Automated Glaucoma Detection Using Vision and Swin Transformers: Advancing Ophthalmic AI

[1]D.Sakunthala, [2]Dr.N.Gireesh
[1]*Research Scholar, Mohanbabu University, Tirupati, India*
[2]*Professor, Mohanbabu University, Tirupati, India*

KEYWORDS

ABSTRACT

Purpose: Glaucoma is one of the most common causes of permanent blindness in the world; early detection and precise diagnosis are essential to successful treatment.Convolutional Neural Networks (CNNs) are one of the deep learning techniques that have shown excellent results in the processing of medical images. **Methodology:**Using a dataset of 1,650 fundus pictures from the REFUGE, ORIGA, and ACRIMA databases600 glaucoma-positive and 1,050 glaucoma-negative samplesthis study assesses the effectiveness of three cutting-edge deep learning models for glaucoma classification. Pretrained models like ResNet-50, VGG16, GoogLeNet, Vision Transformer (ViT), and Swin Transformer are investigated, emphasizing on their capacity to extract key variables such as the optic cup-to-disc ratio, retinal nerve fiber layer thickness, and vascular patterns.**Result:**Swin Transformer outperformed other models, achieving 100% accuracy, precision, recall, and F1-score in perfect classification. ViT and GoogLeNet similarly showed remarkable performance, achieving 92.92% and 91.54% accuracy. ResNet-50 and VGG16, on the other hand, had lower accuracy percentages of 74.62% and 78.77%. A number of drawbacks were found in all the models. ResNet-50 suffered from underfitting, which resulted in incorrect classifications and lower validation accuracy. While VGG16 was effective in standard image classification tasks, it showed inadequate recall and substantial validation loss, especially in situations when the patient had glaucoma. GoogLeNet struggled with overfitting, which limited its ability to generalise to new data, whereas ViT needed a lot of processing power and initially had trouble correctly categorising cases of glaucoma.On the other hand, Swin Transformer used hierarchical feature maps and changing windows to efficiently capture both local and global picture information, such as blood vessel patterns and the structure of the optic nerve head. A confusion matrix verified that there were no misclassifications in the model's flawless generalisation. **Conclusion:**In summary, the study found that Swin Transformer was the most dependable and resilient model for glaucoma diagnosis, even though models like as ViT and GoogLeNet shown potential. This study highlights the promise of transformer-based topologies, in particular Swin Transformer, as a state-of-the-art remedy for ophthalmology-related medical image classification challenges.

1. Introduction

One of the main causes of permanent blindness in the globe, glaucoma is brought on by gradual damage to the optic nerve. Early identification and treatment are critical to avoid visual loss. Nevertheless, glaucoma is frequently missed by conventional diagnostic techniques including visual field testing and intraocular pressure monitoring. Automated glaucoma identification has become a viable way to enable earlier and more precise diagnoses with the development of large-scale retinal imaging databases and deep learning algorithms.

Convolutional Neural Networks (CNNs) are one of the deep learning techniques that have shown excellent results in the processing of medical images. Popular models for a variety of tasks, including the diagnosis of retinal disorders, include ResNet-50, VGG16, and GoogLeNet. Each of these CNN models has its limits despite its success.

For image classification applications, ResNet-50 performs well by addressing the vanishing gradient issue in deep networks through the use of residual connections. On smaller medical datasets, however, it frequently overfits and necessitates a significant amount of processing power. Furthermore, its dependence on convolution processes, it is less able to capture long-range relationships in visuals, which is a crucial capability for glaucoma detection of subtle characteristics.

Smaller convolutional filters and deep layers are used by VGG16 architecture, which is simple and consistent in order to enhance feature extraction. However, because of its huge number of parameters, it has a slow training time and is computationally costly. Furthermore, its inability to handle vanishing

gradients internally may restrict its use on deeper levels. Its efficacy in identifying glaucoma is further impacted by its incapacity to represent intricate retinal structures with efficiency.

By applying convolutions of various sizes in parallel, the GoogLeNet produces Inception modules, which allow the network to collect multi-scale characteristics. Although this improves computing efficiency and lowers the possibility of overfitting, the intricacy of the model may make it more difficult to train and fine-tune for certain applications, such as the detection of glaucoma. Furthermore, the model may not be as successful when addressing global structural alterations in the optic nerve, which are crucial for the diagnosis of glaucoma, due to its dependence on local feature extraction.

Despite their great success in medical image analysis, CNN-based models are limited by their emphasis on local characteristics, which frequently ignores the overall context of the picture. This restriction may be crucial for the diagnosis of glaucoma, as precise diagnosis requires the collection of both local and global retinal characteristics. Transformer-based designs, such the ViT and Swin Transformer, have been developed recently to overcome this constraint. These architectures use self-attention methods to capture both local and global dependencies in images.

In order to function, ViT divides an image into patches and models the links between these patches using self-attention. Because of this, ViT is able to collect features more efficiently than CNNs, both local and global. But without extensive data augmentation or pre-training, its performance tends to deteriorate on smaller medical datasets, necessitating big datasets for training. Additionally, the model lacks inductive biases like translational invariance, which increases its reliance on large amounts of data and processing capacity.

Swin Transformer is a hierarchical Transformer model that uses a shifted window technique for hierarchical feature extraction and local attention to overcome some of the issues raised by ViT. As a result, the Swin Transformer can operate more effectively while preserving global context knowledge on datasets of any size. Nevertheless, its implementation might be computationally demanding and intricate, necessitating meticulous parameter adjustment to guarantee peak efficiency. Additionally, research on its effectiveness on medical pictures is currently ongoing, with limited relevance to the precise identification of glaucoma.

In this work, we evaluate the efficacy of ViT and Swin Transformer, two transformer-based models, against the standard CNN-based models ResNet-50, VGG16, and GoogLeNet for automated glaucoma identification. We want to ascertain the advantages and disadvantages of each model via this comparison study and ascertain the best architecture for precisely identifying glaucoma from retinal fundus pictures. The goal of this research is to better understand the trade-offs involved in these models in order to help build automated glaucoma screening methods that are more accurate and efficient. This might lead to the early identification and treatment of this crippling condition.

2. Literature Review

P. Jibhakate et al. present a discussion on the early identification of glaucoma. A comparative study is conducted between two distinct transfer learning algorithms ResNet-50 and VGG16. Machine learning techniques were utilised to quantify the morphological indicators of glaucoma, which are predictive of the onset of anomalies in the disease. The process of glaucoma screening is costly, labour-intensive, and prone to human error. There are fewer eye specialists in underdeveloped and underprivileged communities. By increasing public accessibility to glaucoma screening, this initiative will save time, money, and resources. Early diagnosis is crucial for glaucoma because it is the primary cause of blindness in the US.

In addition to aiding in the early diagnosis of glaucoma illness, the technique of A. Sallam et al. makes use of pre-trained models like as AlexNet, VGG11, VGG16, VGG19, GoogLeNet (Inception V1), ResNET-18, ResNET-50, ResNET-101, and ResNet-152. The Large-scale Attention based Glaucoma (LAG) dataset was used to assess the suggested approach. Using AlexNet, VGG11, VGG16, VGG19, GoogLeNet (Inception V1), ResNET-18, ResNET-50, ResNET-101, and ResNet-152 models, satisfactory results of 81.4%, 80%, 82.2%, 80.9%, 82.9%, 86.7%, 85.6%, 86.2%, and 86.9% were observed on the LAG dataset. The ResNet-152 model was determined to be the best among these outcomes, achieving a high accuracy with recall of 86.9% and precision of 86.9%.

This work, which was reported by A. Serener et al., is based on the automated identification of advanced and early glaucoma by the use of fundus photos. Transfer learning is used to train and optimise the deep convolutional neural network algorithms ResNet-50 and GoogLeNet for classification. It is proven that GoogLeNet model outperforms ResNet-50 for the identification of early as well as advanced glaucoma detection.

It's still unclear exactly colour fundus-based deep learning algorithms identify glaucoma. Consequently, D.-W. Lu et al. study examined the deep features retrieved by the deep learning models in order to provide the frameworks of the deep convolutional neural network (DCNN) in glaucoma evaluation a visual interpretability. 986 fundus photos from National Taiwan University Hospital Hsin-Chu Branch were used in the study. These photos were divided into two groups: 512 glaucomatous cases with impaired ganglion cell complex (GCC) and 474 non-glaucomatous cases with normal GCC thickness.According to the experimental findings, deep learning models primarily target the optic nerve head (ONH) regions for glaucoma diagnosis, in line with clinical guidelines for glaucoma evaluation. Remarkably, even with reduced pictures of the macular regions alone, the DCNN models are still able to achieve good prediction accuracy in identifying glaucomatous patients. The model's focal regions match the region with GCC impairment in a number of situations. The findings suggest that deep learning algorithms are capable of identifying morphologically intricate changes in fundus photos that may be difficult for professionals to visualise.

3. Methodology

This section explains about the methodology of various CNN's and the emerging transformer model in the application glaucoma detection.

3.1 ResNet-50

The architecture of ResNet-50 is displayed in Fig.3a which identifies the presence or absence of glaucoma in the given retinal fundus.

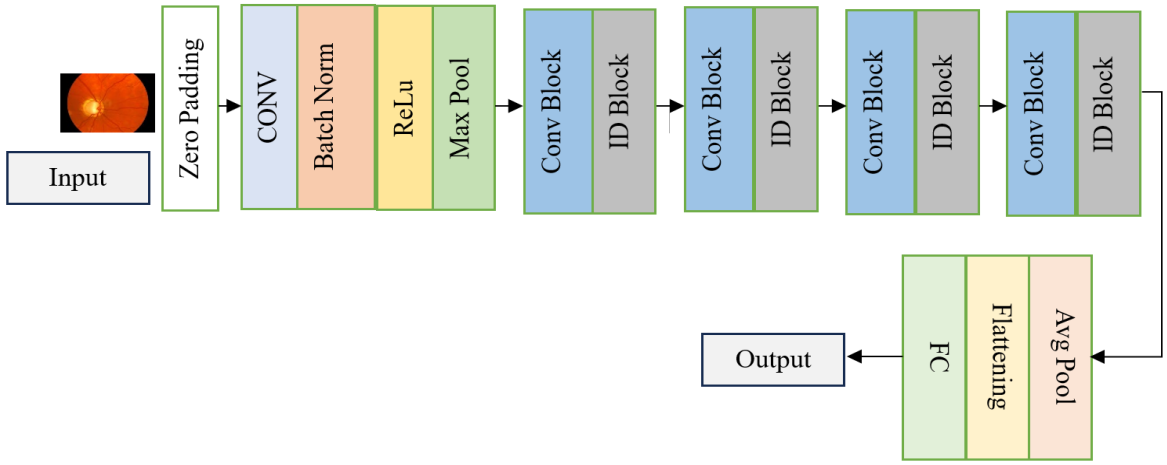


Fig. 3a – Block diagram of ResNet-50 for the detection of glaucoma

To efficiently diagnose glaucoma, the model examines pre-processed retinal fundus pictures, which are usually scaled to 224x224 pixels and normalised. These images include important features that can be detected, including the optic nerve head, cup-to-disc ratio, and structural anomalies.

Zero-padding is used to retain the spatial dimensions during convolution processes, which is essential for accurate detection since it preserves significant edge and boundary characteristics which is expressed as,

$$\text{Output size} = \frac{I+2P-K}{S} + 1 \text{----------(1)}$$

P is the size of the padding; S is the stride and K is the kernel size.

First convolutional layer extracts low-level features like as edges, textures, colour gradients, and simple patterns by using 3x3 filters, which are widely used in ResNet-50. The convolution operation at each layer is,

$$Y_{ij}^{k} = \sum_{m=1}^{M} \sum_{n=1}^{N} X_{i+m,j+n}\ W_{mn}^{k} + b^{k} \text{----------(2)}$$

$Y_{ij}^{k}$is the output feature map, $X_{i+m,j+n}$ is the input pixel value, W_{mn}^{k} is the kernel weight, $b^{k}$is the bias term, M and N are the height and width of the kernel.

After that, the output is normalised using Batch Normalisation, which minimises overfitting and speeds up training so that the model can more effectively generalise to new images and it is expressed as,

$$Y_{ij}^{k} = \frac{Y_{ij}^{k} - \mu^{k}}{(\sigma^{k2} + \in).^{\frac{1}{2}}} \text{----------(3)}$$

$\mu^{k}$is the mean value of the feature map, σ^{k2} is the value of variance, $\in$ is the numerical stability constant.

ReLU activation function provides non-linearity to the model, allowing it to learn intricate properties like nerve fibre thinning and optic disc deformationwhich is given as,

$$f(x) = \max(0,x) \text{----------(4)}$$

x is the input of the activation function which helps in detecting the complex patterns such as optic disc deformation and nerve fiber thinning.

After that, Max Pooling focusses on the most important data, such the optic cup and disc, by reducing the spatial dimensions of the feature maps, it is expressed as,

$$Y_{ij}^{k} = \max(X_{i:i+p,j:j+p}^{k}) \text{----------(5)}$$

$Y_{ij}^{k}$is the pooled output, $X_{i:i+p,j:j+p}^{k}$is the input patch.

Regarding feature extraction, the model encompasses a broad range of attributesEdges, textures, colour gradients, and patterns that delineate anatomical elements like blood arteries and the optic disc are examples of low-level characteristics. The cup-to-disc ratio, vascular alterations, and optic nerve head characteristics are the main mid-level traits that point to underlying diseases. Elevated characteristics include optic disc distortion, thinning of the nerve fibre layer, mild alterations in the retina, and general optic nerve shape and health, all of which are important markers for glaucoma diagnosis.

In summary, the first convolution operations identify low-level characteristics like edges and textures. Deeper convolution layers extract mid-level information such as optic nerve head properties and cup-to-disc ratio. Residual blocks and pooling layers record high-level information such as optic disc distortion and nerve fibre thinning.

By averaging all values, Average Pooling eventually reduces the final feature map to a single vector, highlighting global aspects like the general form and condition of the optic nerve. For classification, this output is subsequently flattened into a 1D vector, which is given as,

$$Y^{k} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} X_{ij}^{k} \text{----------(6)}$$

$Y^{k}$is the global average pooled value.

As the last classifier, the fully connected layers use the flattened feature vector to provide a probability score that indicates whether or not the picture is glaucomatous and it is expressed as,

$$z = W^{T} x + b \text{----------(7)}$$

W is the weight matrix, x is the flattened input vector, b is the bias term.

In order to give probabilities for each class and help the model determine whether glaucoma is present in the retinal images, the SoftMax function is commonly used in the output layer which is given as,

$$P(y=i \mid x) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \text{----------(8)}$$

P (y=i | x) is the probability of class I for the input x, $z_i$is the score value, C is the number of classes.

3.2 VGG16

The architecture of VGG16is displayed in Fig.3b which identifies the presence or absence of glaucoma in the given retinal fundus.

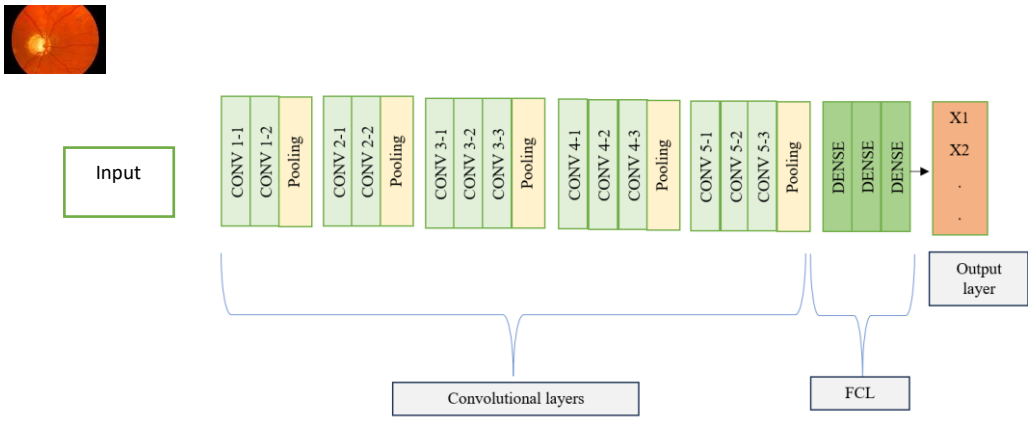


Fig. 3b – Block diagram of VGG16 for the detection of glaucoma

Pre-processed retinal fundus images, which are usually normalised and shrunk to a set size, such as 224x224 pixels, are processed using the VGG16-based architecture for glaucoma identification. In order to extract low-level data, including as edges, textures, and colour gradients, which aid in the identification of fundamental components like blood vessels and the optic disc, the model first applies convolutional layers. Convolution layers (Conv 1-1 and Conv 1-2) in the first block identify these fundamental characteristics, and a Max Pooling layer shrinks the feature map's spatial dimensions while preserving important details like the cup-to-disc ratio and the limits of the optic nerve.The same process is repeated across a number of convolutional blocks, with each new block gradually extracting increasingly complex and abstract characteristics.

Deeper convolutional layers concentrate on obtaining mid-level and high-level characteristics necessary for the diagnosis of glaucoma in blocks 2, 3, 4, and 5. Conv 2-1 and Conv 2-2, for example, concentrate on mid-level elements like the optic nerve head and cup-to-disc ratio, whereas Conv 3-1 through Conv 3-3 capture more intricate patterns like alterations in the optic disc structure and retinal nerve fibre layer. The model recognises high-level traits such as optic disc deformation, nerve fibre thinning, and retinal alterations associated with glaucoma as it moves through Conv 4-1 to Conv 4-3 and Conv 5-1 to Conv 5-3. Max Pooling layers are used to down sample the feature maps after each convolutional block, preserving only the most important features.

The output of the convolutional layers is flattened into a one-dimensional vector and sent to the fully connected (dense) layers after all relevant features have been retrieved and processed. The high-level characteristics are combined by these layers to create a thorough feature vector, which is then utilised to determine the classification. Using the learnt characteristics, the fully connected layers determine whether glaucoma is present in the image. Usually, a SoftMax activation function is used in the final output layer to assign probabilities to the various classes (glaucoma or non-glaucoma). The final categorisation is decided by the model based on the highest probability.

In summary, a variety of characteristics are extracted by the VGG16 architecture, ranging from high-level diagnostic indicators like optic disc deformation and nerve fibre thinning to low-level visual signals like edges and textures. The fully connected layers incorporate these characteristics for the final classification after the pooling layers minimise spatial dimensions while preserving important information. With the use of this organised, hierarchical feature extraction, the model can correctly identify glaucoma from retinal fundus photos.

3.3 GoogLeNet

The architecture of GoogLeNet is displayed in Fig.3c which identifies the presence or absence of glaucoma in the given retinal fundus.

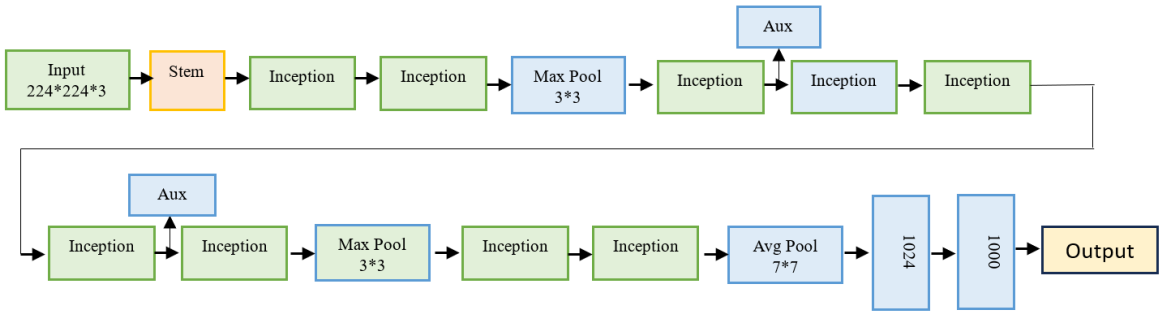


Fig. 3c – Block diagram of GoogLeNet for the detection of glaucoma

The first step in the GoogLeNet architecture for glaucoma diagnosis is to process retinal fundus images that have already been pre-processed and scaled to 224x224x3. Using simple convolutional techniques, the first stem layer extracts low-level information like as edges, textures, and patterns, concentrating on the borders of the optic disc and the retinal blood vessels. After then, the network as a whole uses Inception modules to collect data at various sizes. The model is able to learn both local and global characteristics that are essential for diagnosing glaucoma, such as the anatomy of the optic nerve and general health of the retina, by applying filters of size 3x3, simultaneously in each Inception module. As the image moves forward into the network, these modules extract low, mid, and high-level information. Basic retinal patterns are the focus of low-level features, whereas optic nerve thinning, nerve fibre layer flaws, and optic disc deformation are among the important diagnostic metrics captured by mid-level and high-level features, respectively.

After a number of Inception modules, max pooling procedures are used to minimise the spatial dimensions of feature maps while keeping important details about structural anomalies that aid in the diagnosis of glaucoma. Auxiliary Classifiers are included after specific Inception modules to help with training and avoid gradient vanishing. They provide early predictions on the existence of glaucoma and regularise the training process. The network's total performance is improved by these auxiliary outputs, which guarantee that the network picks up useful characteristics at intermediate stages.

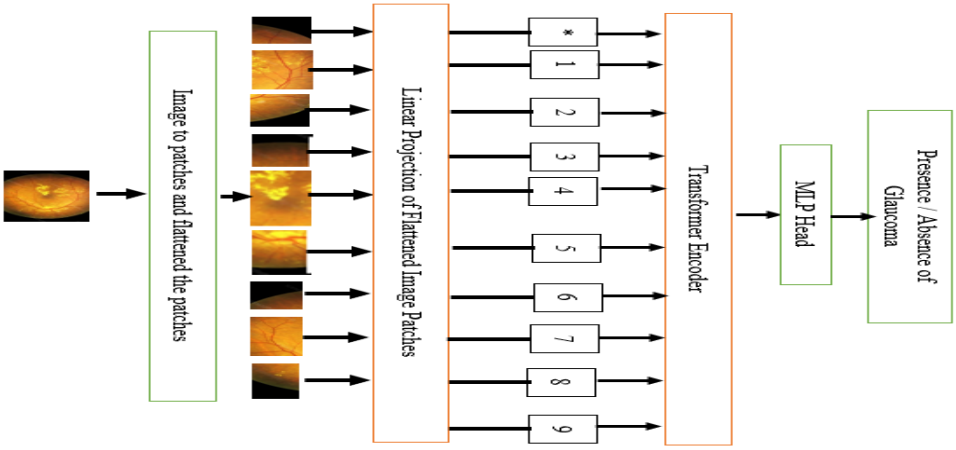
The network employs Average Pooling 7x7to reduce the size of the feature maps while focussing on the most important global features, such as the general shape and health of the optic nerve, after passing through the deeper Inception modules, which concentrate on high-level feature extraction like optic disc deformation and nerve fibre thinning. The final classification is achieved by aggregating high-level features in fully connected layers (FCL) using the reduced feature map. Whether or whether glaucoma is visible in the retinal fundus picture is determined by these totally linked layers. A SoftMax layer is used to create the final output. It offers a probability distribution across the potential classes (glaucoma or non-glaucoma), with the final classification determined by the class with the highest probability.

In conclusion, GoogLeNet can efficiently extract features from retinal images on several scales using Inception modules. This enables it to capture both local and global data. Both high-level diagnostic characteristics and low-level visual signals that are essential for glaucoma detection are gradually learnt by the architecture. In order to provide precise identification of glaucomatous anomalies, fully linked layers make the final choice based on learnt characteristics, with the assistance of auxiliary classifiers to optimise training.

3.4 Vision Transformer

The architecture of ViT[7] is displayed in Fig.3d which identifies the presence or absence of glaucoma in the given retinal fundus.

Fig. 3d – Block diagram of ViT for the detection of glaucoma



In order to identify glaucoma, the ViTarchitecture first processes retinal images, which are separated into smaller, non-overlapping patches. Like token embeddings in language models, each of these image patches gets flattened into a vector, enabling linear projection onto a fixed-dimensional representation. As transformers are permutation-invariant and cannot record the spatial connections between patches, positional embeddings are added to the patch embeddings to preserve the image's spatial structure.

Each patch $x_p \in R^{P^2 \, X \, C}$ is projected into a lower dimensional embedding space :

$Z_p = W_e .\text{flatten}(x_p) + b_e$ ----------(9)

b_e is the bias term, $W_e \in R^{D \, X \, C \, .P^2}$ is the learnable weight matrix, D is the embedding dimension space.

Followed by the above patches, positional encodings are added to retain the spatial structure, which is

$z_0^i = z_p^i + E_{pos}^i$; i=1,2,…N ----------(10)

The core element of this design is the transformer encoder, which is made up of several layers of feedforward and self-attentional neural networks. The model can focus on important regions of the retinal vision by capturing long-range relationships between image patches due to the self-attention process.

For every patch embedding, the self-attention value is calculated in terms of query, key and value which are represented by Q,K,V.

$Q = zW_q$; $K = zW_k$; $V = zW_v$----------(11)

W_q ,W_k , W_v€ R^{DXD} are learnable projection matrices. The attention scores between the patches are calculated as follows

$$\text{Attention}(Q,K,V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{dk}}\right) V \qquad \text{----------(12)}$$

where Q-Query, K-Key and V-Value,

$\left(\frac{QK^T}{\sqrt{dk}}\right)$ is the softmax function which gives probabilities out of scores

Additional emphasis gets placed on areas of the optic disc, blood vessels, or other retinal characteristics that may be indicative of glaucoma. This allows the model to extract characteristics like the cup-to-disc ratio, other minor anomalies associated with glaucoma, and the size, shape, and texture of the optic nerve head.

The model integrates these retrieved characteristics into higher-level representations. These comprise both global information from the entire retinal picture and local features from particular patches, so the model can identify both large-scale indications such as injured nerve fibres and small-scale anomalies such as structural alterations in the optic disc. These collected properties are further refined by the feedforward neural networks, which contribute to the creation of abstract representations that are essential for the categorisation of glaucoma.

The transformer encoder's output is sent to a Multi-Layer Perceptron (MLP) head after going through many stages of self-attention and feature refinement processing. After processing the encoded patches, the MLP generates a classification prediction that indicates whether or not glaucoma is present in the retinal picture. To sum up, this ViT-based architecture makes use of its capacity to learn complex patterns, concentrating on both local and global retinal aspects, to enable precise glaucoma diagnosis by utilising the image's delicate visual signals.

3.5 Swin Transformer

The architecture of Swin transformeris displayed in Fig.3e which identifies the presence or absence of glaucoma in the given retinal fundus.

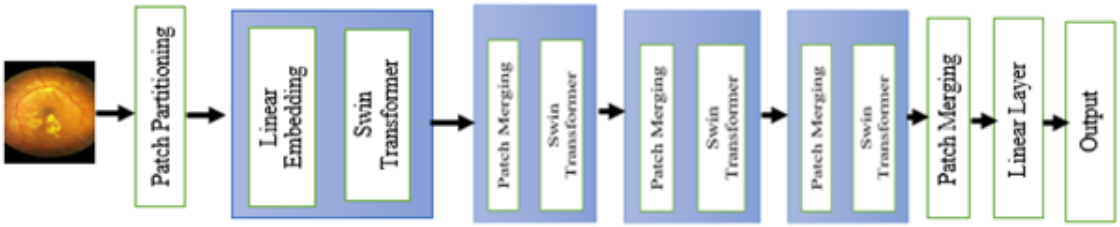


Fig. 3e – Block diagram of Swin transformer for the detection of glaucoma

Feature extraction is an essential stage in the Swin Transformer's glaucoma detection process, when the model recognises and learns the most significant visual patterns from the input fundus image. Unlike conventional CNNs, the Swin Transformer uses special methods for feature extraction.

First, the fundus image is split up into minuscule, fixed-size 4x 4 patches. The model can focus on local characteristics of the fundus images, such as alterations in the optic nerve or cupping of the optic disc, that are essential for identifying early indicators of glaucoma since it works with patches rather than individual pixels.

Instead of having a fixed patch size, an adaptive partitioning mechanism is used. The patch size varies depending on high contrast region especially around optic disc.

The number of patches is mathematically computed by adaptive patch partitioning as

$$\text{Adaptive patch count} = \sum_{i=1}^{N}\left(\frac{H}{p_i} x \frac{W}{p_i}\right) \quad \text{----------(13)}$$

Where H and W mentions the height and width of the image and p_i adapts to the content, varying per region. Each patch is then flattened into a vector of size $P^2 x C$, where C represents the number of channels. These vectors serve as initial input tokens for the Swin transformer.Patch partitioning effectively breaks down the fundus image into manageable sections, ensuring that the Swin Transformer can focus on smaller regions for focused feature extraction.

After being flattened into a 1D vector, each patch is then passed through multiscaleembeddings. This layer creates a higher-dimensional feature space both fine and coarse features from each patch's raw pixel data which is expressed as

$$Z_i = W_{emb,small}.x_i + W_{emb,large}.x_i + b_{emb} \text{----------(14)}$$

Where $W_{emb,small}$ and $W_{emb,large}$represents the learnable weight matrix which captures different scales of information which leads to richer patch representations,$b_{emb}$is the bias vector,Z_i is the embedded vector for each patch.Each patch has a feature vector as a consequence, which represents fundamental visual attributes like colour, texture, and intensity and acts as the basis for additional research.

The Swin Transformer uses window-based self-attention techniques after patches are implanted. The model learns the associations between patches in a given local area by performing self-attention inside each local window, which is a set of patches. The optic disc, blood vessels, and retinal layers are some of the key glaucoma-related components that the model may be able to identify in the fundus image due to this attention mechanism.

Here in both the swin transformer blocks, dynamic-attention mechanism is applied to the patches as that weights the attention score based on the important regions as optic nerve and it is defined as

$$\text{Attention (Q, K,V)} = \text{SoftMax}\left(\alpha_i.\frac{QK^T}{\sqrt{dk}}\right) V \quad \text{----------(15)}$$

where Q-Query, K-Key and V-Value,

$\left(\frac{QK^T}{\sqrt{dk}}\right)$ $is\ the\ softmax\ function\ which\ gives\ probabilities\ out\ of\ scoresand\alpha_i$is a learned weight that prioritizes medically relevant areas.

It may also identify elements like edges, forms, and textures.The model concentrates on several parts of the fundus picture at the same time due to the multi-head attention mechanism, which guarantees a more thorough knowledge of the local characteristics.

$$\text{MultiHead (Q,K,V)} = \text{Concat}(\text{head}_1, \text{head}_2 \ldots\ldots \text{head}_h)\, W_o \text{----------(16)}$$

$Head_i$ represents the attention score calculated using query, key and value.

The Swin Transformer modifies each local window's characteristics before moving them significantly to the next tier. This change makes it possible for patches from nearby areas to communicate with one another, guaranteeing that the model can accurately represent global dependencies across the fundus image. It helps in the extraction of broader, more global information from the fundus image, such as the general optic disc structure or the glaucoma-associated patterns of retinal thinning.The Swin Transformer extracts fine details and large-scale patterns from the fundus picture by switching between local (window-based) and global (shifting window) attention.

Three-tiered feature representations are learnt when the fundus picture moves through the Swin Transformer's levels. Early on, the model records low-level characteristics such as blood vessel architecture, optic disc boundaries, and edges and textures. Higher level and more abstract characteristics, such the optic cup's form or indications of nerve fibre layer loss, are derived from the deeper layers. The human visual system, which begins by recognising basic patterns and then advances to more sophisticated structures that are crucial for glaucoma diagnosis, is modelled by this hierarchical feature extraction method.

Hierarchical patch merging is performed here which means patch merging is performed in a hierarchical manner where patches are merged differently based on their level of information content. It provides more granularity in significant regions and less in other regions. As a result, the fundus image's spatial resolution decreases but its feature dimensionality rises. Hierarchical patch merging aids in gradually condensing the features while keeping the most significant information. This allows the model to concentrate on the characteristics that are most important for identifying glaucoma, including unusual retinal layer thinning or cupping of the optic disc.

This is as follows, Z^{l+1} = HierarchicalMerge ($W_{merge,high}.z1 + W_{merge,low}.z2$)

where z1 is important region and z2 is non-important region, $W_{merge,high}$ and $W_{merge,low}$ represents patches with varying information levels.

$$Z^{l+1} = W_{merge}.\text{Concat}(z_1^l, z_2^l, \ldots\ldots z_k^l) \text{----------(17)}$$

Z^l is the input feature map at layer l, z_i^l is a patch representation at layer l, W_{merge} is the learnable merging weight matrix.

The model strikes a compromise between computational efficiency and feature richness by shrinking the spatial size of the patches while enhancing the feature representations.

Following several layers of Swin Transformer processing and patch merging, the model generates a fundus image condensed feature representation. Rich information on regional and worldwide patterns associated with glaucoma may be found in these characteristics. A fully connected linear layer receives the final information and converts them into a format that could potentially use for classification (glaucoma positive or negative). This classification is formulated as

$$\acute{y} = \text{SoftMax}(W_{cls}. Z + b_{cls}) \text{----------(18)}$$

where W_{cls} is the weight matrix for the classification layer, b_{cls} is the bias term, z is the final feature vector.

A customized loss function is implemented that weights the classification loss based on the severity level.

$$L_{weighted} = \sum_{i=1}^{N} Weight\,(Severity_i).CrossEntropy(\tilde{y_i}, y_i) \text{----------(19)}$$

The severity weight emphasizes higher penalties for misclassification in severe cases, guiding the model to prioritize detection of critical stages.

4. Results and Discussions

This study comparesmany cutting-edge deep learning architectures, including as VGG16, GoogLeNet, ViT, and Swin Transformer, to classify images from the REFUGE, ORIGA, and ACRIMA datasets that are glaucoma-positive and glaucoma-negative. The models underwent five epochs of training, and their efficacy was assessed by the application of training and validation accuracy, as well as many evaluation indicators like precision, recall, F1 score, and AUC.

4.1 Database

In this work, REFUGE, ORIGA, and ACRIMA three publically available datasets are used which is shown in Table 1. These databases consist of fundus images, which are frequently utilised in medical contexts to diagnose glaucoma and other visual conditions.

Table 1 Database Description

S.No.	Database Name	Total number of images	Field of view(in degrees)
1	REFUGE	1,200 fundus images (600 glaucoma-positive and 600 glaucoma-negative images)	45 - 50
2	ORIGA	650 fundus images (200 glaucoma-positive and 450 glaucoma-negative images)	45
3	ACRIMA	400 fundus images (300 glaucoma-positive and 100 glaucoma-negative images)	45

The dataset is divided into training and validation sets in order to create a reliable and broadly applicable model. Twenty percent of the images (330 samples) were set aside for validation, and the remaining eighty percent (1,320 samples) were utilised to train the model. This division guarantees that the model's capacity to generalise is tested on untested data.

i. Training and Validation Accuracy

The models showed differing levels of effectiveness in identifying presence or absence of glaucoma which is shown in Table 2.

Table 2 Model-wise Training and Validation Accuracy Progression

No. of Epochs	Training Accuracy					No. of Epochs	Validation Accuracy				
	ResNet-50	VGG16	GoogLeNet	ViT	Swin Transformer		ResNet-50	VGG16	GoogLeNet	ViT	Swin Transformer
1	0.74	0.7	0.73	0.71	0.73	1	0.74	0.74	0.8	0.8	0.74
2	0.74	0.74	0.84	0.75	0.74	2	0.78	0.74	0.94	0.84	0.74
3	0.78	0.74	0.92	0.8	0.75	3	0.82	0.74	0.99	0.87	0.85
4	0.83	0.74	0.98	0.87	0.78	4	0.88	0.76	1	0.79	0.88
5	0.85	0.76	0.99	0.9	0.84	5	0.78	0.74	1	0.92	0.91

The Swin Transformer performs well when it applies to glaucoma detection. The Swin Transformer progressively develops more effectively attaining 0.84 training accuracy and 0.91 validation accuracy by the fifth epoch, whereas GoogLeNet reaches perfect validation accuracy (1.0) by the third epoch. The cup-to-disc ratio, optic nerve head form, retinal nerve fibre layer thickness, and other important glaucoma markers may all be captured by this model due to its shifting window technique and hierarchical attention mechanism. The Swin Transformer's ability to dynamically concentrate on various picture areas sets it apart from typical CNNs, enhancing its generalisation and making it an effective tool for medical imaging. While GoogLeNet works incredibly well, the transformer-based design of the Swin Transformer provides more reliable feature extraction for medical workloads by striking a balance between accuracy and efficiency without overfitting.

ii. Training and Validation loss

Table 3 provides the model-wise comparison of training and validation loss, showcasing how each model reduces error over time.

Table 3 Model-wise Training and Validation loss

No. of Epochs	Training loss					No. of Epochs	Validation Loss				
	ResNet-50	VGG16	GoogLeNet	ViT	Swin Transformer		ResNet-50	VGG16	GoogLeNet	ViT	Swin Transformer
1	0.57	0.62	0.54	0.6	0.58	1	0.51	0.57	0.39	0.5	0.53
2	0.53	0.56	0.37	0.5	0.56	2	0.47	0.6	0.21	0.4	0.49
3	0.5	0.55	0.22	0.4	0.51	3	0.38	0.55	0.08	0.4	0.36
4	0.42	0.53	0.11	0.3	0.47	4	0.29	0.53	0.02	0.4	0.38
5	0.31	0.52	0.05	0.3	0.39	5	0.46	0.5	0.01	0.3	0.27

The training and validation loss tables give additional insight into the models' performance. With its lowest training loss of 0.05 by the fifth epoch and validation loss of 0.01 by the same epoch, GoogLeNet demonstrates its exceptional error minimization and data generalization capabilities. Additionally, the Swin Transformer performs well, demonstrating strong generalisation and resilient learning skills by reducing its training loss to 0.39 and validation loss to 0.27 by the fifth epoch. The architecture of the Swin Transformer helps minimise loss by concentrating on important retinal properties like optic nerve shape and cup-to-disc ratio, which are essential for glaucoma detection. It does this by utilising hierarchical attention and shifting window processes.

Comparatively, VGG16 observes the slowest loss decline by the fifth epoch, validation loss is still unusually high at 0.50, suggesting that it has difficulty generalising as well as the other models. Swin Transformer is a great choice for medical image analysis tasks like glaucoma detection because it can balance lower loss with effective feature extraction and learning. ViT and ResNet-50 also demonstrate notable improvements, with ViT reaching a low validation loss of 0.30 by the end of training.

iii. Confusion Matrix

A confusion matrix in Fig. 4 is utilised to assess the extent to which the various models identify glaucoma. Correctly classified instances are represented by the diagonal members of the matrix, whereas incorrectly classified instances are represented by the off-diagonal elements.

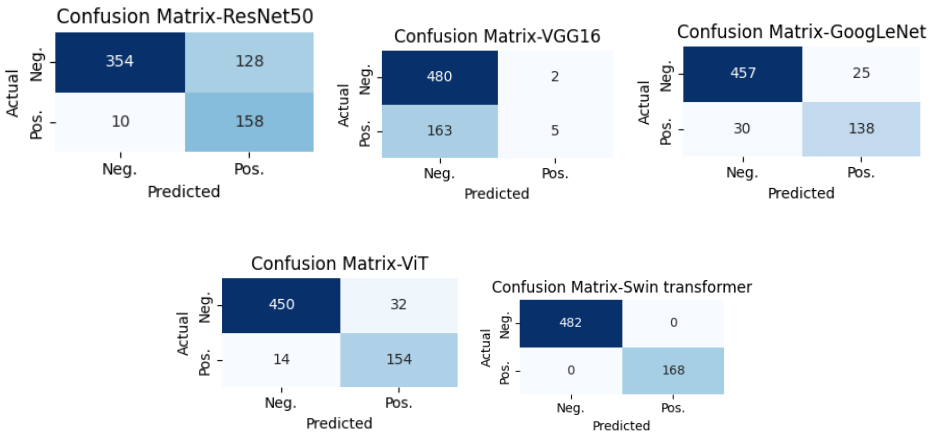
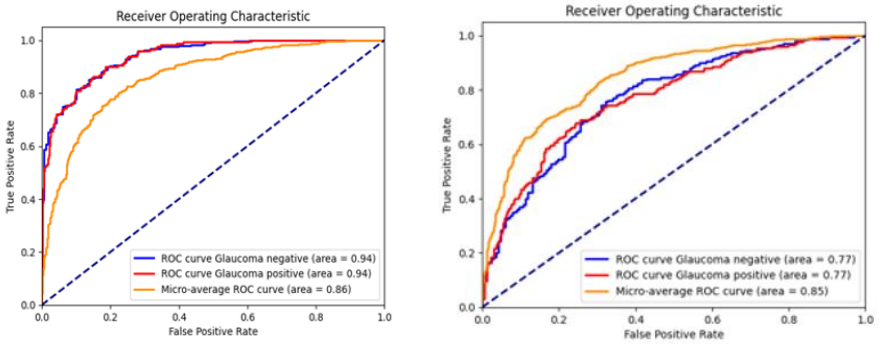


Fig. 4 – Confusion matrix of the models

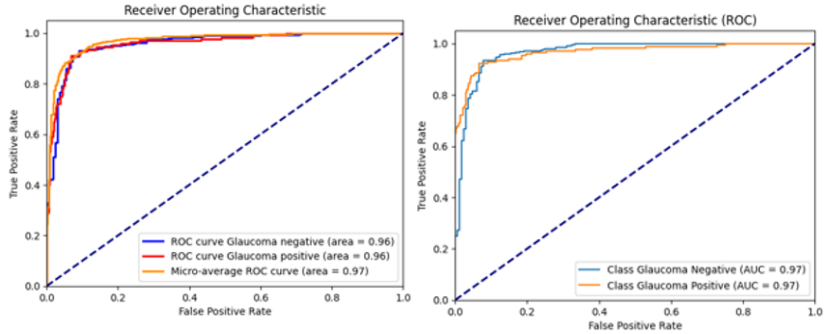
With zero false positives and false negatives and nearly flawless classification, the Swin Transformer model performs the best in glaucoma detection. Its ability to recognise minor symptoms of glaucoma is largely due to its ability to record both local and global dependencies within retinal pictures, which is made possible by its hierarchical feature representation and self-attention mechanism. On the other hand, because of their reliance on convolutional layers, which restricts their capacity to capture long-range relationships, models such as ResNet50 and GoogLeNet perform somewhat well. Despite having a high true negative rate, VGG16's simpler architecture makes it difficult to identify positive instances of glaucoma, which results in a high proportion of false negatives. Although it works effectively, ViT is sensitive to lesser datasets, which can lead to false positives. Overall, the Swin Transformer has a clear advantage over other models in correctly identifying glaucoma due to its capacity to process multi-scale input and concentrate on important retinal characteristics.

iv. Receiver Operating Characteristics curve (ROC)

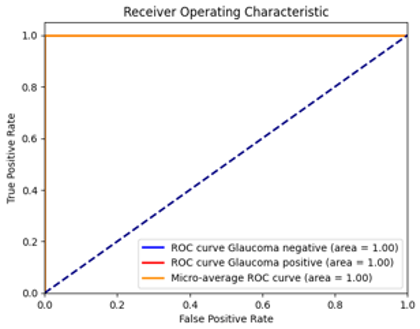
The five models' different ROC curves in Fig. 5 show the variance in detecting glaucoma.



a. ResNet50 b.VGG16



c.GoogLeNet d.ViT



e. Swin Transformer

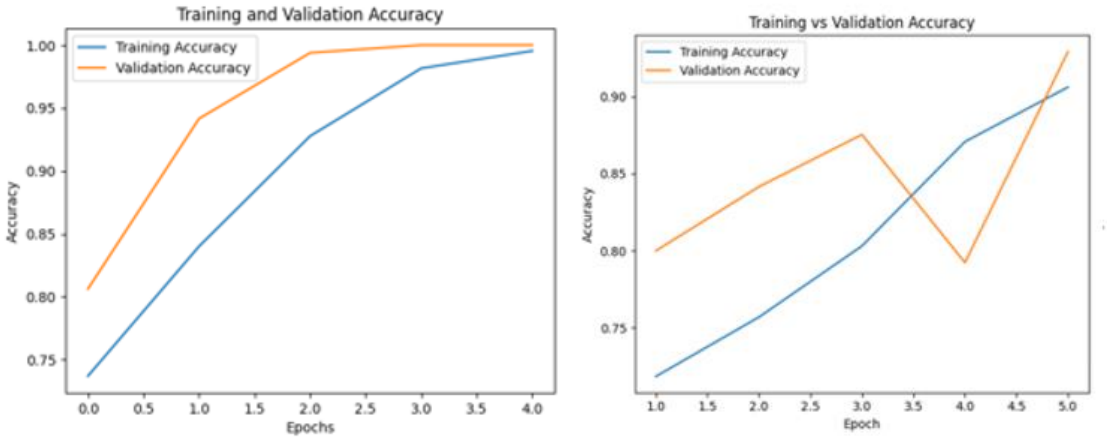
Fig. 5 – ROC of the models

Due to their deep architectures and effective multi-scale feature extraction, ResNet-50 and GoogLeNet perform well, however they mostly rely on local features. VGG16's restricted depth and feature extraction capabilities cause it to lag behind the others while having a simpler structure. ViT does rather well, but its inability to catch finer information is hampered by its patch-wise processing without hierarchical aggregation. The Swin Transformer, on the other hand, performs well and has the greatest AUC. It is the greatest model for identifying anomalies associated to glaucoma because of its hierarchical feature representation and self-attention processes, which enable it to successfully capture both local and global trends. This capacity to generalize better at multiple scales is what sets Swin Transformer apart from the other models.

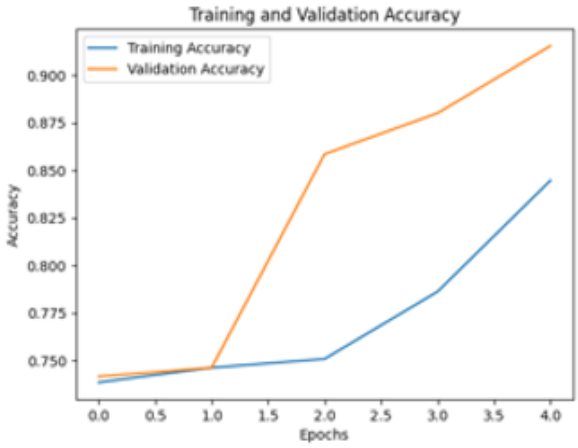
v. Training accuracy and validation accuracy

Fig. 6 shows the training and validation accuracy curves for five models. These curves show the amount that it learns from the training set and the way well it applies that learning to unseen data.

a. ResNet50 b.VGG16



c. GoogLeNet d. ViT



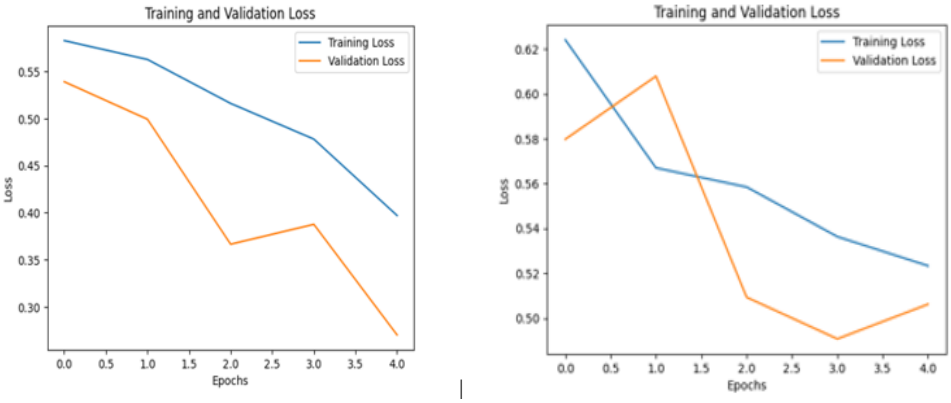
e. Swin Transformer

Fig. 6 – Training and validation accuracy of the models

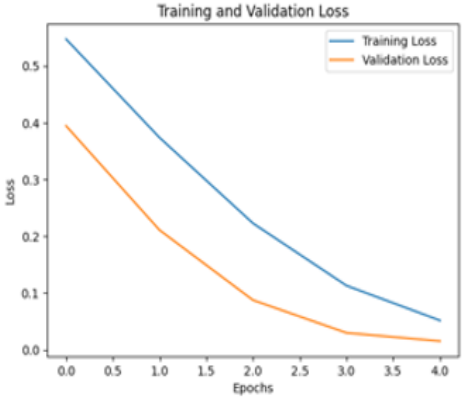
Strong learning and generalisation are suggested by the constant and consistent accuracy increases that GoogLeNet and ResNet-50 show over time, with steady increasing trends in both training and validation accuracy. On the other hand, VGG16 and ViT show varying accuracy, suggesting possible problems with learning stability and overfitting, in which case the models work well on training data but have trouble with generalisation. Swin Transformer successfully generalises to previously encountered data, as seen by a sharp spike in training accuracy and a gradual improvement in validation accuracy. Its self-attention technique and hierarchical feature extraction enable it to collect both global dependencies and local features in retinal images, contributing to its outstanding performance. This provides it an advantage over more conventional CNN-based models that concentrate more on local feature extraction, such as ResNet-50, GoogLeNet, and VGG16, as well as over ViT, which does not include hierarchical feature aggregation.

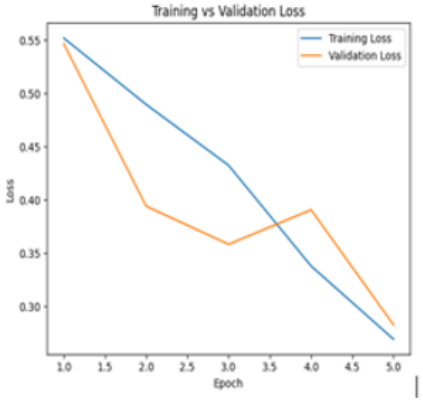
vi. Training accuracy and validation loss

The training and validation loss curves in Fig. 7 demonstrate the performance of the models.



a. ResNet50 b.VGG16





c. GoogLeNet d. ViT

e. Swin Transformer

Fig. 7 – Training and validation loss of the models

With the lowest validation loss among the others, Swin Transformer performs noticeably better than the others, suggesting better generalisation to untested data. This is mainly due to its hierarchical architecture, which enables it to concentrate on small details and more significant patterns in fundus pictures by efficiently capturing local and global information through shifting windows.The stated Convolutional models are limited in their capacity to comprehend complicated linkages in glaucoma detection because they rely on stacked convolutional layers, which are effective for extracting local features but ineffective at capturing long-range dependencies. Although GoogLeNet's inception modules enable multi-scale feature extraction, they only marginally improve performance when compared to transformers' self-attention processes.Higher validation loss results from ViT's use of picture patches rather than Swin Transformer's hierarchical feature extraction. Because of its hierarchical structure, Swin Transformer is more capable of tackling problems related to medical imaging, such as glaucoma diagnosis, where accurate feature localisation is crucial. It also allows for better spatial analysis and lowers computing complexity. Swin Transformer performs better due to the combination of global context modelling and local attentiveness.

5. Conclusion

For the purpose of classifying glaucoma, this study included an extensive evaluation of many advanced deep learning models, including ResNet-50, VGG16, GoogLeNet, ViT, and Swin Transformer.With 100% results in accuracy, precision, recall, and F1-score, among other important performance parameters, Swin Transformer proved to be the most successful of the models. Shifted window approaches enabled its hierarchical feature extraction, which improved the precision of capturing both local and global retinal data, including vascular patterns and optic nerve head shapes. Accordingly, Swin Transformer offers a great degree of generalisation without sacrificing classification performance, making it the ideal model for identifying subtle glaucoma signs in fundus. GoogLeNet and ViT were two models with great potential but faced challenges. ViT was resource-intensive and had trouble early on in the training process differentiating between glaucoma instances, whereas GoogLeNet struggled with overfitting, which limited its capacity to generalise to new data. Despite being commonly utilised in picture classification tasks, ResNet-50 and VGG16 performed

poorly in this application. ResNet-50 had underfitting, while VGG16 showed substantial validation loss, especially in more complicated glaucoma cases. The study's findings highlight the expanding significance of transformer-based designs in the categorisation of medical images. Compared to conventional CNN-based methods, Swin Transformer in particular showed a balance of accuracy, computational efficiency, and feature extraction capabilities. This demonstrates its potential for wider application in the field of ophthalmology, especially in automating the early diagnosis of glaucoma, which is one of the main causes of permanent blindness. Subsequent investigations may focus on improving these transformer models, maximising their computational effectiveness, and expanding their use to additional retinal conditions. This might result in better patient outcomes and more advanced diagnostic resources for medical practitioners.

References:

(Dewa et al. 2023), "Investigating Self-Attention in Swin-Unet Model for Disc and Cup Segmentation," *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, 2023, pp. 401-406, doi: 10.1109/ICITACEE58587.2023.10276855.

(Grover &Kapoor 2023), "Detection of Glaucoma and Diabetic Retinopathy Using Fundus Images and Deep Learning," *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Hamburg, Germany, 2023, pp. 407-412, doi: 10.1109/ICCCMLA58983.2023.10346704.

(Haouli et al. 2023), "Exploring Vision Transformers for Automated Glaucoma Disease Diagnosis in Fundus Images," *2023 International Conference on Decision Aid Sciences and Applications (DASA)*, Annaba, Algeria, 2023, pp. 520-524, doi: 10.1109/DASA59624.2023.10286714.

(Islam et al. 2022), "Deep Learning-Based Glaucoma Detection With Cropped Optic Cup and Disc and Blood Vessel Segmentation," in *IEEE Access*, vol. 10, pp. 2828-2841, 2022, doi: 10.1109/ACCESS.2021.3139160.

(Jibhakate et al. 2022), "Early Glaucoma Detection Using Machine Learning Algorithms of VGG-16 and Resnet-50," *2022 IEEE Region 10 Symposium (TENSYMP)*, Mumbai, India, 2022, pp. 1-5, doi: 10.1109/TENSYMP54529.2022.9864471.

(Lu *et al*.2022), "Visual Interpretability of Deep Learning Models in Glaucoma Detection Using Color Fundus Images," *2022 IET International Conference on Engineering Technologies and Applications (IET-ICETA)*, Changhua, Taiwan, 2022, pp. 1-2, doi: 10.1109/IET-ICETA56553.2022.9971605.

(Mallick et al. 2022), "Study of Different Transformer based Networks For Glaucoma Detection," *TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON)*, Hong Kong, Hong Kong, 2022, pp. 1-6, doi: 10.1109/TENCON55691.2022.9977730.

(Sallam *et al*. 2021), "Early Detection of Glaucoma using Transfer Learning from Pre-trained CNN Models," *2021 International Conference of Technology, Science and Administration (ICTSA)*, Taiz, Yemen, 2021, pp. 1-5, doi: 10.1109/ICTSA52017.2021.9406522.

(Serener&Serte 2019), "Transfer Learning for Early and Advanced Glaucoma Detection with Convolutional Neural Networks," *2019 Medical Technologies Congress (TIPTEKNO)*, Izmir, Turkey, 2019, pp. 1-4, doi: 10.1109/TIPTEKNO.2019.8894965.

(Tang et al. 2022), "A Deep Learning Approach for the Detection of Neovascularization in Fundus Images Using Transfer Learning," in *IEEE Access*, vol. 10, pp. 20247-20258, 2022, doi: 10.1109/ACCESS.2022.3151644.

(Virbukaitė et al. 2024), "Glaucoma Identification Using Convolutional Neural Networks Ensemble for Optic Disc and Cup Segmentation," in *IEEE Access*, vol. 12, pp. 82720-82729, 2024, doi: 10.1109/ACCESS.2024.3412185.