

HYBRID EFFICIENTNET-TRANSFORMER ARCHITECTURE FOR AUTOMATED DETECTION OF GLAUCOMA

MALLA SIREESHA¹, MEKA JAMES STEPHEN², P.V.G.D. PRASAD REDDY³

¹Department of Information Technology and Computer Applications, Andhra University, Visakhapatnam, Andhra Pradesh, India

²Dr. B. R. Ambedkar Chair, Andhra University, Visakhapatnam, Andhra Pradesh, India

³Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India

E-mail: mallasireesha72@gmail.com, jamesstephenm@gmail.com, prasadreddy.vizag@gmail.com

ABSTRACT

In recent years, the early and accurate diagnosis of Glaucoma, a leading cause of permanent blindness, has highlighted the importance of effective treatment and better patient care. Conventional deep learning methods that only use Convolutional Neural Networks (CNNs) have shown significant performance in classification tasks involving fundus images, but they often fail to capture global contextual dependencies that are crucial for fine-grained disease categorization. Addressing this limitation, a Hybrid CNN-Transformer architecture has been proposed that uses the local feature extraction capability of a pretrained EfficientNetB0 backbone and augments it with the global modeling power of Transformer encoders. The EfficientNetB0 module extracts robust spatial features, which are improved using Transformer layers to model inter-feature relationships throughout the image. The proposed Hybrid EfficientNet-Transformer model was trained and validated on a dataset created by combining five datasets containing Fundus images, employing standard preprocessing pipelines and label encoding strategies. Extensive experiments show that the proposed hybrid approach attains a 99% high classification accuracy on the unseen test set, outperforming many existing HybridCNNs and transfer learning models. These results indicate that the fusion of CNNs and Transformers offers a powerful framework for high-precision glaucoma detection, creating new prospects for the development of automated ophthalmic screening tools.

Keywords: *Glaucoma, Convolutional Neural Networks (CNNs), EfficientNetB0, Fundus images, Transformer encoders.*

1. INTRODUCTION

Vision is one of the five primary senses and plays a major role in how people perceive and make sense of their surroundings. It is the result of your eyes and brain working in harmony, whereas blindness is the loss of vision [1]. Mainly there are two types of blindness: one is sudden blindness, and the other is blindness from birth. Congenital blindness is present from birth and is frequently more difficult or impossible to heal than sudden blindness, which can come on by an injury or illness. There are different reasons, such as eye conditions, injuries or infections, and nutritional deficiencies that causes sudden blindness. Globally, more than 2.2 billion people have eyesight loss, of which at least 1 billion cases are cured or improved, but many cases do not receive treatment. While the WHO defines blindness as having visual acuity (PVA) of 3/60 or

worse and eye damage as PVA less than 6/12, a Central European study considers different degrees of visual acuity and field defects [2].

The optic nerve, which transmits information from the eye to the brain, is harmed by glaucoma, a chronic, progressive eye condition. This damage often results from pressure inside the eye higher than normal, which leads to blindness that cannot be cured [3]. In the beginning, glaucoma usually doesn't cause noticeable symptoms like pain or vision changes. Some people may show very slow-developing symptoms. Because of this, many people don't realize they have it. Even though treatments and understanding have improved, glaucoma still causes around 10% of global blindness [4].

Detecting glaucoma involves several advanced imaging techniques, which are optical coherence

tomography (OCT): It helps detect changes to the optic nerve head and retinal nerve fiber layer. Fundus photography is a specialized camera that captures fine-grained images of the retina, optic disc, and blood vessels. SLP, or scanning laser polarimetry: This method measures the thickness of the retinal nerve fiber layer using laser light. Another method is known as confocal scanning laser ophthalmoscopy (CSLO). A 3D image of the optic nerve head is provided by this imaging technique, enabling an accurate assessment of its depth and form. These methods used fundus images because those are crucial in diagnosing various eye diseases [5]. When intraocular pressure (IOP) increases in the eye, it leads to structural changes inside the eye. This method measures the thickness of the retinal nerve fiber layer using laser light. The retina is composed of two key components: The optic cup is a central, bright area inside the optic disc (OD), which is where blood veins and nerve fibers travel to reach the retina. Higher cup-to-disc ratios are frequently indicative with glaucoma. [6].

Ophthalmologists typically depend on various manual techniques to diagnose glaucoma, such as gonioscopy, pachymetry, tonometry, and perimetry. While gonioscopy measures the angle between the cornea and iris, pachymetry measures corneal thickness, and tonometry measures intraocular pressure (IOP), a crucial sign of glaucoma. Despite their effectiveness, these techniques take a lot of time, subjective, and heavily reliant on the presence of experienced specialists, who are not available, especially in underserved or remote regions. That resulted in a demand for automated diagnostic systems that can detect glaucoma accurately and also improve accessibility and efficiency in eye care [7].

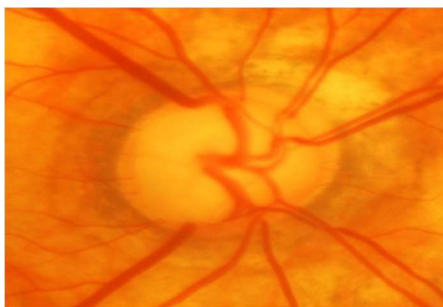


Figure 1: Glaucoma image taken from Acrima Dataset

The center of the eye, where the optic nerve fiber merges and leaves the eye, is seen in Figure 1. For the diagnosis of glaucoma, this area is crucial. If not detected and treated quickly, this aggressive kind of glaucoma that damages the optic nerve can result in permanent eye loss. Changes in the surface of the

eye, especially pupil dilation, are often the first noticeable sign of cataract. The optic cup is the center of the eye, and as glaucoma progresses, the cup becomes darker due to the gradual death of retinal cells. The cup-to-disc test (CDR), which gauges the diameter and thickness of the optical cup, is one of the most widely used procedures for glaucoma diagnosis. The cup is tiny and the CDR ranges from 0.2 to 0.3 in a healthy eye. However, in eyes with glaucoma, the cup can be more than half the diameter of the eye, resulting in a CDR of less than 0.6. In the image provided, the cup appears quite large, indicating a high CDR, which can be an indication of visual confusion.

In recent years, Artificial Intelligence (AI) has shown remarkable advancements with increasing integration into various healthcare applications [8] [9] [10]. A large number of automated tools that use AI are now employed for practical diagnostic and treatment purposes. One such application is the use of Computer-Aided Diagnostic (CAD) systems for the automated detection of glaucoma. The use of machine learning, especially deep learning, has grown significantly in medical image analysis due to its ability to make accurate predictions. This work has looked at several deep learning and hybrid models in an effort to find and improve models for precise glaucoma prediction.[11] [12] [13] [14].

2. LITERATURE REVIEW

There are many different research methods for glaucoma detection. Some pieces combine machine learning and image processing methods, while others only use deep learning. Patil et al. developed preprocessing and clustering techniques using the public dataset RIM-ONE and the private dataset DRSON-DB, respectively. [15]. Furthermore, Al-Kamdi et al. proposed a sequence segmentation-based method for early glaucoma detection that utilizes the RIM-ONE database. In order to provide datasets with sparse labels, a supervised CNN model called VGG16 was developed. The labeled data was then used to predict the raw data using a different classifier. A self-learning model, according to the authors, is one that is trained on unlabeled data to predict a significant amount of unlabeled data [16]. To detect glaucoma, Thakoor et al. employed a number of CNN models, both pre-trained and untrained. They use CNN models that have already been trained. A smaller dataset was utilized to evaluate the learned model [9]. In order to predict glaucoma in semi-supervised learning, Diaz-Pinto et al. used labeled image pairs and the DCGAN with semi-supervised learning algorithm (SS-DCGAN). Their suggested

approach has been evaluated on fourteen sets of glaucoma datasets, both public and private [17].

According to Glaucoma is one of the leading causes of irreversible vision loss worldwide, and the number of cases is continuously rising, according to a study by Lauren J. et al. Early detection is essential because it enables prompt intervention, which can stop additional loss of visual field. The evaluation of the optic cup and disc borders is at the core of a fundus imaging examination of the optic nerve head, which can be used to identify glaucoma. Although fundus imaging is expensive and noninvasive, picture analysis depends on expensive, time-consuming, and subjective expert evaluations [18]. Using fundus photos, Ajitha S. et al. presented a precise deep learning-based method for automatic glaucoma identification. An 1113-image fundus dataset from four databases was used to train a 13-layer CNN model. The 70:20:10 split is used for training, validation, and testing the model, correspondingly. The study showed how deep learning could help with early glaucoma detection. This model is compared using 2 classifiers: Softmax and Support Vector Machine (SVM). Among these classifiers SVM classifier achieved high accuracy of 95.61% which indicates high reliability in the detection of glaucomatous cases. However, it lacked external validation, used only a single dataset, and did not apply explainable AI techniques, making it difficult to interpret model decisions [19].

With a deep learning-based methodology, Shoukat A et al. concentrated on the early diagnosis of glaucoma in its early stages. Due to the shortcomings of the present manual assessment methods, glaucoma, a primary cause of permanent blindness, is frequently detected too late. The detection method involves identifying characteristics in retinal images that physicians frequently overlook. In order to generate a large collection of fundus images that may be used in a number of ways to train the CNN model, the suggested method uses fundus image gray channels and a processing technique. This method has been demonstrated to increase the glaucoma detection accuracy on the G1020, RIM-ONE, ORIGA, and DRISHTI-GS datasets using the ResNet-50 architecture. This model achieved an accuracy of 98.48% using this model on the G1020 dataset. The system is able to help hospital doctors diagnose early-stage glaucoma and provide early intervention [7].

Braganza et al. presented a public fundus image dataset to investigate glaucoma pattern identification with deep learning (DL). The Brazilian glaucoma labeling dataset contains 2,000 images taken from 1,000 volunteers, divided into two parts equally i.e., glaucoma and non-glaucoma. All images were acquired using a smartphone connected to an eye tracker. Additionally, a neural network segmentation model was trained with additional data to produce a DL approach for automated glaucoma identification. Results demonstrate that the suggested approach has a 90% accuracy rate in identifying glaucoma in fundus images. Thus, the total accuracy of visual glaucoma examination is positively aided by the combination of fundus photos taken with a smartphone, a portable panoptic ophthalmoscope, and an artificial intelligence system. Therefore, the suggested method may be useful for developing technologies for diagnosis [20].

Mamta Juneja, et al., had presented research focused on self-diagnosis of glaucoma through deep learning, particularly on convolutional neural networks (CNNs) for optic disc and cup segmentation of fundus images [21]. This work presents a retinal expert system that uses two CNN architectures to separately segment retinal discs and retinal cups, facilitating early retinal detection. Following testing on 50 fundus images, the model demonstrated exceptional accuracy, achieving 95.8% disc segmentation and 93% cup segmentation. Despite these promising findings, the study's small dataset size restricts the model's ability to generalize across a range of demographics. Future work could involve expanding the dataset, improving segmentation precision, and integrating other diagnostic parameters. The study also failed to prove the system's effectiveness in real-time clinical use and did not benchmark its performance against established clinical methods or expert evaluations [22]. Thisara Shyamali and Dulani report on a study on the identification of glaucoma by the segmentation and categorization of retinal fundus pictures utilizing a method based on deep learning computational model. This work focused on optic disc (OD) and optic cup (OC) using the U-Net attention model with three key CNN models: Inception-v3, VGG50, and ResNet50. Glaucoma is one of the leading causes of blindness, affecting about 65 million people globally. To enhance performance and reduce overfitting, the study employs a range of image preprocessing and data augmentation techniques. The models were tested

on the RIM-ONE dataset. The ResNet50-based attention U-Net achieved the best OD segmentation accuracy (99.58%), while Inception-v3 demonstrated remarkable performance in glaucoma classification (98.79%) [23].

Many academics are very interested in the automatic classification of fundus abnormalities for early diagnosis, which has been made possible by recent developments in artificial intelligence. In order to determine the cup-to-disc ratio (CDR), the study attempts to identify the margins of the optic disc and the optic cup in fundus images obtained from glaucoma patients. On a number of fundus datasets, we apply an enhanced U-Net model architecture and assess the model using segmentation measures. To improve the visibility of the optic disc and cup, we post-process the segmentation using edge detection and dilatation. The ORIGA, RIM-ONE v3, REFUGE, and Drishti-GS datasets serve as the foundation for our model's output. Our results show that the segmentation efficiency of our method is encouraging for CDR analysis [24].

Kim S.J. et al. set out to create machine learning models for glaucoma diagnosis that were both interpretable and had a high prediction power based on the thickness of the retinal nerve fiber layer (RNFL) and visual field (VF). Based on the assessment of the VF and the retinal nerve fiber layer RNFL thickness, a number of choices were found. Additionally, they created synthesized features from the original features. After that, they evaluated the features and chose the best ones for classification (diagnosis). 100 data cases were used for testing, while 399 data cases were used for validation and training. They looked at five machine learning techniques to create the glaucoma prediction model: k-nearest neighbor (KNN), random forest (RF), C5.0, and support vector machines (SVM). Using the training dataset, they repeatedly created a learning model, which was then assessed using the validation dataset. Ultimately, the learning model with the highest validation accuracy was determined to be the best. They used a range of metrics to assess the models' quality. The random forest model outperforms the SVM, C5.0, KNN models, despite their similar accuracy. The classification accuracy, specificity, sensitivity, AUC of the random forest model are, in that order, 0.98, 0.983, and 0.979. High levels of accuracy, specificity, sensitivity, and AUC are demonstrated by the built prediction models for distinguishing glaucoma from healthy eyes. Using

unidentified test results, it will be used to predict glaucoma. By using the prediction results, clinicians can make better decisions. Several learning models may be combined to improve prediction accuracy. Predictive decision rules are part of the C5.0 model. It can be applied to provide justifications for particular forecasts [25].

To avoid vision loss from glaucoma, early detection and appropriate screening are required. In recent years, glaucoma in color fundus images has been successfully detected automatically using convolutional neural networks (CNNs). When it comes to directly extracting distinguishing traits from fundus images, CNNs outperform the current automatic screening techniques. To distinguish between glaucomatous and normal fundus images, the researchers' CNN-based deep learning method was put out. To extract the discriminative features from the fundus images, an 18-layer CNN is built and trained. Four convolutional layers, two max pooling layers, and one fully connected layer make up this system. A two-phase tuning method is advised to ascertain the ideal batch size and initial learning rate. The proposed network is tested using the extensive attention-based glaucoma (LAG) databases DRISHTI-GS1, ACRIMA, ORIGA, and RIM-ONE2 (version 2). The dataset is expanded using the rotation data augmentation technique. Thirty percent of the randomly chosen images are used for testing, and seventy percent are used to train the model. The RIM-ONE2, ORIGA, ACRIMA, LAG and DRISHTI-GS1 databases yield an overall accuracy of 86.62%, 94.43%, 78.32%, 96.64% and 85.97% respectively [26]. For the ACRIMA database, the suggested approach has obtained 97.74% precision, 96.07% sensitivity, 97.39% specificity, and 96.64% accuracy. Compared to other existing architectures, the proposed method is more robust against Gaussian and salt-and-pepper noise [21][27].

Chai et al. (2018) proposed a deep learning-based model for glaucoma diagnosis that incorporates both hidden features learned from data and domain knowledge related to glaucoma. The model utilized a hybrid approach, combining CNNs with traditional knowledge-based features, such as clinical indicators, to enhance diagnosis accuracy. The study achieved high accuracy and sensitivity, demonstrating the value of integrating domain expertise with deep learning techniques. However, the approach was not evaluated on external datasets or real-world clinical environments, and the interpretability of the model's decision-making

procedure was not explored, limiting its practical application [28].

3. ABOUT DATASETS

In this study, the dataset utilized was sourced from Kaggle (<https://www.kaggle.com/datasets/hindsaud/datasets-higancnn-glaucoma-detection>). This publicly accessible dataset, which was especially selected for glaucoma classification tasks, includes a set of retinal fundus photos divided into two groups: normal and glaucoma. The images are provided in high resolution and are a valuable resource for training and evaluating deep learning models for automated glaucoma detection. The overall dataset consists of 30000 retinal images. Figure 2 shows a sample image of each type taken from each dataset. Below is a description of each subset of the datasets included in this set:

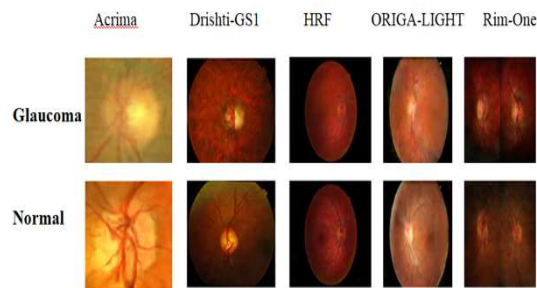


Figure 2: Glaucoma and Normal images taken from all five sub-Datasets

The Acrima dataset is a publicly available dataset that provides labeled retina fundus images for the task of glaucoma classification. This included glaucoma classifications that did not have the same glaucoma group. These images are high-resolution and provide a clear picture of the optic disc, making them ideal for training experts to identify anatomical abnormalities that cause glaucoma. The Acrima dataset is well known for its image quality and label accuracy, and is often used to evaluate segmentation methods in the diagnosis of eye diseases.

The Drishti-GS1 dataset is a well-known example of optic disc and cup segmentation and glaucoma detection. Ophthalmologists have classified it based on the accuracy of optic disc and cup alignment, as well as scores based on glaucoma diagnostic labels. These datasets are particularly useful for tasks that require segmentation-based analysis, such as calculating the cup-to-disc ratio (CDR), an important parameter in glaucoma diagnosis. The combination of segmentation masks

and diagnostic labels allows Drishti-GS1 to effectively apply supervised learning models to classification and segmentation tasks. Figure 3 shows the optic cup and optic disc in each variant of fundus image. The inner circular area represents the optical up and the outer circle represents the optical disc.

The HRF dataset contains high-resolution color images of the eye, primarily for the detection and diagnosis of retinal diseases, including glaucoma. The glaucoma subset in HRF contains images with expert annotations, which are necessary for a comprehensive model of retinal conditions. Due to the high resolution of HRF images, subtle artifacts can be removed, making them suitable for detailed examination of the optic nerve head which is useful in the diagnosis of glaucoma.

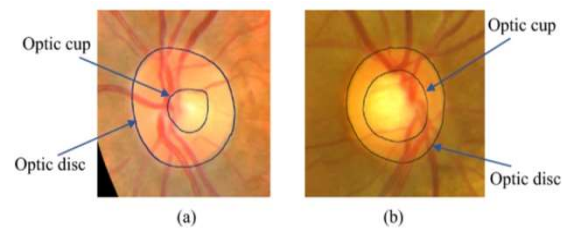


Figure 3: (a) Normal fundus image (b) Glaucoma image

The RIM-One (Retinal Images for Optic Nerve Evaluation) dataset is the largest dataset developed for glaucoma diagnosis using visual examination. It includes fundus images with associated clinical labels and, in some cases includes optic disc/cup annotations. RIM-One is composed of several versions (RIM-One r1, r2, r3) and datasets used for this purpose typically contain images labeled as glaucomatous or normal. This technique has been recognized as a standard of care worldwide and is widely used in competitions and research studies focused on automated glaucoma screening.

The ORIGA-LIGHT dataset was obtained from the ORIGA(Online Retinal Image Collection for Glaucoma Analysis) collection. It contains large number of fundus images with annotations for both optic disc and cup boundaries and glaucoma labels. ORIGA-LIGHT is best suited for supervised deep learning models where classification and segmentation are important. The images are stored in controlled environments by professionals, which increases reliability. This dataset is expected to play a major role in training models to understand the

structural and morphological changes in the optic disc and cup related to glaucoma.

4. METHODOLOGY

This research presents a novel hybrid model that combines the advantages of transformer-based self-attention mechanisms with convolutional neural networks (CNNs) for glaucoma detection from fundus images. A composite dataset that was covered in the previous part is used to train the model. The collection's fundus photos are labeled as either glaucoma or normal. Among the preprocessing techniques include reading and converting images to RGB format, scaling all images to 224 by 224 pixels, and normalizing pixel values using ImageNet mean and standard deviation. The dataset was split into training, validation, and test sets with relative ratios of 70%, 15%, and 15%. The proposed model employs CNN-based EfficientNetB0 for feature extraction and a transformer encoder for global attention.

4.1 EfficientNetB0

EfficientNetB0 is a lightweight model designed to achieve high performance while using fewer computational resources. It is mainly composed of three key components: the stem, body, and head. The stem is responsible for the initial processing of the input image by applying a 3×3 convolutional layer with 32 filters and a stride of 2, which also reduces the spatial dimensions. This operation is followed by batch normalization and the ReLU6 activation function to introduce non-linearity and improve training stability. After the stem, the body handles feature extraction through a series of specialized blocks known as Mobile Inverted Bottleneck Convolutions (MBConv). Each MBConv block has specific configurations, including an expansion ratio, kernel size, and stride. The expansion step first increases the number of channels using a 1×1 convolution, mathematically represented as shown in equation 1:

$$X_{\text{expanded}} = \text{ReLU6}(XW_{\text{exp}}) \quad (1)$$

where X is the input, W_{exp} are the expansion weights, and ReLU6 is the activation function applied after expansion. Following expansion, depth wise convolution is performed, applying a single convolutional filter per input channel,

reducing computational complexity. This operation can be expressed as shown in equation (2):

$$Y_c(i,j) = (K_c * X_c)(i,j) \quad (2)$$

where X_c is the channel's input and K_c is the channel's kernel. Squeeze-and-Excitation (SE) blocks, which introduce channel-wise attention to highlight important characteristics and suppress less relevant ones, are also incorporated into each MBConv block. Equation (3) illustrates the two completely connected layers and global average pooling that are used in this method.

$$Z = \text{GAP}(X)$$

$$S = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z)) \quad (3)$$

where σ is the sigmoid function, W_1 and W_2 are the weights of the fully connected layers, and z is the feature vector obtained after global average pooling. Moreover, Swish activation is sometimes used to further enhance learning, defined as shown in equation (4):

$$\text{Swish}(x) = x \cdot \sigma(x) \quad (4)$$

where the sigmoid activation function is indicated by $\sigma(x)$. Each MBConv block also includes a residual connection to facilitate better gradient flow during backpropagation, improving training dynamics. The head of EfficientNetB0 is responsible for final feature processing and classification. A global average pooling layer reduces each feature map to a single value after a 1x1 convolutional layer projects the features to the desired dimension. The class probabilities are then output using a softmax function in conjunction with a fully connected layer. Compound scaling, a unique feature of EfficientNetB0, optimizes accuracy and efficiency by equally scaling the model's depth (number of layers), breadth (number of channels), and input resolution using a single compound coefficient. In the proposed architecture, EfficientNetB0 serves as a feature extractor, efficiently capturing rich representations from the input images for downstream glaucoma detection tasks. Figure 4 illustrates the layers used to construct EfficientNetB0 for feature extraction.

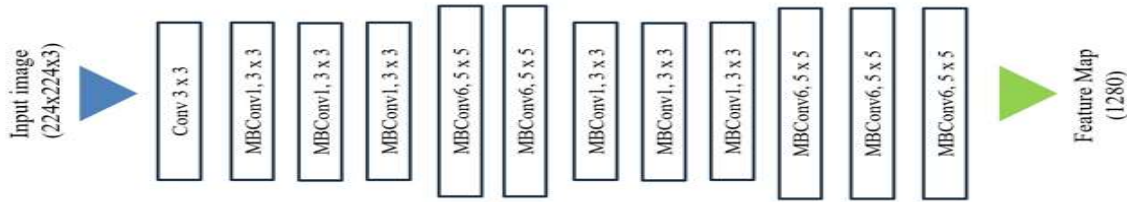


Figure 4: Architecture of EffecientNetB0 Feature Extractor

4.2 Transformer Model

In the proposed architecture, there is another model transformer encoder, which is used to learn global relationships between the sequence of the input features, which are the output of EfficientNetB0. Here, The model can focus on several aspects of the input sequence simultaneously thanks to the multi-head self-attention mechanism, which captures many dependencies and interactions. Attention scores between every pair of locations in the input sequence are calculated by the process. The model uses several attention heads to capture distinct features from separate representation subspaces. The representations obtained from the self-attention mechanism are further transformed by the other method, Position-wise Feed-Forward Network (FFN), which adds nonlinearity. This layer ensures consistency between tokens by applying two linear transformations with a ReLU activation in between at every place in the sequence. The next layer of residual connections assists gradient flow and prevents issues like vanishing gradients and is implemented by adding the input of a sublayer to its output. Layer normalization stabilizes and improves the training by normalizing the outputs of sublayers; this was applied after the addition of residual connections. The transformer architecture lacks recurrence; it doesn't inherently capture the order of tokens, so to prevent this problem, positional encoding was introduced. This adds positional encoding to input embeddings to provide information about all tokens in a sequence.

Figure 5 explains the Transformer Encoder architecture used for feature refinement. The input features (B, 1, 1280), extracted by EfficientNetB0, are first combined with positional encoding to preserve spatial information. The model consists of two Each transformer encoder layer has eight heads for multi-head self-attention. Feed-forward networks and normalization layers come next. By enabling the model to concentrate on significant portions of the input features, the self-attention method enhances representation learning. The final output features maintain the same dimension (B, 1,

1280), which are then used for downstream tasks such as classification.

4.3 Workflow of Proposed Model

Data Loading: collects images from 5 sub-datasets in the main dataset, with labels and dataset source. Creates balanced train, validation, and test sets by label and source The dataset was loaded for preparation and label encoding, which holds image paths and labels.

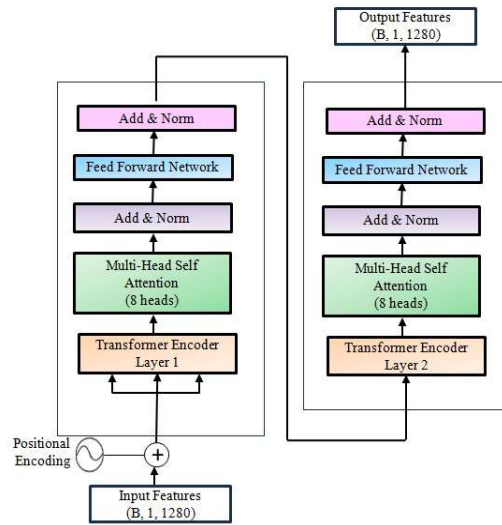


Figure 5: Architecture of the Transformer Model

Here, label encoding is used for categorical string labels like glaucoma and normal, which is needed for cross-entropy loss. The dataset loads the images using PIL and applies optional transformations, which return the tensor images and their encoded labels. In image transformation, it resizes the image to 224x224 for consistency with efficientnetb0 input and normalizes using imagenet stats. pretrained EfficientNetB0 only extracts the features from the image and produces the output of a 1280-dimensional feature vector per image. Transformer Encoder takes the input, i.e., a sequence of CNN features, and performs the 2-layer transformer encoder with 8 attention heads.

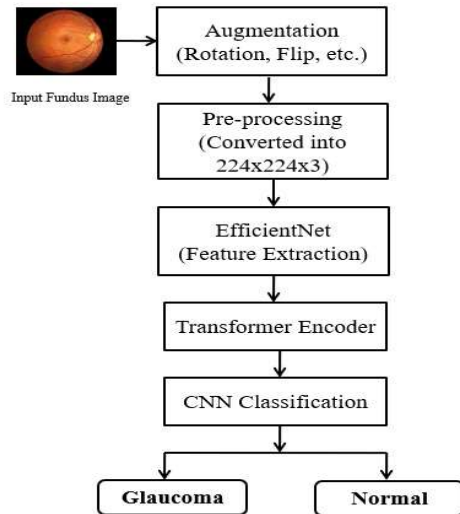


Figure 6: Workflow of the Proposed Model

After the transformer, the classification head performs average pooling that uses some variants. Using a loss function, optimizer (Adam), and device (GPU), the model was trained, and loss was monitored over epochs. Figure 6 represents the workflow of the proposed hybrid model.

5. RESULTS

The performance of the proposed hybrid model with curated dataset has been evaluated in terms of training and validation loss, accuracy, and confusion matrix analysis. Both training and validation losses are near to zero and also close to each other, suggesting that the model is neither overfitting nor underfitting. Since both losses are very small and nearly overlap at most points, it implies that the model has learned the data distribution effectively without memorizing it and also achieved high predictive accuracy. The model is compared with various transfer learning models on the same composite dataset, where transfer learning models have failed to achieve good predictive accuracy. This highlights the continued efficacy of the hybrid model in capturing relevant glaucoma-specific features. These findings suggest that this model is reasonable for use in real-world settings with increased complexity and variability in different data.

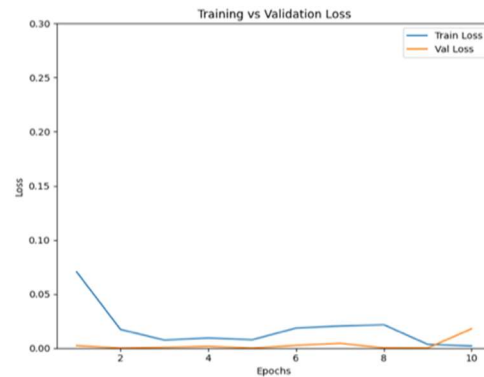


Figure 7: Training Vs Validation loss of Hybrid EffecientNet-Transformer

From Figure 7, it is clear that training and validation loss steadily decreased over the epochs. The training loss drastically dropped within the first few epochs and then gradually stabilized, reaching near to the zero value by the tenth epoch. Additionally, the validation loss was continuously low during the training phase, suggesting that the model did not significantly overfit and generalized well.

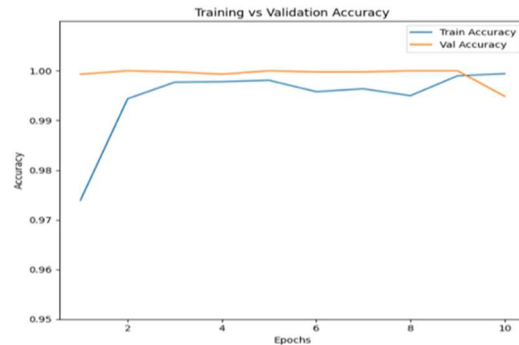


Figure 8: Training Vs Validation Accuracy of Hybrid EffecientNet-Transformer

The accuracy curves for training and validation of the Hybrid CNN-Transformer model are shown in Figure 8. During the first epochs, the model's training accuracy increased quickly, eventually reaching about 99.7%. The validation accuracy remained consistently high, around 99.7% across most epochs, suggesting strong generalization capability and minimal overfitting or underfitting. Figure 9 presents the confusion matrix for the Hybrid CNN-Transformer model's classification performance. The model accurately identified 2250 glaucoma cases and 2249 normal cases out of 4500 test samples, with only one misclassification. This corresponds to a very high classification accuracy, indicating the model's robustness in distinguishing between glaucoma and normal cases.

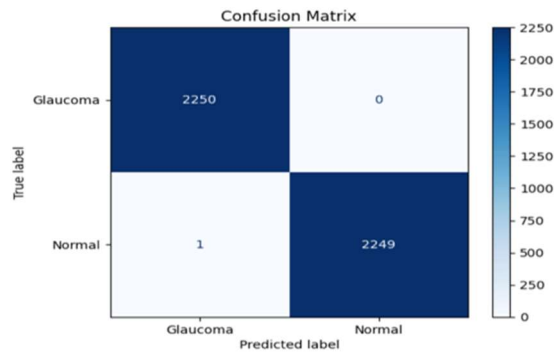


Figure 9: Confusion Matrix of Hybrid EfficientNet-Transformer model

The proposed model's performance is evaluated using accuracy, precision, recall, and F1 score, achieving 99.7% and more for all metrics. Table 1 shows the comparative analysis of various deep learning models implemented and evaluated for glaucoma detection, including ResNet50, VGG16, EfficientNet, SVM Classifier, and Custom-CNN. Transfer learning models VGG16 and ResNet50 demonstrated relatively lower performance on the curated dataset, achieving accuracies of 89% and 82.3%, respectively, compared to other models. The remaining models delivered strong results, with the SVM Classifier achieving an accuracy of 95.61% and the Custom-CNN reaching 91.32% accuracy. However, the proposed model, outperformed all other models, with 99.7% sensitivity, 99% specificity, 100% specificity, and 99% F1 score. These results demonstrate the ability of the proposed system to detect glaucoma with high sensitivity. The combination of EfficientNet as feature extractor and transformer encoder with CNN allows us to design more efficient algorithms, which makes it easier to solve the classification problem. Moreover, the hybrid design demonstrates potential for extension to other ophthalmic diseases, paving the way for scalable AI-powered diagnostic systems.

Table 1: Comparative Analysis of Various Models on curated dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
ResNet50	82.3	82	82	82
VGG16	89	89.7	92.6	91.1
EfficientNet	93.8	99	85.4	92.1
SVM Classifier	95.6	99	89.5	94.5
Custom CNN	91.3	92.2	91.3	91.3
Hybrid EfficientNet-Transformer	99.7	99.7	100	99.7

6. CONCLUSION

This study applied retinal fundus images, with deep learning models, particularly convolutional neural networks (CNNs) like VGG16 and ResNet, to detect glaucoma with a high level of accuracy (maximum of 99%). Datasets, such as ACRIMA, Drishti-GS1, HRF, RIM-One, and ORIGA-LIGHT enabled the models to generate high accuracy. The models reflected the ability to detect signs of glaucoma through analysis of the optic nerve, cup-to-disc ratio, among others. The use of deep learning models and traditional image process methods provided better performance for both image segmentation and feature extraction. The model, however, may not perform as well on smaller datasets, because it can have difficulty learning adequately. In this situation, a simpler or smaller CNN model might perform better. Also, the present model may miss some patterns that can cover the whole image. Using transformer-based models in the future may contribute to the system's ability to learn the larger image. In my view, this study demonstrates the ideal results, but testing of the model with real clinical-world data and populations should be done to see how well this works in practice. Including additional types of medical images such as OCT scans, in addition to fundus images, would increase accuracy and enhance the likelihood of detection at earlier stages. In short, combining new technology and many purposes of data sources will increase and improve the clinical reproducibility of the future glaucoma detection systems.

REFERENCES:

- [1] Vision. [(accessed on 16 July 2021)]. Available online: <https://my.clevelandclinic.org/health/articles/21204-vision>.
- [2] Vashist P, Senjam SS, Gupta V, Gupta N, Shamanna BR, Wadhvani M, et al. (2022) "Blindness and visual impairment and their causes in India: Results of a nationally representative survey." PLoS ONE 17(7): e0271736. <https://doi.org/10.1371/journal.pone.0271736>.
- [3] Guangzhou An, Hideo Yokota, et al.: (2019) "Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images." Hindawi Journal of Healthcare Engineering, Volume 2019, Article ID 4061313, 9 pages <https://doi.org/10.1155/2019/4061313>.
- [4] Mona Ashtari-Majlan, Mohammad Mahdi Dehshibi, and David Masip (2024) "Glaucoma Diagnosis in the Era of Deep Learning: A

- Survey." *Expert Systems With Applications* 256 (2024) 124888.
- [5] Jisy N K, Md. Hasnat Ali, SirishaSenthil& M.B. Srinivas (2024) Early detection of glaucoma: feature visualization with a deep convolutional network, *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 12:1, 2350508, DOI: 10.1080/21681163.2024.2350508.
- [6] RutujaShinde (2021) "Glaucoma detection in retinal fundus images using U-Net and supervised machine learning algorithms." *Intelligence-Based Medicine* 5 (2021) 100038, <https://doi.org/10.1016/j.ibmed.2021.100038>.
- [7] Shoukat A, Akbar S, Hassan SA, Iqbal S, Mehmood A, Ilyas QM. Automatic Diagnosis of Glaucoma from Retinal Images Using Deep Learning Approach. *Diagnostics (Basel)*. 2023 May 14;13(10):1738. doi: 10.3390/diagnostics13101738. PMID: 37238222; PMCID: PMC10217711.
- [8] Saxena A., Vyas A., Parashar L., Singh U. A Glaucoma Detection using Convolutional Neural Network; *Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*; Coimbatore, India. 2–4 July 2020; pp. 815–820.
- [9] Thakoor K.A., Li X., Tsamis E., Sajda P., Hood D.C. Enhancing the Accuracy of Glaucoma Detection from OCT Probability Maps using Convolutional Neural Networks; *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; Berlin, Germany. 23–27 July 2019; pp. 2036–2040.
- [10] Maheshwari S., Kanhangad V., Pachori R.B. CNN-based approach for glaucoma diagnosis using transfer learning and LBP-based data augmentation. *arXiv*. 20202002.08013.
- [11] Rehman A., Khan M.A., Saba T., Mehmood Z., Tariq U., Ayesha N. Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture. *Microsc. Res. Technol.* 2021;84:133–149. doi: 10.1002/jemt.23597.
- [12] Saeed J., Zeebaree S. Skin lesion classification based on deep convolutional neural networks architectures. *J. Appl. Sci. Technol. Trends*. 2021;2:41–51. doi: 10.38094/jastt20189.
- [13] Sobocki P., Józwiak R., Sklinda K., Przelaskowski A. Effect of domain knowledge encoding in CNN model architecture—A prostate cancer study using mpMRI images. *PeerJ*. 2021;9:11006–11021. doi: 10.7717/peerj.11006.
- [14] Medeiros F.A., Jammal A.A., Mariottoni E.B. Detection of progressive glaucomatous optic nerve damage on fundus photographs with deep learning. *Ophthalmology*. 2021;128:383–392. doi:10.1016/j.ophtha.2020.07.045.
- [15] D. D. Patil, R. R. Manza, G. C. Bedke, and D. D. Rathod, "Development of primary glaucoma classification technique using optic cup disc ratio," in 2015 International Conference on Pervasive Computing (ICPC), 2015, pp. 1–5.
- [16] M. Al Ghamdi, M. Li, M. Abdel-Mottaleb, and M. AbouShousha, "Semi-supervised transfer learning for convolutional neural networks for glaucoma detection," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3812–3816.
- [17] A. Diaz-Pinto, A. Colomer, V. Naranjo, S. Morales, Y. Xu, and A. F. Frangi, "Retinal image synthesis and semisupervised learning for glaucoma assessment," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2211–2218, 2019.
- [18] Lauren J. Coan, Bryan M. Williams, Venkatesh Krishna Adithya, Swati Upadhyaya, AlaAlkafri, SilvesterCzanner, RengarajVenkatesh, Colin E. Willoughby, Srinivasan Kavitha, Gabriela Czanner, Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review, *Survey of Ophthalmology*, Volume 68, Issue 1,2023, Pages 17-41, ISSN 0039-6257, <https://doi.org/10.1016/j.survophthal.2022.08.05>.
- [19] Ajitha S, Akkara JD, Judy MV. Identification of glaucoma from fundus images using deep learning techniques. *Indian J Ophthalmol*. 2021 Oct;69(10):2702-2709. doi: 10.4103/ijo.IJO_92_21. PMID: 34571619; PMCID: PMC8597466.
- [20] Bragança CP, Torres JM, Soares CPA, Macedo LO. Detection of Glaucoma on Fundus Images Using Deep Learning on a New Image Set Obtained with a Smartphone and Handheld Ophthalmoscope. *Healthcare (Basel)*. 2022 Nov 22;10(12):2345. doi: 10.3390/healthcare10122345. PMID: 36553869; PMCID: PMC9778370.
- [21] Pardhasaradhi, Mittapalli&Kande, Giri. (2016). Segmentation of optic disk and optic cup from digital fundus images for the assessment of glaucoma. *Biomedical Signal Processing and Control*. 24. 34-46. 10.1016/j.bspc.2015.09.003.

- [22] MamtaJuneja, Shaswat Singh, Naman Agarwal, Shivank Bali, Shubham Gupta, Niharika Thakur & Prashant Jindal “Automated detection of Glaucoma using deep learning convolution network” Volume 79, 1573-7721, <https://doi.org/10.1007/s11042-019-7460-4>.
- [23] ThisaraShyamalee and DulaniMeedeniya Glaucoma Detection with Retinal Fundus Images Using Segmentation and Classification November 2022Machine Intelligence Research 19(6):563-580 DOI:10.1007/s11633-022-1354-z.
- [24] J. Kim, L. Tran, E. Y. Chew and S. Antani, "Optic Disc and Cup Segmentation for Glaucoma Characterization Using Deep Learning," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 2019, pp. 489-494, doi: 10.1109/CBMS.2019.00100.
- [25] Kim S.J., Cho K.J., Oh S. Development of machine learning models for diagnosis of glaucoma. PLoS ONE. 2017;12:177726–177742. doi: 10.1371/journal.pone.0177726.
- [26] Singh, L.K., Pooja, Garg, H. et al. Deep learning system applicability for rapid glaucoma prediction from fundus images across various data sets. Evolving Systems 13, 807–836 (2022). <https://doi.org/10.1007/s12530-022-09426-4>
- [27] Elangovan P., Nath M.K. Glaucoma assessment from color fundus images using convolutional neural network. Int. J. Imaging Syst. Technol. 2021;31:955–971.doi: <https://doi.org/10.1002/ima.22494>.
- [28] Chai Y, Liu H, Xu J. Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models. Knowl Based Syst. 2018;161:147–56.