev@gmail.com*

[4]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, rauniyargautam777@gmail.com*

[5]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal, pralhadchapagain@kec.edu.np*

Abstract

Glaucoma is one of the major causes of permanent loss of vision in the world. Blindness can best be prevented by mass screening of the disease at its early stages. Nonetheless, the manual checking of eye photos is time consuming, and it is also vulnerable to human error. This paper describes a hybrid deep learning architecture that can be used to detect glaucoma by examining eye photographs. The suggested system isolates the procedures of locating the eye anatomy as well as categorizing the disease to simplify the outcomes and make them easy to comprehend. First, an Attention U-Net masks Optic Disc and Optic Cup to produce accurate black and white masks. The masks are added to the original color image to form a 5-channel input. The resulting combined input is then used by a Data-efficient Image Transformer (DeiT-Tiny) that then determines the physical distances between the eye structures. The hybrid architecture was tested on the SMDG-19 dataset of 12,449 images with accuracy of 89.77%, sensitivity of 84.41% and Area Under the Curve (AUC) of 0.9564. The results of the experiment indicate that the local feature extraction and global classification should be separated to produce an extremely useful and effective medical screening program tool. This study specifically focuses on glaucoma detection rather than general retinal diseases.

Keywords: Glaucoma Detection, Hybrid Architecture, Attention U-Net, Optic Disc Segmentation

1. Introduction

Glaucoma is an eye condition characterized by progressive optic nerve head degeneration, thus resulting in irreversible loss of vision [1]. Predictions regarding the condition indicate that 111.8 million cases of glaucoma would occur around the world by the year 2040 [2]. Since the disease often does not manifest any symptoms until it is in advanced stages, large-scale screenings that include color fundus photographs are necessary as a preventive measure. Manual assessment of the CDR ratio and the neuroretinal rim state is a fully subjective procedure [3]. As part of diagnosis, ophthalmologists analyze the optic nerve head by assessing the CDR and the neuroretinal rim according to the ISNT (Inferior, Superior, Nasal, Temporal) rule. A CDR greater than 0.6 is typically considered indicative of glaucoma. Automated detection of this feature is essential for screening purposes.

Automated Computer-Assisted Diagnosis (CAD) approaches were developed based on the application of machine learning techniques [4] [5]. The standard workflow for such algorithms involves the application of deep learning approaches based solely on CNNs [6] [7]. Even though this approach ensures effective classification based on probabilities, their non-explainability prevents identification of those features that played an important role in the classification [8] [9]. Furthermore, the localized nature of convolution filters does not allow for a thorough global geometry evaluation of the optic nerve head. In this paper, we propose an explainable model that can overcome this engineering problem. This model uses Attention U-Net to

* *Corresponding author*

localize anatomical structures of interest and Data-efficient Image Transformer (DeiT-Tiny) for classification purposes only. This way, a global analysis by the Transformer based on explicitly bound structural information simulates the human diagnostic approach. The presented model produces high accuracy rates and has low inference costs allowing for remote deployment. Unlike existing hybrid architectures such as TransUNet and UTNet, which integrate segmentation and classification into a single model, the proposed system enforces a strict separation between anatomical localization and classification. This improves interpretability and aligns more closely with clinical diagnostic workflows.

2. Related Works

2.1 Convolutional Neural Networks in Ophthalmology

Early successes in deep learning for automated ophthalmology were mainly dependent on standard CNNs to categorize raw fundus images. As the medical database was relatively smaller compared to the commercial image dataset, early work was highly dependent on transfer learning [7]. Models such as ResNet [6] were pre-trained using large image databases such as ImageNet and then fine-tuned for detecting retinal disease. Although these deep residual neural networks addressed the issue of vanishing gradient and attained great accuracy, there is one drawback with their architecture: they depend on localized receptive fields. The process of standard pooling reduces the image size in a systematic way, which causes the loss of hierarchical spatial information. When considering glaucoma diagnosis, the physical distance between the edge of the optic cup and the edge of the Optic Disc) becomes the determining factor. Previous studies have also used VGG-16 as a feature extractor for medical image classification. However, deeper architectures like ResNet50 provide improved gradient flow and better feature representation, which motivated its selection in this work.

2.2 Semantic Segmentation and Attention Mechanisms

To obtain explicitly measured morphology, the use of whole-image classification in clinical research was replaced by semantic segmentation. The emergence of the U-Net model [10] set the benchmark in medical image segmentation. U-Net adopts symmetrical encoder-decoder architecture augmented with skip connections that circumvent the bottleneck, delivering high-resolution spatial information from the contracting path to the expanding path. It turned out to be very effective in detecting the Optic Disc [11]. The detection of the optic cup is a much more challenging task due to the presence of retinal blood vessels, which makes the segmentation much harder. To overcome this issue, the Attention U-Net was invented [12]. Its key feature lies in introducing Attention Gates (AGs) within the skip connections. Rather than sending everything across the skip connection, AGs create an attention map, where mathematically important details (the optic cup) are highlighted, while distracting information (blood vessels and other parts of the retina) is suppressed. This specific feature selection becomes the primary localization step of the hybrid method under consideration.

2.*3 Data Standardization and Class Imbalance Handling*

Both the performance of CNNs and segmentations directly depends on the quality of the input images. Fundus photos are notorious for their tendency towards nonuniform lighting, light reflection from the center, and contrast variability due to the specific camera hardware used. Multiple studies showed that preprocessing pipelines must be automated [13]. Using Contrast Limited Adaptive Histogram Equalization (CLAHE) [14], it is possible to suppress noise amplification and increase the contrast between retinal vessels and the background tissue. In addition, medical databases are characterized by extremely severe class imbalance, where a significant portion of samples correspond to healthy individuals. This leads to an oversampling of the majority class during learning, causing high overall accuracy at the price of unacceptable false negatives. To alleviate class imbalance, training algorithms currently utilize intensive augmentations through rotation and flips of images. Most importantly, current approaches rely on weighted cross-entropy (WCE). This approach allows adding a particular mathematical cost to a minority class and, therefore, making the gradient optimizer more inclined towards predicting it [15].

2.4 Vision Transformers and Data Efficiency

As an attempt to eliminate all spatial limitations caused by local convolutions, Vision Transformers (ViTs) [16] were adapted for use in medical imaging [17]. Unlike convolutional networks, ViT models omit convolutions. Instead, they split images into flattened patches and use Multi-Head Self-Attention (MHSA). Using MHSA, ViTs can exchange information between all image patches, creating a global understanding of image geometry. While ViTs offer superior architecture for capturing long-range dependencies, they do not have many inductive biases common to CNNs such as translation invariance, thus requiring large datasets to avoid overfitting during training. To make Transformers practical for small-scale medical data sets, researchers proposed a new model called the Data-efficient Image Transformer (DeiT) [18]. One of the main innovations brought by DeiT is the distillation token which allows training Transformers with a teacher CNN model. Thus, DeiT provides excellent global classification performance while using fewer parameters, which makes it highly suitable for implementation in Medical IoT environments.

2.5 Hybrid Architectures and Explainability

Considering that convolutional networks are good at extracting features locally, and Transformers are good at modeling context globally, recent research in the field has resulted in hybrid models being used as the advanced approach. Networks like TransUNet [19] and UTNet [20] manage to incorporate Transformer self-attention mechanisms into bottleneck layers of convolutional networks. Other advanced models utilize gating axial attention mechanism [21] and non-local neural networks [22] for capturing spatial information during the feature extraction step. However, there is one significant drawback associated with applying these models to a medical setting: explainability [8] [9]. Although hybrid architectures such as TransUNet and UTNet integrate convolutional and Transformer-based mechanisms, they typically combine segmentation and classification within a single end-to-end model. This limits interpretability, as the decision-making process remains embedded within a unified architecture. In contrast, the proposed approach introduces a strict functional separation between anatomical localization and disease classification. The Attention U-Net explicitly generates Optic Disc and Optic Cup masks, which are then used as structured inputs for the DeiT classifier. This design enforces transparency by allowing visual verification of intermediate outputs, thereby aligning the model's behavior with clinical diagnostic procedures. This explicit two-stage separation represents the primary novelty of the proposed framework.

3. Methodology

The proposed framework operates as a two-stage sequential pipeline which is shown in figure 1 which is below the page.

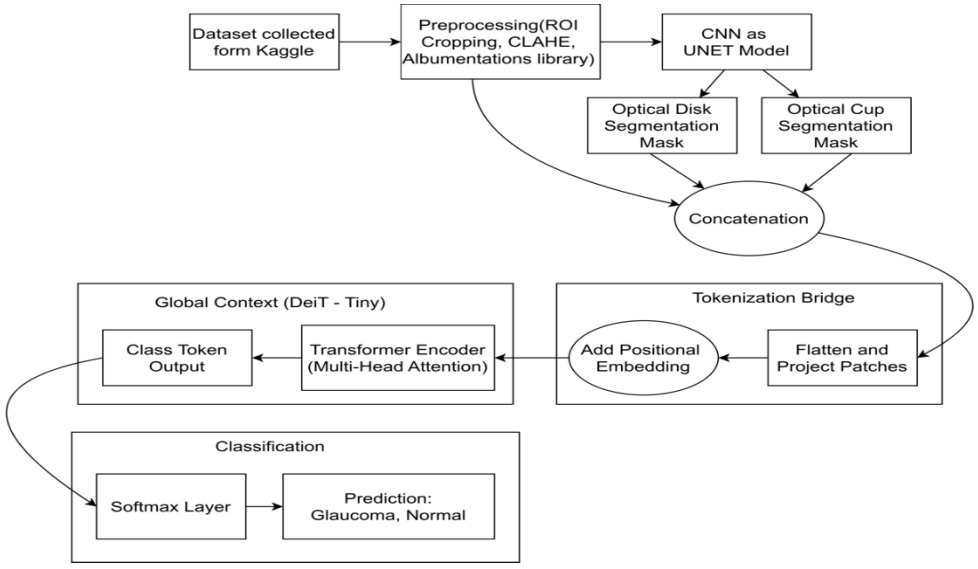

Figure 1. Working Mechanism of Glaucoma Disease Detection System

3.1 Data Preprocessing and Standardization

The system was tested and trained using the SMDG-19 standardized data set [23]. This is a large database that has been constructed using 19 different public sources of fundus images. The data set consists of a total of 12,449 images that have been proven and consist of 7,549 normal images and 4,767 glaucomatous images. Glaucoma suspect images were not allowed to be included in the training process to ensure that there is an evident distinction between normal and diseased. The open source CV pipeline takes care of the image processing. The green chromatic channel of the raw image is isolated to enhance the retinal vascular and Optic Disc structure, which then undergoes CLAHE to balance uneven lighting. A dynamic foreground thresholding technique estimates the ocular focal point, cropping the ROI to a standard size of 512 × 512 square pixels to reduce periphery noise.

3.2 Anatomical Localization via Attention U-Net

In the first stage, an adapted Attention U-Net, utilizing the pre-trained ResNet50 encoder, is employed. This encoder extracts the hierarchical semantic features while the decoder restores spatial resolution. Information flows through skip connections from the encoder to the decoder, delivering high-resolution spatial maps. In these connections, Attention Gates selectively inhibit irrelevant background activation. The output of the neural network includes probability maps that have been converted to discrete binary masks for the Optic Disc and Optic Cup by means of an argmax function.

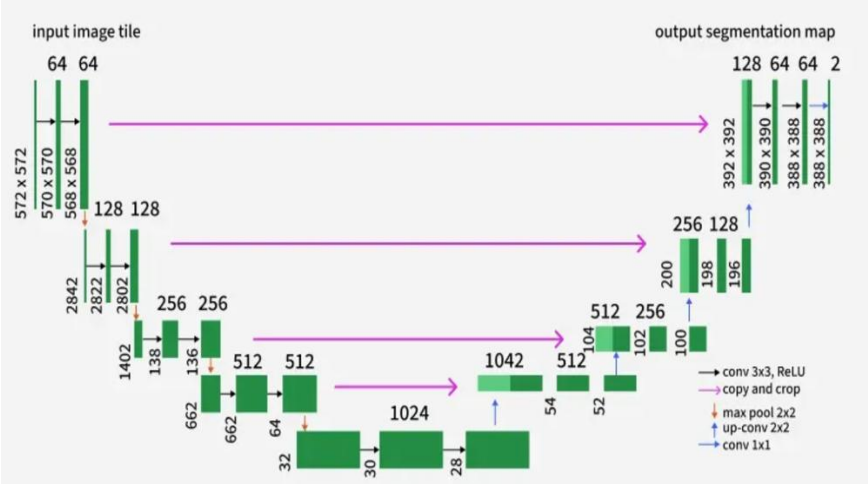

Figure 2. Architecture of Attention U-net

An Attention U-Net, as the major feature extractor, was utilized, where the pre-trained ResNet50 encoder was utilized. A probability map of the pixels was produced by the model. It follows an encoder-decoder structure. The encoder consists of repeated Convolution2D layers followed by ReLU activation and max-pooling operations for down-sampling, which reduces spatial dimensions while increasing feature depth. The decoder performs up-sampling using interpolation followed by convolution operations to reconstruct spatial resolution. Skip connections transfer high-resolution features from encoder to decoder.

Attention Gates are applied to filter irrelevant features and focus on important anatomical regions such as Optic Disc and cup. An argmax mathematical operation was performed to produce discrete black and white masks for the Optic Disc and Optic Cup.

3.2.1 Basic Operations in U-Net

Convolution2D applies learnable filters to extract spatial features. Down-sampling is performed using max-pooling, which reduces spatial resolution and helps capture global context. Up-sampling restores spatial resolution using interpolation techniques, allowing precise localization. These operations collectively enable accurate segmentation of retinal structures.

$$S_{i,j} = (I * K)i,j = \sum_m \sum_n I_{m,n} K_{i-m,j-n} \quad \text{(Equation 1)}$$

Where I denote the input image, K denotes the two-dimensional filter (kernel), and S denotes the output feature map. After extracting these features, a probability map of the pixels was created.

3.3 Multi-Spectral Tensor Fusion

To bridge localization and classification, the system constructs a 5-channel multi-spectral tensor. The original RGB fundus image is down sampled to 224×224 pixels and concatenated along the channel axis with the two extracted binary masks. This input structure forces the subsequent classification model to evaluate the spatial relationship between the highlighted anatomical structures explicitly.

3.4 Global Classification via DeiT-Tiny

The 5-channel tensor undergoes processing through a DeiT-Tiny architecture. The pre-trained projection layer is modified to handle the five channel inputs. The tensor undergoes partitioning into 14 x 14 non-overlapping 16 x 16 patches which undergo flattening and tokenization. The MHSA computes the distance and deformation between the cup and the disc boundaries at the same time using this tokenization. The diagnosis comes out of the MLP head in the form of binary classification. Using this DeiT-Tiny architecture greatly simplifies computation, in line with medical IoT edge computing [24].

The most innovative idea proposed in this paper is the idea of data mixing. The first RGB image used was downscaled into 224x224 and normalized. The three-channel RGB image was concatenated with the separated masks producing the 5-channel input image.

The underlying framework utilized to build the system is PyTorch [25]. Specifically, the pre-trained layer of the DeiT-Tiny model was modified to accept a 5-channel input. The input image was further split into 14×14 grid and 196 unique patches were extracted. The self-attention mechanism employed in each transformer block compared the scores between the patches, allowing the model to determine the physical distance and relations between the optic cup and disc.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(Equation 2)}$$

Where Q is a query matrix, K is a key matrix, V is a value matrix, and dk is the dimension scaling factor to avoid vanishing gradients while training.

4. Experiment and Result

4.1 Dataset Overview

The suggested architecture has been trained and validated on SMDG-19 standardized benchmark dataset [25]. The dataset comprises 12,449 color retinal scans with manual annotations performed by experts, including 7,549 healthy eyes and 4,767 cases diagnosed with glaucoma and the remaining 133 were glaucoma suspects. To ensure a proper evaluation and avoid any kind of data leak, the dataset was carefully separated into 3 subsets: 70% for training, 15% for internal cross-validation during training, and 15% for external blind evaluation. Data characteristics of 10 retinal images are presented in the following table.

Table 1. SMDG-19 Dataset Overview (10 Samples)

ID	File Name	Age	Sex	Eye	Label	Diagnosis	OD Mask	OC Mask	Vertical CDR
1	OIA-ODIR-TRAIN-201.png	55	Male	Right	0	Non-Glaucoma	False	False	Null
2	ORIGA-103.png	62	Female	Left	0	Non-Glaucoma	True	True	0.51
3	OIA-ODIR-TRAIN-2109.png	48	Female	Right	0	Non-Glaucoma	False	False	Null
4	G1020-706.png	71	Male	Left	1	Glaucoma	True	True	0.85
5	OIA-ODIR-TRAIN-2362.png	50	Male	Right	0	Non-Glaucoma	False	False	Null
6	sjchoi86-HRF.png	45	Female	Left	1	Glaucoma	True	False	Null
7	REFUGE-Train-015.png	68	Male	Right	1	Glaucoma	True	True	0.78
8	ORIGA-055.png	42	Female	Right	0	Non-Glaucoma	True	True	0.44
9	G1020-022.png	74	Female	Left	1	Glaucoma	True	True	0.82
10	DRISHTI-GS-002.png	59	Male	Right	1	Glaucoma	True	True	0.75

4.2 Experimental Setup

The experiments were conducted on an NVIDIA Tesla P100 GPU using the PyTorch framework in Linux as well as Windows operating systems [24]. The dataset was divided into training, validation, and test sets using a 70:15:15 ratio to ensure unbiased evaluation and prevent data leakage. The split was performed randomly while maintaining class distribution. The Attention U-Net model was initialized with a pre-trained ResNet50 encoder (ImageNet weights), while the DeiT-Tiny model also initialized with pre-trained weights and modified to accept 5-channel input. Training was performed for 50 epochs with a batch size of 16. The AdamW optimizer was used with an initial learning rate of 1e-4, combined with a cosine annealing learning rate scheduler to improve convergence stability [26].To improve generalization and address class imbalance, data augmentation techniques such as horizontal flipping, rotation (±15 degrees), and scaling were applied during training [27].

4.3 Evaluation Criteria

Performance evaluation metrics included accuracy, sensitivity, specificity, precision, and F1-score. As there is an imbalance in the dataset between healthy and glaucomatous subjects, weight cross-entropy loss function was employed as the objective loss function of the classification model. The minority class, Glaucoma, was prioritized through the weighting mechanism:

The primary metrics used to evaluate the classification models depend on the calculation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The mathematical formulas for these metrics are defined as follows:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad \text{(Equation 3)}$$

$$\text{Sensitivity (Recall)} = TP / (TP + FN) \quad \text{(Equation 4)}$$

$$\text{Specificity} = TN / (TN + FP) \quad \text{(Equation 5)}$$

$$\text{Precision} = TP / (TP + FP) \quad \text{(Equation 6)}$$

$$\text{F1-Score} = (2 \times \text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity}) \quad \text{(Equation 7)}$$

Sensitivity is particularly crucial in this medical context, as minimizing false negatives ensures that diseased patients are not mistakenly cleared as healthy during the screening process. To ensure reproducibility and clarity, the evaluation was performed on the held-out test set (15% of total data), which was not used during training or validation. Class imbalance was handled using weighted cross-entropy loss, where the glaucoma

class was assigned a higher weight (1.58) to penalize false negatives more heavily. All evaluation metrics were computed on the test set after training convergence. No cross-dataset evaluation was performed, which remains a direction for future work.

4.3.1 Loss Functions and Optimization

To assess the effectiveness of the pixel segmentation, Binary Cross-Entropy loss was used.

$$L_{BCE} = -[y \log(p) + (1 - y) \log(1 - p)] \quad \text{(Equation 8)}$$

Where y represents the actual medical ground truth label (1 for the target eye anatomy and 0 for the background), and p is the predicted probability that the pixel belongs to that target anatomy.

To have a balanced dataset, as there are much more healthy background images than diseased ones, a Weighted Cross-Entropy loss was employed to do the final classification [15]. The weight of 1.58 was added to the Glaucoma class so that it greatly penalizes the model when it failed to detect a positive glaucoma. The formula is expressed as:

$$L_{WCE} = -\sum_i w_i \, y_i \log(p_i) \quad \text{(Equation 9)}$$

Where w_i represents the specific weight assigned to class i (such as 1.58 for the Glaucoma class), y_i is the actual ground truth label for that class, and pi is the model's predicted probability for that class.

The AdamW optimizer was used to optimize the model with Cosine Annealing learning rate schedule [26].

4.4 Training and Validation Performance

Monitoring the training and validation loss and accuracy over the various epochs provided a clear view of how well each model learned. Four different setups were tested for comparison.

4.4.1 Hybrid Model (U-Net-DeiT):

Figure 3. Training and Validation Performance of the Proposed Hybrid Model

The accuracy of the model for testing data: 89.77%
The loss of the model for testing data: ~0.0510
The ROC AUC score of the model for testing data: 0.956

4.4.2 Comparative Visual Analysis

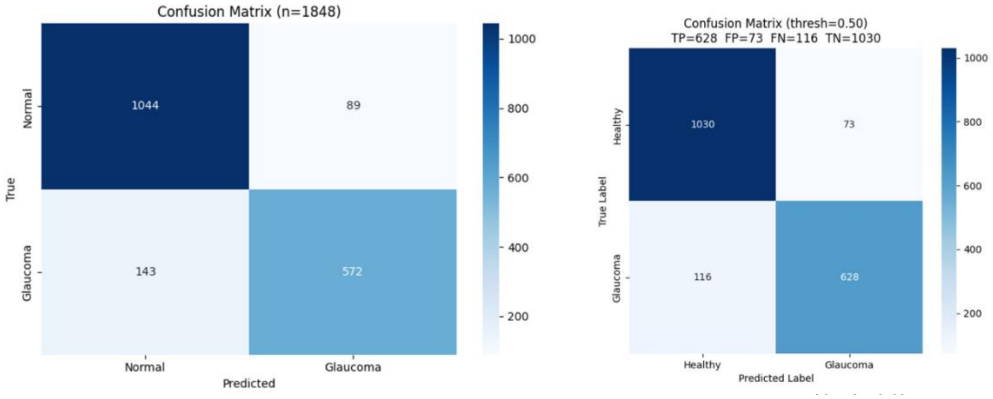

Figure 4. Comparative Confusion Matrices demonstrating the reduction of False Negatives in the proposed Hybrid Model (right) compared to the standalone ViT baseline (left)

For a visual representation of the advancements made by the novel framework, Figure 4 shows a side-by-side contrast of the confusion matrices from the best-performing baseline (ViT Only) and the Hybrid U-Net-DeiT framework. Even though the individual ViT achieves excellent accuracy across all cases, it is evident that there is a larger number of false negative predictions in identifying actual cases of glaucoma patients with healthy eyes. The false negative rate is considered the most hazardous type of error when working with diagnostic medical procedures in an actual screening setting. By integrating the information obtained regarding the Optic Disc and Optic Cup boundaries as input to the Transformer, the Hybrid framework forced the self-attention mechanism to directly analyze rim distortions, resulting in an impressive reduction in the false negative rate, with the sensitivity reaching 84.41%.

4.5 Results of the Experiment

Table 2. Comparative Performance Metrics of the Experimental Models

Model Architecture	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
U-Net Only	0.7527	0.6629	0.7343	0.7643	0.6967	0.8063
U-Net with SVM	0.7549	0.6770	0.7007	0.7891	0.6887	0.7875
Standard CNN with SVM	0.8063	0.7734	0.7063	0.8694	0.7383	0.8766
ViT Only	0.8745	0.8654	0.8000	0.9214	0.8314	0.9469
Hybrid Model	**0.8977**	**0.8959**	**0.8441**	**0.9338**	**0.8692**	**0.9564**

4.5.1 Architectural Evolution and Relative Analysis

The validation of the final hybrid model involved an analysis of the architecture used in the experiments to assess any improvements made to the accuracy of the system. The baseline U-Net algorithm had moderate diagnostic accuracy, which was 75.27%, and an Area Under the Curve (AUC) score of 0.8063 [28]. Adding SVMs to perform classification on the features identified by the U-Net improved overall accuracy by only 75.49% while decreasing sensitivity by a small amount (70.07%). As can be seen from the results, simply performing classification through standard statistical techniques does not allow for adequate processing of the spatial relations in a medical image.

With the introduction of attention mechanisms, there were notable improvements. Using a Vision Transformer (ViT) on its own to classify regular RGB images provided an accuracy of 87.45% and an AUC of 0.9469. The last and most efficient model, Hybrid Model, managed to outperform other models in all measures. Combining the strong ability to identify anatomical boundaries in a medical image using U-Net with a comprehensive global analysis of the image allowed achieving 89.77% accuracy, 84.41% sensitivity, 93.38% specificity, and 0.9564 AUC.

4.5.2 Qualitative Error Analysis

Even with the superior performance provided by the proposed hybrid system, it is evident that the review of the false negatives recorded (15.59%) presents certain clinical problems. It was found that the system performed poorly on two kinds of difficult images. First, there were images where glaucoma was severely present such that the

optic nerve head became invisible, making the detection of the anatomical boundary impossible. Secondly, with images of patients having glaucoma at an extremely early stage, the changes in structure within the optic cup did not meet the criteria set by the classifier.

4.5.3 Real-World Application and Computational Efficiency

While building an accurate diagnostic model is important, the second crucial step in implementing automated healthcare is making sure that such a solution can be physically implemented in a medical setting. Many current deep learning models rely on expensive custom hardware to work. This makes it impossible for them to use them in small clinics or underfunded health care centers. To make sure that this problem is overcome, DeiT-tiny was chosen as the core classifier. DeiT-tiny is designed in a way that minimizes computational load while still providing great accuracy. Thanks to the pre-processing filter, which removes corrupted images from further analysis, and the fact that calculations are focused only on crucial anatomical regions, this solution is stable and fast enough to run on regular hospital computers.

5. Conclusion and Future Enhancements

The study introduces a highly precise, completely interpretable hybrid framework for automated detection of glaucoma. In this study, the separation of anatomical characteristics detection (Attention U-Net) from global disease classification (DeiT-Tiny) overcomes the "black box" problem that is prevalent in conventional CNN architectures. On the SMDG-19 dataset, the proposed algorithm attained an accuracy of 89.77%, ROC-AUC of 0.9564. The adoption of DeiT-Tiny guarantees minimal computational expenses, thus ensuring the viability of the system in edge computing applications in resource-limited settings.

Further advancements will involve the implementation of state-of-the-art artifact correction algorithms that address the opacity of retinal images caused by glaucoma, resulting in a lower false negative rate. Furthermore, to ensure scalability while complying with existing patient data protection regulations, future versions will implement Federated Learning techniques [29]. While alternative architectures such as VGG-based models and advanced U-Net variants may further improve performance, the proposed system prioritizes interpretability and computational efficiency.

References

[1] T. Aung, R. Bourne, A. Bron, R. Ritch, and S. Panda-Jonas J. Jonas, "Glaucoma," The Lancet, vol. 390, 2017.

[2] X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng Y.-C. Tham, "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] Y. He, S. Keel, W. Meng, R. T. Chang, and M. He Z. Li, "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.

[4] J. E. Koh, J. H. Tan, S. V. Bhandary, A. Tyagi, H. Fujita, and U. R. Acharya Y. Hagiwara, "Com-

puter-aided diagnosis of glaucoma using fundus images: A review," Computer Methods and Programs in Biomedicine, vol. 165, pp. 1-12, 2018.

[5] D. S. W. Ting et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," Jama, vol. 318, no. 22, pp. 2211-2223, 2017.

[6] N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?" IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1299-1312, 2016.

[7] X. Zhang, S. Ren, and J. Sun K. He, "Deep residual learning for image recognition," in IEEE conf. on computer vision and pattern recognition, 2016, pp. 770-778.

[8] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793-4813, 2021.

[9] A. Holzinger et al., "Toward interpretable machine learning: transparent deep neural networks and beyond," Brain Informatics, vol. 4, pp. 59-70, 2017.

[10] P. Fischer, and T. Brox O. Ronneberger, "U-net: Convolutional networks for biomedical image segmentation," in Int. Conf. on Medical image computing and computer-assisted intervention, 2015, pp. 234-21.

[11] J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao H. Fu, "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] O. Oktay et al., "Attention u-net: Learning where to look for the pancreas," arXiv preprint, 2018.

[13] M. Abid, M. R. Ardali, J. Steen, and E. Amjadian R. Kiefer, "Automated fundus image standardization using a dynamic global foreground threshold algorithm," in 8th Int. Conf. on Image, Vision and Computing (ICIVC), 2023, pp. 460-465.

[14] K. Zuiderveld, "Contrast limited adaptive histogram equalization. USA: Academic Press Professional," Inc, pp. 474-485, 1994.

[15] A. Maki, and M. A. Mazurowski M. Buda, "A systematic study of the class imbalance problem in convolutional neural networks," Neural networks, vol. 106, pp. 249-259, 2018.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in ICLR, 2021.

[17] R. Fan et al., "Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization," Ophthalmology Science, vol. 3, 2022.

[18] M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou H. Touvron, "Training data-efficient image transformers & distillation through attention," in Int. conf. on machine learning (PMLR), 2021, pp. 10347–10357.

[19] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint, 2021.

[20] M. Zhou, and D. N. Metaxas Y. Gao, "Utnet: a hybrid transformer architecture for medical image segmentation," in Int. conf. on medical image computing and computer-assisted intervention, 2021, pp. 61-71.

[21] J. M. J. Valanarasu et al., "Medical transformer: Gated axial-attention for medical image segmentation," in MICCAI, 2021, pp. 36-46.

[22] R. Girshick, A. Gupta, and K. He X. Wang, "Non-local neural networks," in IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7794-7803.

[23] R. Kiefer. (2023) Smdg, a standardized fundus glaucoma dataset. [Online]. https://www.kaggle.com/ds/2329670

[24] S. Wang et al., "Lightweight deep learning models for edge computing in medical IoT," IEEE Internet Things, vol. 10, no. 5, 2023.

[25] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019.

[26] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," arXiv preprint, 2016.

[27] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of big data, vol. 6, no. 1, pp. 1-48, 2019.

[28] T. Fawcett, "Introduction to roc analysis," Pattern Recognition Letters, vol. 27, pp. 861-874, 2006.

[29] N. Rieke et al., "The future of digital health with federated learning," NPJ Digit. Med, vol. 3, no. 119, 2020.