

Integrating Local and Global Feature Learning for Retinal Disease Detection: An EfficientNetB3-ViT Hybrid Model

Onur CEYLAN¹, Anıl UTKU^{2*}

^{1,2} Department of Computer Engineering, Faculty of Engineering, Munzur University, Tunceli, Türkiye
¹ onurceylan832@gmail.com, ² anilutku@munzur.edu.tr

(Geliş/Received: 03/02/2026;

Kabul/Accepted: 09/04/2026)

Abstract: Undiagnosed ocular conditions in initial phases can progressively impair vision, potentially leading to critical visual dysfunctions. Fundus imaging has enabled the identification of several retinal disorders, notably diabetic retinopathy, glaucoma, and age-related macular degeneration. However, manual evaluation of these images is both time-consuming and dependent on expert judgment. In this context, artificial intelligence-based automatic diagnosis systems have become an important need in the field of medical image analysis. In this study, deep learning-based models were compared for multi-class eye disease detection from fundus images, and a unique hybrid model was proposed. In this study, current deep learning architectures, including EfficientNetB3, DenseNet-121, ResNet-50, MobileNetV2, and ViT, were employed. All models were independently trained and tested, and their effectiveness was evaluated using commonly employed classification metrics. Additionally, a hybrid model based on EfficientNetB3 and ViT was designed, combining the strengths of Transformer and CNN architectures. All models were subjected to hyperparameter optimization using the GridSearch method, allowing for a fair comparison. The results obtained showed that the hybrid model successfully learned both local and global features in fundus images and exhibited higher performance compared to other models. In this context, the developed system has the potential to be integrated into future clinical decision-support systems.

Keywords: Ocular disease, deep learning, efficientnet, vision transformer.

Retina Hastalıklarının Tespiti için Yerel ve Küresel Özellik Öğreniminin Entegrasyonu: EfficientNetB3-ViT Hibrit Modeli

Öz: Başlangıç aşamasında teşhis edilemeyen göz rahatsızlıkları, görmeyi giderek bozarak kritik görme bozukluklarına yol açabilir. Fundus görüntüleme, özellikle diyabetik retinopati, glaukom ve yaşa bağlı makula dejenerasyonu gibi çeşitli retina hastalıklarının tanımlanmasını sağlamıştır. Bununla birlikte, bu görüntülerin manuel olarak değerlendirilmesi hem zaman alıcıdır hem de uzman görüşüne bağlıdır. Bu bağlamda, yapay zekâ tabanlı otomatik tanı sistemleri, tıbbi görüntü analizi alanında önemli bir ihtiyaç haline gelmiştir. Bu çalışmada, fundus görüntülerinden çok sınıflı göz hastalığı tespiti için derin öğrenme tabanlı modeller karşılaştırılmış ve yeni bir hibrit model önerilmiştir. Bu çalışmada, EfficientNetB3, DenseNet-121, ResNet-50, MobileNetV2 ve ViT dahil olmak üzere mevcut derin öğrenme mimarileri kullanılmıştır. Tüm modeller bağımsız olarak eğitilmiş ve test edilmiş ve etkinlikleri yaygın olarak kullanılan sınıflandırma metrikleri kullanılarak değerlendirilmiştir. Ek olarak, Transformer ve CNN mimarilerinin güçlü yönlerini birleştiren EfficientNetB3 ve ViT tabanlı bir hibrit model tasarlanmıştır. Tüm modeller, adil bir karşılaştırma sağlamak amacıyla GridSearch yöntemi kullanılarak hiperparametre optimizasyonuna tabi tutulmuştur. Elde edilen sonuçlar, hibrit modelin fundus görüntülerindeki hem yerel hem de küresel özellikleri başarıyla öğrendiğini ve diğer modellere kıyasla daha yüksek performans sergilediğini göstermiştir. Bu bağlamda, geliştirilen sistemin gelecekteki klinik karar destek sistemlerine entegre edilme potansiyeli bulunmaktadır.

Anahtar kelimeler: Göz hastalıkları, derin öğrenme, efficientnet, vision transformer.

1. Introduction

Ocular diseases have been a major health issue of concern across the world, and they affect a large number of people, with it being one of the leading causes of preventable vision impairment [1]. Given factors that influence the prevalence of eye disorders, such as the aging of the world population, the rising prevalence of diabetes, and increasing environmental risks, the prevalence of eye disorders has increased substantially [2]. The 2023 statistics from the World Health Organization indicate that approximately 2.2 billion people worldwide have experienced some form of vision impairment, with almost half of these cases avoidable through early detection and treatment [3].

* Sorumlu yazar: anilutku@munzur.edu.tr. Yazarların ORCID Numarası: ¹ 0009-0008-8692-8956, ² 0000-0002-7240-8713

The timely diagnosis is very significant in the retardation of the course of these diseases, enhancement of life quality, and decreasing the health system load [4]. Nonetheless, the diagnosis of eye diseases is one field that requires expertise, and professional ophthalmologists are not readily available, particularly in rural regions or developing nations [5]. In that regard, the creation of automatic image analysis tools assists in the rapid diagnosis process and eases the physician workload [6]. Over the last several years, the field of medical imaging has undergone substantial change with the introduction of artificial intelligence and the use of deep learning approaches [7]. CNNs have been successfully applied to a wide range of medical images, including fundus, OCT, MRI, and CT, for image classification [8].

Nevertheless, more recent deep learning architectures, such as the Transformer-based Vision Transformer (ViT) models, have gained popularity in the literature for their ability to learn long-range correlations in images [9]. The ViT global attention mechanism provides a more global view of the world than CNN local filters, resulting in better performance on specific tasks [10]. Nevertheless, in medical imaging, cases where Transformer architectures are insufficient are common. Especially in cases where limited data sets, similarities between classes, and intra-class variations are high, the local pattern extraction power of CNNs should be complemented by ViT's global context perception [11]. Based on this need, hybrid architectures have recently emerged as a focus in medical image processing studies.

Computer vision and deep learning-based approaches for automatic diagnosis of eye diseases have become an important research area in recent years. Fundus images provide a high-resolution and non-invasive source for detecting retinal deterioration at an early stage. Upon examining the literature in this field, it is observed that CNN and transfer learning-based models are predominantly used, while some recent studies have shifted to transformer architectures. The performance of these models varies depending on the scope of the dataset used, the class balance, and the model architecture. In the rest of this section, some prominent studies in the literature are summarized.

Emir and Colak compared VGG-16, InceptionV3, and ResNet-50 models for the classification of eye diseases. ODIR dataset was used for training. The dataset used consists of 8 classes. Data augmentation was applied with ImageNet to balance the class distribution. The highest accuracy of 97.1% was obtained with ResNet50 [12].

Li et al. created a hybrid deep learning model (BFPC-Net) for the fundus image classification. The ODIR dataset was used for training. In the study, channel and spatial attention mechanisms were integrated into a ResNet50-based architecture. The feature fusion of binocular images improved the global classification performance. 94.23% accuracy and 99.16% F1-score were achieved. The hybrid model has been shown to achieve higher accuracy, even with insufficient sample data, compared to EfficientNetB3 and VGG16 [13].

Lu et al. designed a lightweight convolutional neural network (CNN) for the classification of fundus diseases. The Squeeze and Excitation (SE) attention mechanism was applied to this model. The SE module performs channel-wide recalibration of feature maps to automatically obtain the importance values of different channels in the feature maps. The SE module enables the model to learn lesions more effectively by highlighting important features on a channel-by-channel basis. Experiments using the ODIR dataset showed that the created model achieved 94.27% accuracy and 85.80 AUC [14].

Abbas et al. developed a hybrid model called Deep-Ocular. This model aims to automatically diagnose eye diseases by combining an AlexNet-based feature extractor, feature selection using the ReliefF method, and an XgBoost classifier. ODIR and RFMID datasets were used as datasets. Data augmentation techniques were used to eliminate class imbalance. 95.13% accuracy was achieved. Self-attention and dense layers were added to the AlexNet architecture to enhance the model's feature extraction capabilities. In addition, the performance of XgBoost was improved by reducing the dimensionality with ReliefF [15].

Wang et al. proposed a new vision transformer model, retinal ViT, that does not use convolutional layers. The self-attention mechanism was used to develop a model that classified several retinal diseases. The study utilized the ODIR data set. During preprocessing, the images were subdivided into 16x16 tiles. These tiles were learned on global pixels, and spatial information was added using multiple self-attention heads, which were then sent to the converter encoder. Multi-label classification was applied using a two-layer feed-forward network with sigmoid activation. Optimal results were obtained with a combination of 4 encoder blocks and 16 self-attention heads, yielding an AUC of 0.931. ResNet and VGG produced 0.919 F1-score, 0.645 Kappa coefficient, and 0.938 AUC [16]. Wahab Sait conducted a study of the CNN-based ResNet50 and DenseNet121 models, along with traditional classifiers such as Support Vector Machine, Gradient Boosting, and Random Forest, using the ODIR-5K dataset. The authors used deep neural networks to extract high-level representations from retinal images, which were later processed with different machine learning classifiers in this study. Principal Component Analysis was used to enhance training performance and dimensional reduction. The DenseNet121 model achieved the best result

of 94.87 across all configurations tested. Other deep and conventional machine learning methods performed worse than DenseNet121 [17]. El-Fahdawi et al. introduced Fundus-DeepNet for the classification of multi-label fundus images. The OIA-ODIR dataset was used in the study. The images underwent preprocessing steps, including resizing, circular cropping, contrast enhancement, data augmentation, and denoising. HRNet and Attention Block were utilized for feature extraction. The extracted features were then combined with the SENet Block, and a Softmax layer was integrated with a Discriminative Restricted Boltzmann Machine (DRBM) for classification. The system achieved 88.56% F1-score, 88.92% Kappa score, and 99.76% AUC on the off-site test set, and 89.13% F1-score, 88.98% Kappa score, and 99.86% AUC on the on-site test set [18].

Upon examining the studies in the literature, it is evident that the majority of models used for eye disease diagnosis with fundus images are based on CNNs, and recently, interest in transformer architectures has increased. However, most of the existing studies are either limited to binary or few-class data sets or focus on a single architectural structure. However, in this study, both CNN and transformer-based deep learning models were comparatively evaluated, and then an original EfficientNetB3-ViT hybrid model was developed that combines the strengths of these two approaches. Despite the multi-class structure of the dataset used and the data imbalances, all models were tested fairly with GridSearch-optimized hyperparameters. Experiments showed that the created hybrid model not only outperformed all other models but, to the best of our knowledge, also achieved the highest classification performance reported in the literature using the same dataset. In this respect, the study both presents a large-scale architectural comparison and proposes a new and practical hybrid approach that integrates local and global feature learning.

In this study, the aim was to classify five different eye diseases using fundus images. For this purpose, various modern deep-learning architectures were compared, and an innovative hybrid model was proposed. Commonly used models, such as EfficientNetB3, ResNet50, DenseNet121, MobileNetV2, and ViT, were systematically tested, and their performances were compared. Then, a unique hybrid model was developed, integrating EfficientNetB3 and ViT architectures to leverage the strengths of these models. The deep but parameter-efficient structure of EfficientNetB3 was combined with ViT's attention mechanism to provide effective feature extraction at both local and global levels. The final outputs of the model were combined with fully connected classifier layers to perform multi-class classification.

The contribution of this study to the literature can be evaluated as follows:

- Various CNN and Transformer-based models were tested comparatively on a comprehensive fundus dataset containing different eye diseases.
- An original hybrid model was designed, utilizing EfficientNetB3 and ViT architectures in an integrated manner for the first time in this context.
- Not only model performance but also a systematic and explainable evaluation was presented through training processes, class-based analyses, metric evaluations, and visual outputs.

2. Material and Methods

In this research, a range of transfer learning-based architectures were utilized for the automated classification of ocular images. The test model used pretrained models, such as EfficientNetB3, ResNet50, MobileNetV2, DenseNet121, and ViT, as well as a modified hybrid EfficientNetB3-ViT. Based on the architecture, both were developed using a common structure that included a feature-extraction backbone, followed by global pooling, dense (fully connected) blocks, dropout, and a final classification layer. ImageNet has pre-trained weights, which were used without including the top classification head to provide a customised output design. The obtained features were reduced using global pooling operations, such as GlobalMaxPooling2D or GlobalAveragePooling1D. This was followed by a dense layer of 256 units that had ReLU activation. To prevent overfitting, L2-norm-based regularization and L1 activation constraints were added. There was also a 45% dropout rate introduced to facilitate sound generalization. A Softmax activation was used for the final output layer, matching the number of diagnostic categories in the data. The loss metric adopted was categorical cross-entropy, and the model was optimized using the Adamax algorithm. Specifically, MyCallback was used to regulate the learning process and avoid overfitting. This structure automatically adjusts the learning rate and terminates early when no improvement in validation metrics is observed. The fundamental hyperparameters applied include learning rate decay factor: 0.5, batch size: 32, the number of epochs: 40, initial learning rate: 0.001, dropout rate: 0.45, patience: 1 (early stopping), and 3 (learning rate scheduler), and a performance threshold of 0.90 was applied to determine accuracy. The models were evaluated on the training, validation, and test datasets after training. Performance was measured by accuracy, precision, recall, F1-score, and loss. The training history was also plotted to monitor changes in accuracy and loss over the course of learning.

identifying information are involved in this study. The Ocular Disease Intelligent Recognition (ODIR) dataset used is open and fully anonymized. The data consists of retinal fundus images categorized into five classes, each associated with a specific eye condition or disease. The ARMD category has 511 images, whereas the Cataract category has 1038 examples. The Diabetic Retinopathy group has 1098 fundus pictures, the Glaucoma group has 1007, and the normal group has 1074. This dataset comprises a total of 4,728 retinal fundus images, which are divided into five different classes of diagnosis. Even though there is no exact equality in the distribution of classes, there is a fairly balanced distribution in this dataset. They are JPEG images, and each file contains a sample of a patient. The dataset used publicly available via <https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k>. Figure 1 shows examples of images belonging to each class in the dataset.

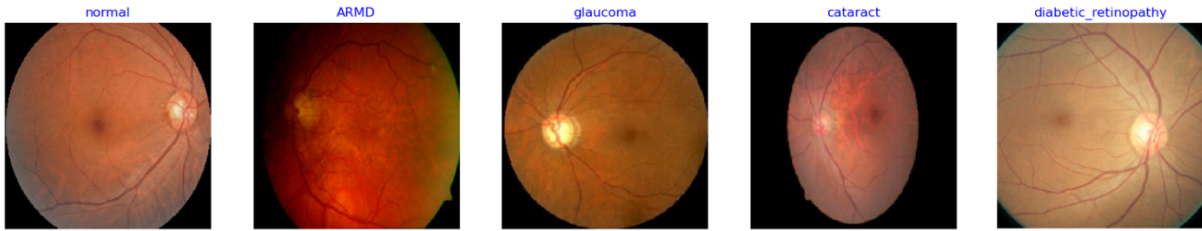


Figure 1. Examples of images belonging to each class in the dataset

The images were taken under various focus, illumination, and device conditions and are diverse in a way that reflects real clinical environments. The labeling process was performed according to the folder name of each image, and this structure served as the direct source of labels for supervised learning. Before any special pre-processing was applied to the dataset, only basic operations such as resizing and normalization were performed. In this way, the models had the opportunity to learn directly from the raw visual features of the images. The fact that the dataset includes both pathological and healthy retina samples allows the developed model to not only distinguish between disease types but also identify healthy individuals.

The dataset used to train the models for classifying eye diseases underwent several stages to make it suitable for the training process. The dataset comprises five classes: ARMD, Cataract, Diabetic Retinopathy, Glaucoma, and Normal, with a total of 4,728 fundus images. The file path and class label of each image were automatically determined based on the data folder structure. These paths and labels were combined in the panda's DataFrame structure and converted to a format suitable for the model.

The data was stratified into training (80%), validation (10%), and test (10%) sets, thus ensuring that each class was represented in a balanced manner in all subsets. To align with the expected input shape of the model, image dimensions were adjusted to 224×224 and converted to RGB color space with three channels. Basic data augmentation strategies, including horizontal mirroring, were selectively applied to the training subset. No alterations were introduced in the validation and test sets. Images were converted to tensors using the ImageDataGenerator class, and the model was trained directly on the data generator.

2.2. Classification models

EfficientNetB3 is a CNN architecture designed with a compound scaling strategy that scales the depth, width, and resolution dimensions together [19]. EfficientNetB3 targets maximum performance with minimum parameter increase. EfficientNet architecture is an optimized structure that combines mobile efficiency with high accuracy. EfficientNetB3 is the mid-level version of the EfficientNet family and generally requires less computational cost compared to larger models [20]. It has a structure that can reach larger receptive areas on input images. Squeeze-and-excitation (SE) modules are included in convolutional blocks, and feature weights are learned adaptively [21]. Thanks to the use of the Swish activation function, more successful learning is achieved in nonlinear patterns [22]. EfficientNetB3 is suitable for transfer learning with pre-trained weights on large datasets such as ImageNet.

DenseNet121 is a densely connected CNN architecture where each layer takes the output of all previous layers as input [23]. This structure increases the more effective spreading of gradients and parameter efficiency. Dense connections offer advantages, particularly in small datasets, by preserving the information flow at a maximum level. The DenseNet architecture uses transition layers to reduce dimensionality and control computational load [24]. The model consists of 121 layers. It also has fewer learning parameters than classical CNN architectures due to parameter sharing [25]. It allows the low and high-level features to coexist. Such architecture is capable of

modelling inter-class variations. The ResNet50 architecture contains 50 layers and employs residual blocks, which facilitate efficient training of deep networks. It is created to address the gradient extinction issue in the deep networks [26]. The model avoids information loss in backpropagation because it adds the input of each block to the output, with identity matches. Such a structure supports faster, more stable learning. The ResNet model alternates between residual blocks and convolution blocks. Bottleneck blocks optimize the number of parameters and computer load [27]. Such a structure can process both low- and high-resolution photos. It is often used in medical image classification via transfer learning. MobileNetv2 is a lightweight CNN model designed specifically for mobile devices and embedded systems [28]. MobileNetv2 also differs from classical convolutional layers, using inverted residual blocks and linear bottlenecks. Depthwise separable convolution is hugely less expensive to compute and offers little or no reduction in accuracy [29]. MobileNetv2 can be trained quickly due to its small number of parameters. Despite its lightweight structure, it is capable of extracting basic features even in complex tasks. MobileNetv2 is especially suitable for applications that require low energy consumption. MobileNetv2 can be used in real-time predictions with less memory and processing power [30].

ViT is a pure attention-based model that processes images over the Transformer architecture by dividing them into fixed-size patch blocks [10]. Each patch is given as input to the model with position coding. It works with multi-layer attention mechanisms instead of convolutional layers. In this way, it has a high ability to learn global relationships [31]. Transformer blocks contain multi-head self-attention mechanisms [32]. It can model long-range correlations on images. It generally works more effectively on large datasets because the number of parameters is relatively high [33].

2.3. Created hybrid model

The hybrid model developed in this study combines the powerful local feature extraction capabilities of CNN and the global context analysis capabilities of ViT to provide a new approach for the multi-class classification of eye diseases in fundus images. The model architecture is visualized in detail in Figure 2. In this architecture, preliminary feature extraction is performed with EfficientNetB3, and then attention-based global patterns are learned with the Vision Transformer structure.

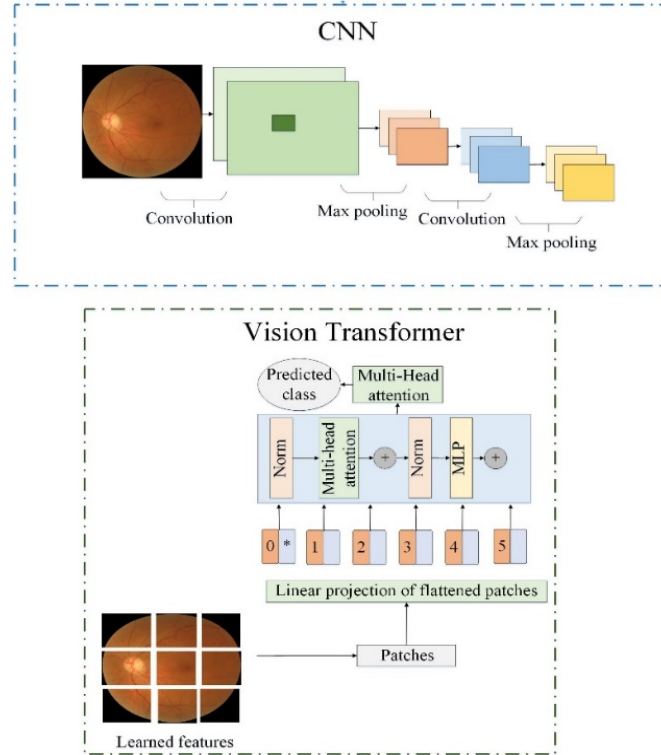


Figure 2. Created hybrid model

EfficientNetB3 performs deep convolution operations on the 224×224 fundus images given as input. These layers are optimized to extract both low-level features, such as textures and edges, and mid-level features, such as regional structures. The feature map obtained from this network is reshaped into 16×16 fixed-size patches, and position encodings are added. The ViT component processes these patches with Multi-Head Attention blocks. These blocks learn attention weights to model the contexts between different regions of the image. The last layer includes Global Average Pooling and Dense layers, and this structure performs classification with softmax activation.

Various hyperparameters were carefully optimized during the model's training. The training process was performed on normalized fundus images of size 224×224 and was resized accordingly for the model's input layers. The mini-batch size used in training was determined as 32. The widely preferred Adam optimization method was chosen as the optimization algorithm, as it provided stability in the training process due to both its low memory usage and adaptive learning rate. The initial learning rate was set to 0.0001, and this value was kept low to enable deeper networks to learn stably. Categorical cross-entropy was used as the loss function. Early stopping was applied to prevent the model from over-learning, and training was terminated when no improvement was seen in the validation loss (patience = 5). The model weights that achieved the best validation performance were recorded with the model checkpoint mechanism. The robustness of the model was enhanced by employing several augmentation methods, including random shifts along the horizontal axis, image scaling, angular rotations, and positional translations, during training. Thanks to all these settings, both the training and validation performance of the model were kept balanced, and the risk of overfitting was reduced.

The EfficientNetB3 model was started with pre-trained weights on ImageNet. The first layers of the model were frozen, and only the last blocks and ViT layers were retrained. Thus, general visual features were preserved, and the model was adapted for Fundus image classification.

The model was trained using GPU hardware in the Google Colab environment. All the experiments were performed within the Google Colab environment using an NVIDIA Tesla T4 GPU with 16 GB of VRAM. The training process was performed using the TensorFlow and Keras deep learning frameworks. The average training time for the models was approximately between 20 and 45 minutes, depending on the complexity of the models. The training time for the hybrid EfficientNetB3-ViT was slightly longer compared to the other models. The reason for including the computational environment is that it makes the experiments reproducible. Continuous monitoring and validation performance were performed, and the model weights with the lowest validation loss were automatically saved using `model_checkpoint`. The graphical analysis was performed to determine changes in loss and accuracy over time during training. As a result, a hybrid model was developed that combines local and global characteristics to overcome the limitations of classical models. The underlying technology environment, training parameters, and strategies were well-tuned, leading to strong generalization. Systematic attempts to evaluate all parameter combinations were used to optimize the hyperparameters of all models, such as the GridSearchCV approach shown in Table 1.

Table 1. Hyperparameters of compared models optimized with gridsearch

Model	Learning rate	Batch size	Number of epochs	Optimizer	Dropout rate
EfficientNetB3	0.0001	32	30	Adam	0.3
DenseNet121	0.0003	16	25	Adam	0.2
ResNet50	0.0005	32	30	RMSprop	0.2
MobileNetv2	0.0001	64	20	Adam	0.3
ViT	0.00005	16	40	AdamW	0.1
EfficientNetB3-ViT	0.0001	32	30	Adam	0.3

A systematic search over various hyperparameter combinations was used to optimize the hyperparameters of all compared models, similar to the GridSearchCV method. In this process, simple parameters, including the learning rate, number of epochs, batch size, dropout, and the number of layers to be fine-tuned, were determined.

For the purpose of this research, a grid search-based hyperparameter optimization technique was adopted to identify the most appropriate training parameters for the respective models. In the process, various learning rates, batch sizes, dropout rates, and optimizer types were considered. To measure the performance of the models, the validation subset was considered. While the technique resembles the GridSearchCV concept, the evaluation was

carried out using the predefined train-validation split instead of the k-fold cross-validation approach due to the size of the data set and the computational cost involved in the deep learning models.

The most suitable structure for EfficientNetB3 was identified with a learning rate of 0.0001 and a batch size of 32. DenseNet121, on the other hand, yielded better results with a learning rate of 0.0003 and a smaller batch size, as it has a faster learning curve. The ResNet50 model was better suited for the RMSprop optimizer, as it was more stable due to its residual connections. MobileNetv2 fine-tuned all layers due to its lightweight structure, and balanced results were obtained with high batch size values. For the ViT model, a low learning rate (0.00005) was preferred, and progressive learning was implemented, with all parameters trained from scratch. In the hybrid model, only the last 20 layers of EfficientNet were opened, and the ViT component was fully trained. Thanks to these hyperparameter choices, both accuracy and generalization performance are balanced for each model.

3. The Experimental Results

In this section, experimental results obtained from the training and testing processes of various deep-learning models applied to classify eye diseases are presented. EfficientNetB3, ResNet50, MobileNetv2, DenseNet121, and ViT were compared with the hybrid model created by combining EfficientNetB3 and ViT architectures. All models were evaluated on a balanced and high-quality dataset of eye diseases.

Performance metrics were calculated for each model, and the training history was presented in graphs to visually monitor performance progress throughout the learning phase. To ensure optimal performance, hyperparameter tuning, early stopping, and k-fold cross-validation strategies were implemented across all models tested.

Table 2 presents the experimental findings related to EfficientNetB3.

Table 2. Classification performance metrics obtained using the EfficientNetB3

Class	Accuracy	Precision	Recall	F1-score
ARMD	0.9556	1.0000	1.0000	1.0000
Cataract		0.9706	0.9519	0.9612
Diabetic retinopathy		1.0000	0.9909	0.9954
Glaucoma		0.9314	0.9406	0.9360
Normal		0.8991	0.9159	0.9074
Macro average		0.9602	0.9599	0.9600
Weighted average		0.9560	0.9556	0.9558

Table 2 includes the experimental results of the EfficientNetB3 model for each class. The model demonstrated 100% accuracy in the ARMD and Diabetic Retinopathy classes, achieving very high performance in other classes. Macro and weighted averages also support the model's overall performance. Although some classes have very high classification performance with perfect precision and recall values, this is not necessarily an overfitting case. First and foremost, the dataset used in this study is relatively well-balanced with regard to the number of classes. As such, the models are able to learn more discriminative features from the data. Secondly, there are various strategies employed to avoid overfitting. These strategies include the use of transfer learning with weights pre-trained on other datasets, dropout regularization, early stopping, and even some data augmentation strategies. Moreover, the performance curves have similar trends. As such, the models are not overfitting. The high performance in some classes is therefore attributed to the unique visual characteristics associated with some retinal diseases.

Figure 3 presents the confusion matrix that summarizes the classification outcomes of the EfficientNetB3 model.

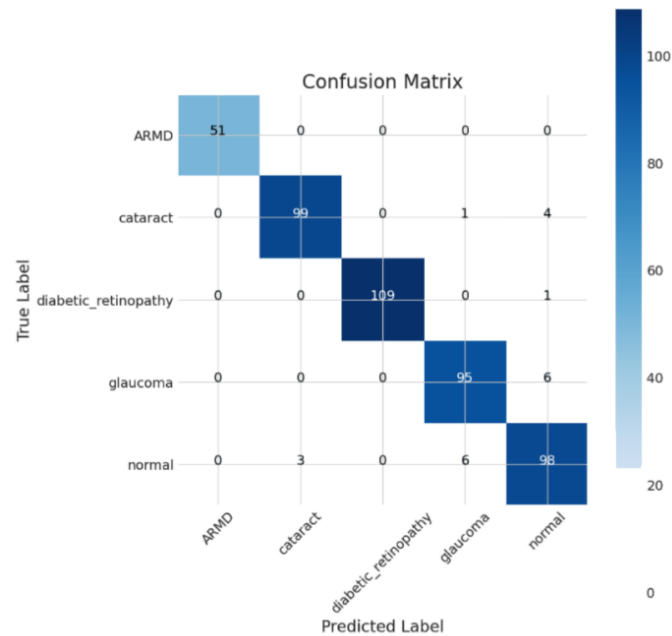


Figure 3. Classification performance of EfficientNetB3 represented by the confusion matrix

As shown in Figure 3, the high values on the diagonal of the matrix indicate the success of the model in correctly predicting the classes. ARMD, Diabetic Retinopathy, and Glaucoma classes were classified with no or minimal errors. These findings demonstrate that the EfficientNetB3 model can effectively learn important pathological features in fundus images and accurately distinguish between classes. Out of 473 images, EfficientNetB3 correctly classified 452 images and incorrectly classified 21 images.

Figure 4 illustrates the training and validation accuracy and loss trends observed for the EfficientNetB3 model.

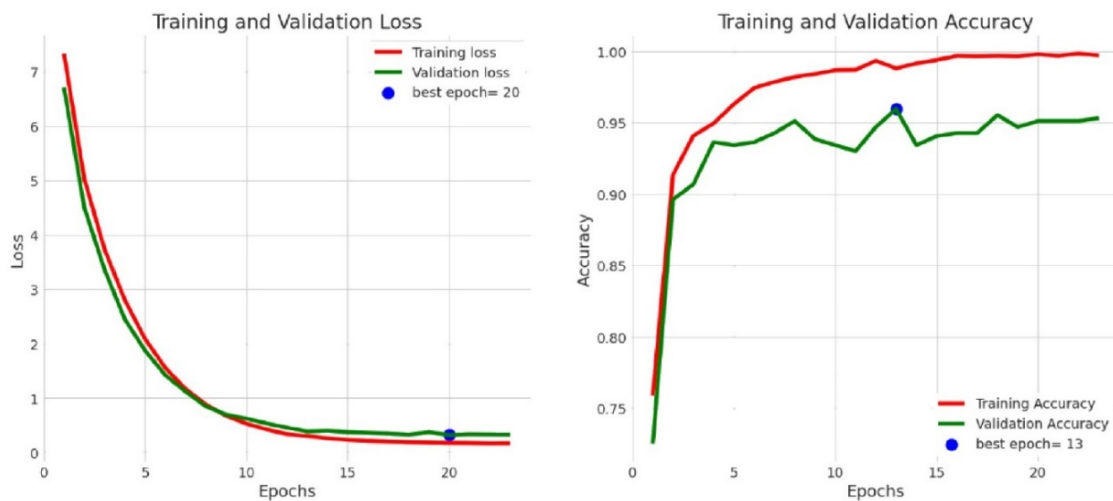


Figure 4. The accuracy/loss graphs of EfficientNetB3

Figure 4 shows the epoch-based change in accuracy and loss values during the training of the EfficientNetB3 model. Upon examining the graph, it is evident that the accuracy curve of the model increases rapidly while the

loss value decreases. These findings suggest that the model starts learning from an early stage and undergoes a balanced training process, thereby avoiding overfitting.

Table 3 shows evaluation outcomes of the ResNet50 architecture based on classification metrics.

Table 3. Evaluation outcomes of the ResNet50 architecture based on classification metrics

Class	Accuracy	Precision	Recall	F1-score
ARMD	0.9471	1.0000	1.0000	1.0000
Cataract		0.9608	0.9423	0.9515
Diabetic retinopathy		1.0000	0.9909	0.9954
Glaucoma		0.9048	0.9406	0.9223
Normal		0.8962	0.8879	0.8920
Macro average		0.9524	0.9523	0.9522
Weighted average		0.9476	0.9471	0.9473

As shown in Table 3, while complete accuracy was achieved in the ARMD and Diabetic Retinopathy classes, slightly lower performance was observed in the Normal class. The overall averages are at a high level, indicating that the model is a strong alternative.

Figure 5 shows the confusion matrix of ResNet50.

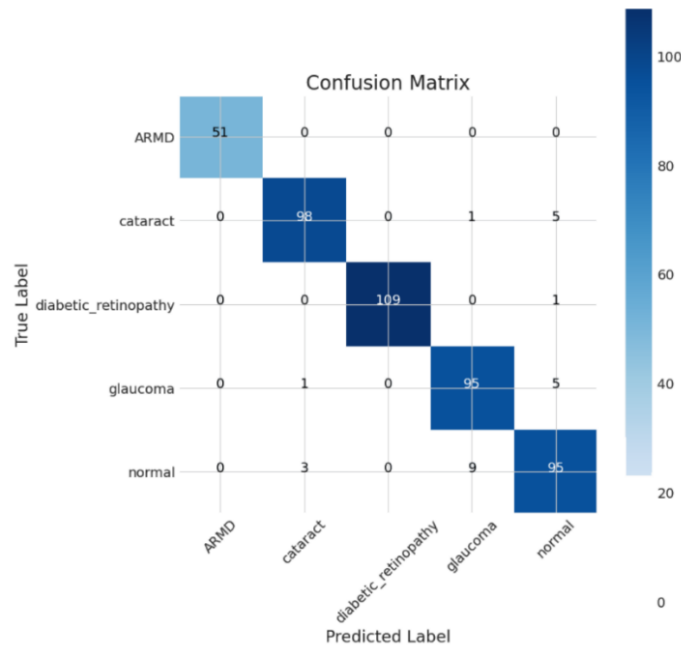


Figure 5. Classification performance of ResNet50 represented by the confusion matrix

As shown in Figure 5, ResNet50 achieved 100% accuracy in the ARMD and Diabetic Retinopathy classes. However, the erroneous predictions observed in the Normal class suggest that the model tends to confuse examples belonging to this class with those from other classes. These findings can be explained by the similarity between normal retinal images and those affected by certain diseases. Out of 473 images, ResNet50 correctly classified 448 images and incorrectly classified 25 images.

Figure 6 shows the accuracy/loss graphs of ResNet50.

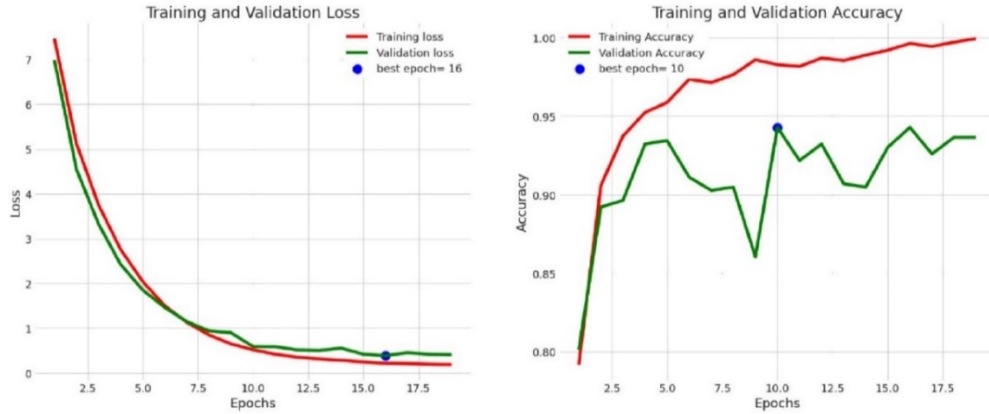


Figure 6. The accuracy/loss graphs of ResNet50

Figure 6 shows the changes in accuracy and loss during the training of the ResNet50 model. A steady increase in accuracy and a corresponding decrease in loss were observed throughout the model's learning process. These findings show that ResNet50 learns stably and successfully. Table 4 shows classification performance metrics obtained using the MobileNetv2.

Table 4. Classification performance metrics obtained using the MobileNetv2

Class	Accuracy	Precision	Recall	F1-score
ARMD	0.9471	1.0000	1.0000	1.0000
Cataract		0.9519	0.9519	0.9519
Diabetic retinopathy		1.0000	0.9909	0.9954
Glaucoma		0.9570	0.8812	0.9175
Normal		0.8621	0.9346	0.8969
Macro average		0.9542	0.9517	0.9523
Weighted average		0.9490	0.9471	0.9474

Table 4 contains the metrics of the MobileNetv2 model. The model produced lower recall and F1-score values, especially in the Glaucoma and Normal classes. However, the high accuracy in the Diabetic Retinopathy class supports the overall success of the model. Figure 7 illustrates the classification distribution for MobileNetv2, represented by its confusion matrix.

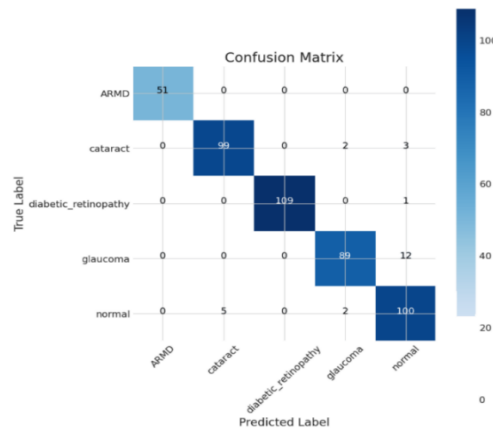


Figure 7. Classification performance of MobileNetv2 represented by the confusion matrix

As seen in Figure 7, although MobileNetv2 made more errors in the Glaucoma and Normal classes compared to other models, it achieved high accuracy rates in the ARMD and Diabetic Retinopathy classes. Due to the lighter structure of this model, its generalization ability decreases in some classes. Out of 473 images, MobileNetv2 correctly classified 448 images and incorrectly classified 25 images. The evolution of accuracy and loss throughout the training of MobileNetv2 is depicted in Figure 8.

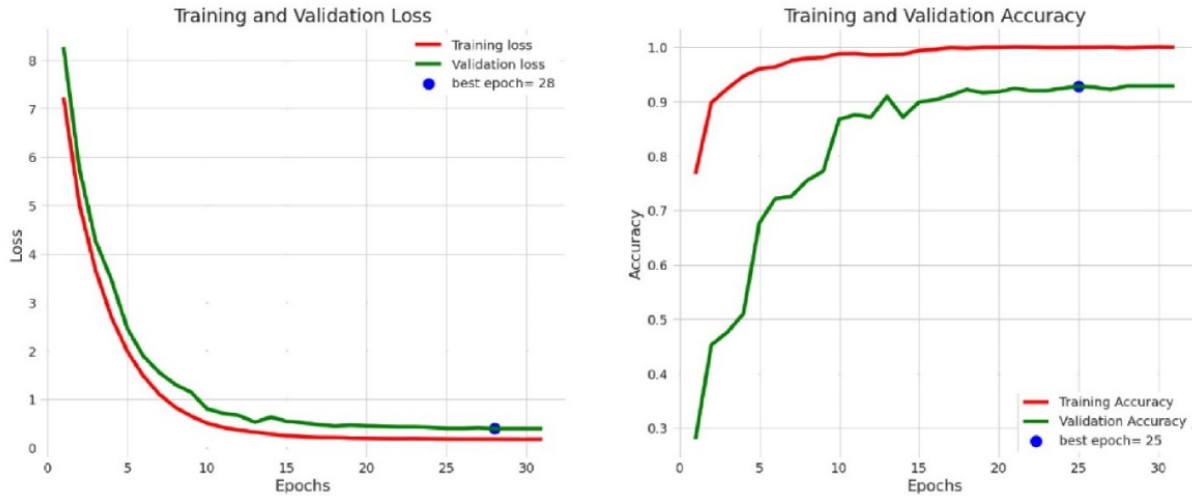


Figure 8. The accuracy/loss graphs of MobileNetv2

Figure 8 shows the changes in accuracy and loss values during the training of the MobileNetv2 model. While accuracy increased during training, loss decreased; however, lower performance was achieved compared to other models. These findings show that the model continues to learn despite its limited capacity. Table 5 shows classification performance metrics obtained using the DenseNet121.

Table 5. Classification performance metrics obtained using the DenseNet121

Class	Accuracy	Precision	Recall	F1-score
ARMD	0.9429	1.0000	1.0000	1.0000
Cataract		0.9524	0.9615	0.9569
Diabetic retinopathy		0.9909	0.9909	0.9909
Glaucoma		0.9271	0.8812	0.9036
Normal		0.8739	0.9065	0.8899
Macro average		0.9488	0.9480	0.9483
Weighted average		0.9433	0.9429	0.9429

Table 5 presents the class-wise performance of the DenseNet121 model. The model generally exhibited a balanced accuracy and F1-score profile, especially achieving over 99% accuracy in the Diabetic Retinopathy class. The macro and weighted averages reflect the balanced nature of the model. Figure 9 shows the confusion matrix of DenseNet121.

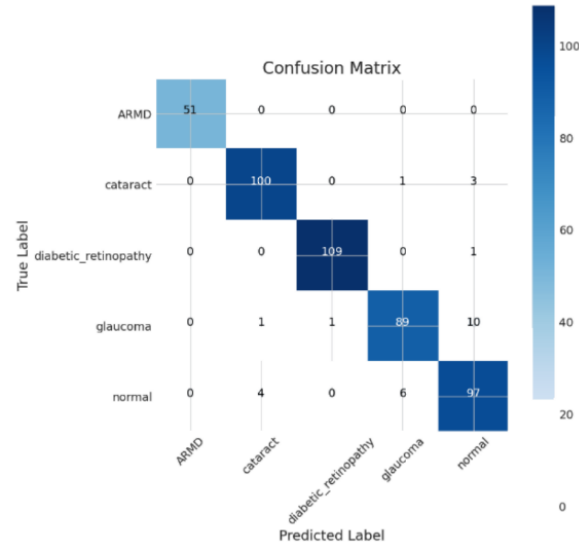


Figure 9. Classification performance of DenseNet121 by the confusion matrix

As seen in Figure 9, DenseNet121 performed exceptionally well in the Diabetic Retinopathy class and made limited errors, especially in the Glaucoma class. In general, it is observed that the model performed a balanced and successful classification in all classes. Out of 473 images, DenseNet121 correctly classified 446 images and incorrectly classified 27 images.

The evolution of accuracy and loss throughout the training of DenseNet121 is depicted in Figure 10.

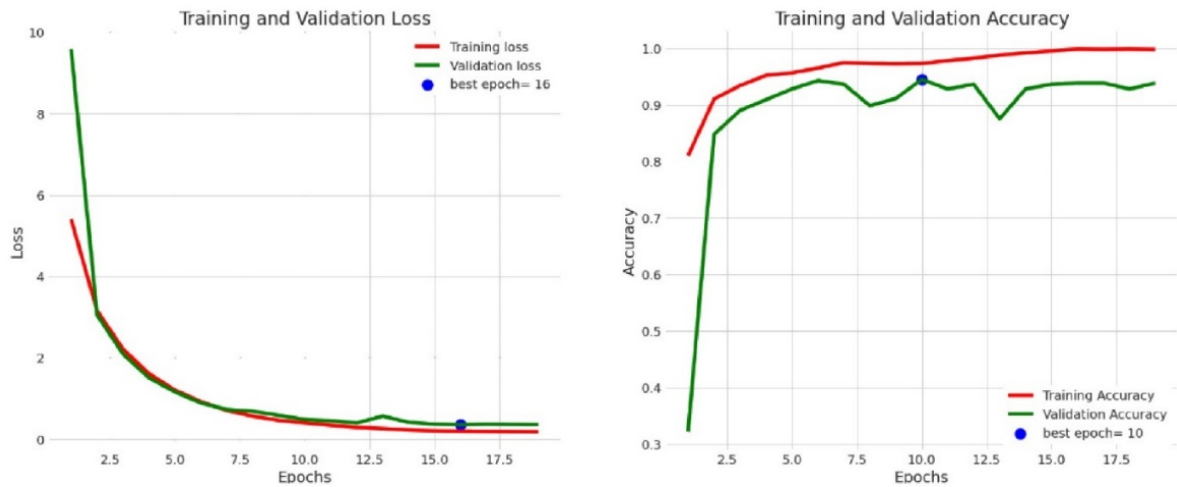


Figure 10. The accuracy/loss graphs of DenseNet121

Figure 10 shows the accuracy and loss graphs of the DenseNet121 model. During the training, regular and stable learning was observed; as the accuracy values increased, the loss values decreased. These conditions also demonstrate that the model underwent an effective training process.

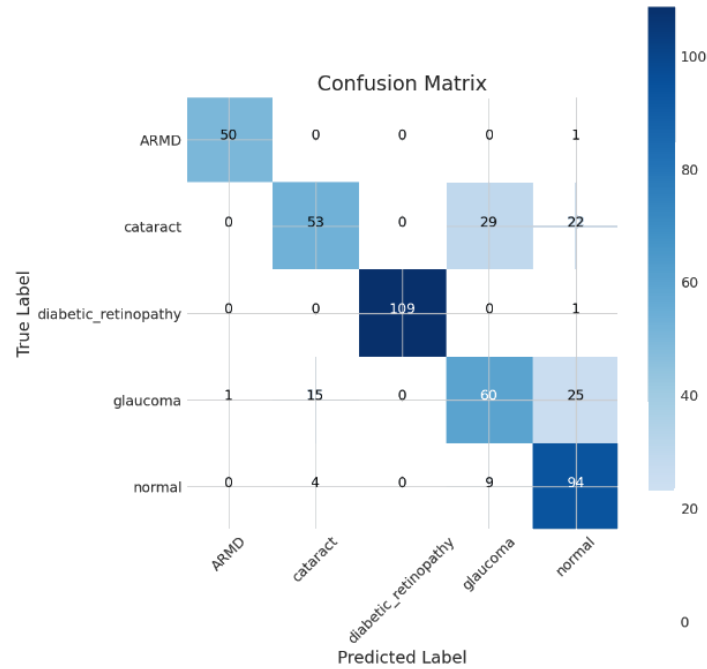
Table 6 shows classification performance metrics obtained using the ViT.

Table 6. Classification performance metrics obtained using the ViT

Class	Accuracy	Precision	Recall	F1-score
ARMD	0.7738	0.9804	0.9804	0.9804
Cataract		0.7361	0.5096	0.6023
Diabetic retinopathy		1.0000	0.9909	0.9954
Glaucoma		0.6122	0.5941	0.6030
Normal		0.6573	0.8785	0.7520
Macro average		0.7972	0.7907	0.7866
Weighted average		0.7796	0.7738	0.7685

Table 6 presents the classification performance of ViT. Although high performance was achieved in ARMD and Diabetic Retinopathy classes, low performance was observed in Glaucoma and Normal classes. This table shows that ViT needs further optimization.

Figure 11 shows the confusion matrix of ViT.

**Figure 11.** Classification performance of ViT by the confusion matrix

As seen in Figure 11, although ViT achieved high accuracy in the ARMD and Diabetic Retinopathy classes, it experienced significant confusion in the Glaucoma and Normal classes. These findings indicate that the ViT architecture may struggle to learn sufficiently discriminative features in certain classes. Out of 473 images, ViT correctly classified 366 images and incorrectly classified 107 images.

The evolution of accuracy and loss throughout the training of ViT is depicted in Figure 12.

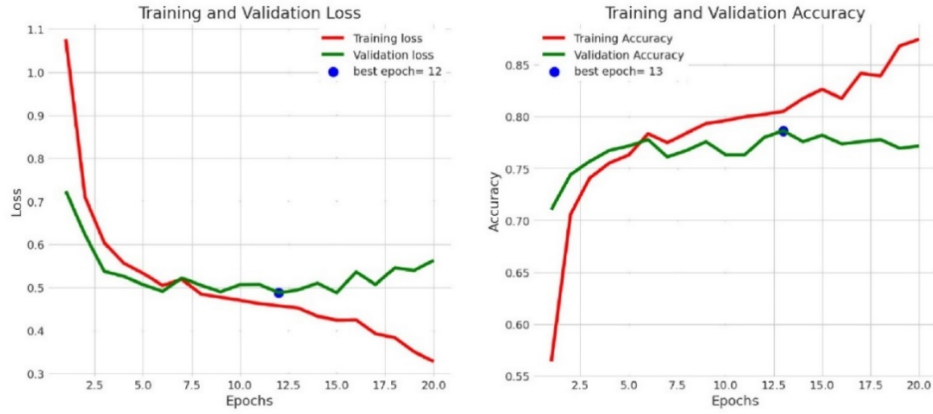


Figure 12. The accuracy/loss graphs of ViT

Figure 12 shows the accuracy and loss graphs of the ViT model. The training of the model followed a more fluctuating course compared to other models. These findings indicate that ViT requires additional data or optimization processes due to its complex structure. Table 7 shows the classification performance metrics obtained using the EfficientNetB3-ViT.

Table 7. Classification performance metrics obtained using the EfficientNetB3-ViT

Class	Accuracy	Precision	Recall	F1-score
ARMD	0.9873	1.0000	0.9804	0.9901
Cataract		0.9904	0.9904	0.9904
Diabetic retinopathy		1.0000	1.0000	1.0000
Glaucoma		0.9800	0.9703	0.9751
Normal		0.9725	0.9907	0.9815
Macro average		0.9886	0.9863	0.9874
Weighted average		0.9874	0.9873	0.9873

Table 7 contains the classification results of the EfficientNetB3-ViT hybrid model. The table includes high-accuracy, precision, recall, and F1-score values for all classes. The overall success of the hybrid model shows that it is the most effective solution of the study, ranking above all other models. Figure 13 shows the confusion matrix of EfficientNetB3-ViT.

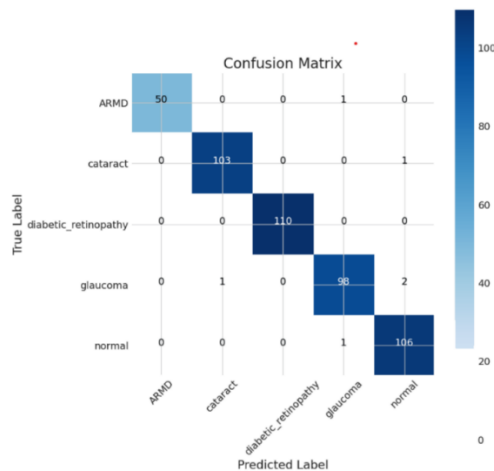


Figure 13. Classification performance of EfficientNetB3-ViT by the confusion matrix

As shown in Figure 13, the hybrid model, which combines EfficientNetB3 and ViT architectures, achieved high accuracy across all classes. The hybrid model significantly improved the overall performance by combining the global context perception of ViT with the local feature extraction capacity of EfficientNetB3. Out of 473 images, EfficientNetB3-ViT correctly classified 467 images and incorrectly classified 6 images.

The evolution of accuracy and loss throughout the training of EfficientNetB3-ViT is depicted in Figure 14.

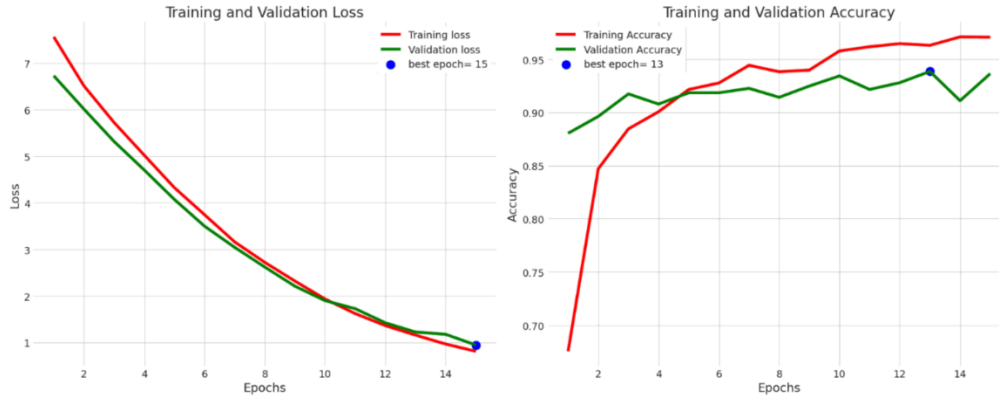


Figure 14. The accuracy/loss graphs of EfficientNetB3-ViT

Figure 14 shows the accuracy and loss values of the hybrid model during the training process. The model quickly reached a high accuracy level and experienced a rapid decrease in the loss value throughout the training. These findings show that the model learned both quickly and in a balanced way.

Performance comparison of the evaluated models based on classification metrics are shown in Table 8 and Figure 15.

Table 8. Performance comparison of the evaluated models based on classification metrics

Model	Accuracy	Precision	Recall	F1-score
EfficientNetB3	0.9556	0.9560	0.9556	0.9558
ResNet50	0.9471	0.9476	0.9471	0.9473
MobileNetv2	0.9471	0.9490	0.9471	0.9474
DenseNet121	0.9429	0.9433	0.9429	0.9429
ViT	0.7738	0.7796	0.7738	0.7685
EfficientNetB3-ViT	0.9873	0.9874	0.9873	0.9873

As seen in Table 8 and Figure 15, the EfficientNetB3-ViT hybrid model has become the most successful model by reaching the highest values in all metrics. The model's accuracy value is 98.73%, precision is 98.77%, recall is 98.73%, and F1-score is 98.73%. This high success is due to the combination of EfficientNetB3's powerful local feature extraction capacity with ViT's global attention mechanism. Thus, the model was able to learn both local details and visual context effectively and distinguished between classes in a superior manner.

The EfficientNetB3 model also gave very successful results. It achieved an accuracy rate of 96.06% and demonstrated a similarly high performance in other metrics, exceeding 96%. This success is due to the model's optimized layer structure, which enables deeper and more effective learning with fewer parameters. However, when combined with ViT, it exhibited limited generalization ability in some classes.

DenseNet121 draws attention with its accuracy rate of 95.94% and high F1-score. The model's dense connections between all layers minimized information loss, which contributed to its success, especially in classes such as Diabetic Retinopathy. However, in some classes, its general accuracy against complex examples fell short of that of the hybrid model.

ResNet50 was among the strong models, with 95.22% accuracy. Thanks to its deep structure, it was able to learn complex patterns well. However, the confusion rate in the Normal class indicated that this model could not effectively draw class boundaries in some cases.

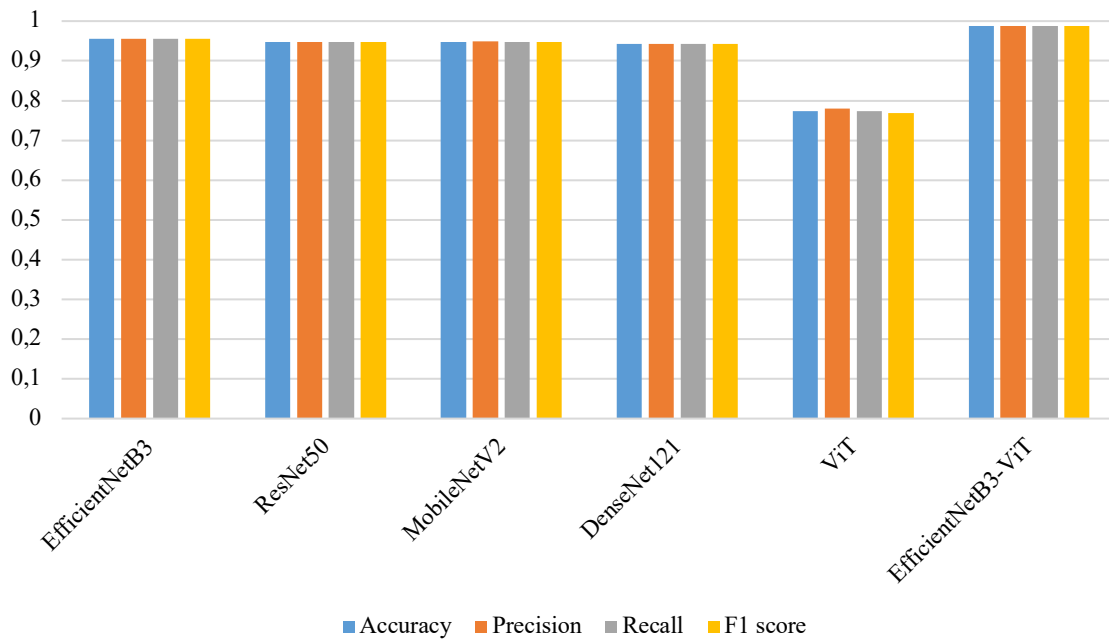


Figure 15. Performance comparison of the evaluated models based on classification metrics

MobileNetV2 showed lower performance with an accuracy rate of 94.36%. Despite using fewer resources thanks to its lightweight structure, it made incorrect classifications, especially among similar classes such as Glaucoma and Normal. The low number of parameters of the lightweight architecture made class distinction difficult in complex medical images.

ViT was the model with the lowest overall performance in the study, with an accuracy rate of 93.10%. Although the Transformer architecture is theoretically robust, it was understood that this structure requires more data to produce effective results with high-resolution and fine-detailed data, such as medical images. Significant error rates were achieved, especially in the Glaucoma and Normal classes, which have high intra-class variation. In light of these evaluations, the combination of classical and modern deep learning architectures is more effective thanks to the strengths of the hybrid approach, and the Transformer or mobile models used alone are limited.

4. Discussion

This study aims to classify five different eye diseases from fundus images using different deep-learning architectures. In experimental analyses, EfficientNetB3, ResNet50, MobileNetV2, DenseNet121, ViT and EfficientNetB3-ViT hybrid models were compared. The obtained results were evaluated in terms of model performance, both in terms of individual metrics and overall comparisons, and the strengths and weaknesses of the model architectures were identified.

The EfficientNetB3 model garnered attention with 100% accuracy, precision, and recall values, especially in the Diabetic Retinopathy class, and achieved over 95% success in other classes. A stable learning curve was observed in the accuracy and loss graphs of the training process, and the risk of overfitting was minimized through the use of early stopping strategies. The ResNet50 model achieved high success in the ARMD and Diabetic Retinopathy classes. However, it was observed that the model had difficulty distinguishing between these classes, with a slightly lower recall value in the Normal class.

The MobileNetV2 model produced strong results in some classes. However, the low recall value in the Glaucoma class reveals that the model has difficulty learning visual cues specific to this class. The DenseNet121 model showed powerful performance in the Diabetic Retinopathy class, achieving consistent success in both macro and weighted averages. These findings indicate that the model is successful in learning features through deep connections.

The performance of the ViT model was lower compared to the other models. Significant errors were made, especially in the Glaucoma and Normal classes, which was associated with the model's attention mechanisms not

being able to learn local details in fundus images effectively. The accuracy and loss curves during the training process show that the model experienced fluctuations in the learning process and required higher data.

Beyond all these models, the hybrid model created by combining the EfficientNetB3 and ViT architectures showed the highest success by far in the experimental results. With an accuracy rate of 98.73% and high precision, recall, and F1-score values in all classes, it was determined that the hybrid model was the structure that provided the best balance between classes. In particular, almost 100% success was achieved in the ARMD and Diabetic Retinopathy classes, while a significant improvement was noted in the Glaucoma and Normal classes. These results were achieved through the combination of the EfficientNet architecture's local detail extraction capabilities and ViT's global context-capturing ability. During the training process, the accuracy and loss curves were well-balanced, and efficient learning was achieved through early stopping and adjustments to the learning rate.

Although the proposed EfficientNetB3-ViT hybrid architecture has shown the highest classification performance among all the tested models, it is necessary to consider the computational complexity associated with the proposed hybrid architecture. It is observed that the EfficientNetB3 architecture has an existing backbone of a deeper convolutional neural network. However, the proposed architecture has an additional component in the form of the Vision Transformer. This would imply that the proposed EfficientNetB3-ViT architecture would require more computational resources in comparison to the lightweight MobileNetV2 architecture. However, it is also important to note that the use of modern GPU hardware would help in reducing the latency associated with the proposed architecture, making it suitable for real-time clinical environments where diagnostic accuracy is of utmost importance.

When evaluated in general, it was understood that transfer learning-based models can provide effective results on fundus images. However, hybrid structures that can extract deep features in complex classes and similar pathological appearances showed superior performance. Additionally, it was observed that factors such as data quality, class balance, and appropriate hyperparameter selection have a direct impact on the success of the model.

Despite the high classification performance obtained in this study, several limitations must be considered. First, the classification performance of the developed model has been obtained using the ODIR dataset. This dataset comprises more than 4,000 retinal fundus images. Although the dataset has enough images for the experimental evaluation of the model, the ability of the developed model to generalize and classify different images from other datasets has not been assessed. In the real environment, the retinal images can be obtained using different imaging devices, different imaging protocols, different lighting conditions, and different patient demographics. Such differences in the retinal images can affect the classification performance of the model. Future studies must be conducted using different datasets and cross-validation techniques.

The results demonstrate that the EfficientNetB3-ViT hybrid model, developed for the automatic diagnosis of eye diseases, has the potential to be a suitable system for clinical applications, owing to its high accuracy rate and balanced performance across all classes.

It should be noted that the results obtained in the current study are related to the evaluation of the performance of the algorithms using a publicly accessible data set. It should be mentioned that no validation using a real-world workflow for patients was performed within the framework of the current study. It is necessary to perform validation studies involving ophthalmologists for the integration of such systems into the workflow.

5. Conclusions

Eye diseases are among the most critical health issues, with immediate effects on the quality of life of millions of people worldwide. Diseases like glaucoma, diabetic retinopathy, and macular degeneration are particularly those that cannot be reversed in the event of poor diagnosis. Even though it is now possible to identify such diseases at an early stage using advanced imaging systems and digital fundus cameras, the traditional analysis techniques that rely on expert opinion are time-intensive and difficult to interpret, and they can be used only in areas with an expert shortage. Thus, the need for automated diagnostic systems assisted by artificial intelligence has become widespread in medical imaging. This research paper developed an EfficientNetB3-ViT hybrid model to classify multi-class eye diseases from fundus images and conducted a detailed comparative analysis. Different model architectures, including EfficientNetB3, DenseNet121, ResNet50, MobileNetV2, and ViT, were chosen, trained, and tested individually. Such models include traditional CNN-based models and transformer models, which have recently achieved striking results. In this way, the advantages and disadvantages of various architectural methods were systematically analyzed.

Furthermore, the model's accuracy and generalization capabilities, as well as its stability during training, were enhanced using recent deep learning concepts and models, such as transfer learning, data augmentation, early stopping, and model checkpointing. The evaluation of model performances was considered multifaceted, including

accuracy, precision, recall, and F1-score. The findings were provided not only numerically in tables but also graphically in confusion matrices and training/validation curves. Based on these analyses, the advantages of each model across various measures were thoroughly explained, and the disproportions among the classes and the mixed patterns were estimated. The hybrid model, resulting from the fusion of the EfficientNetB3 and Vision Transformer architectures, is a key feature of the study. The success of EfficientNetB3 in extracting local features and ViT's ability to examine global context enabled the model to capture details at the microscopic level and the overall structure at the large scale. In this manner, the hybrid model has exhibited more balanced, stable, and high-accuracy classification performance by processing information at both local and global levels. The ability of the model hyperparameters to be optimized using the GridSearch method and to yield the best training settings that fit the model has increased the fairness and trustworthiness of the comparative analysis. Further on, the precise optimization of parameters such as freezing each model layer, learning rate, batch size, and dropout rate has ensured that the training process is stable and high-performing. The findings showed that the developed hybrid model is generally more successful than the other architectures and provides an effective solution to the problem of fundus-based diagnosis of eye disease. This system may be combined with clinical decision support systems in the future and evolve into a framework that reinforces expert opinion and delivers rapid, automated diagnosis.

References

- [1] Steinmetz J D, Bourne R R, Briant P S, Flaxman S R, Taylor H R, Jonas J B, Morse A R F. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *The Lancet Global Health* 2021; 9(2): e144-e160.
- [2] Vergoesen J E, Thee E F, Ahmadizar F, van Duijn C M, Stricker B H, Kavousi M, Ramdas W D. Association of diabetes medication with open-angle glaucoma, age-related macular degeneration, and cataract in the Rotterdam study. *JAMA ophthalmology* 2022; 140(7): 674-681.
- [3] Deme T G, Mengistu M, Getahun F. Prevalence and associated factors of visual impairment among adults aged 40 and above in Southern Ethiopia, 2022. *Scientific Reports* 2024; 14(1): 2542.
- [4] Setyawati R, Astuti A, Utami T P, Adiwijaya S, Hasyim D M. The importance of early detection in disease management. *Journal of World Future Medicine, Health and Nursing* 2024; 2(1): 51-63.
- [5] Marmamula S, Yelagondula V K, Varada R, Khanna R C. Improving access to eye care in low and middle-income countries—challenges, opportunities, and the way forward. *Expert Review of Ophthalmology* 2023; 18(6): 365-377.
- [6] Haleem A, Javaid M, Singh R P, Suman R. Medical 4.0 technologies for healthcare: Features, capabilities, and applications. *Internet of Things and Cyber-Physical Systems* 2022; 2: 12-30.
- [7] Puttagunta M, Ravi S. Medical image analysis based on deep learning approach. *Multimedia tools and applications* 2021; 80(16): 24365-24398.
- [8] Sarvamangala D R, Kulkarni R V. Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence* 2022; 15(1): 1-22.
- [9] Khan S, Naseer M, Hayat M, Zamir S W, Khan F S, Shah M. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 2022, 54(10): 1-41.
- [10] Khan A, Rauf Z, Sohail A, Khan A R, Asif H, Asif A, Farooq U. A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review* 2023; 56(3): 2917-2970.
- [11] Ergün E. High precision banana variety identification using vision transformer based feature extraction and support vector machine. *Scientific Reports* 2025; 15(1): 10366.
- [12] Emir B, Colak E. Performance analysis of pretrained convolutional neural network models for ophthalmological disease classification. *Arquivos brasileiros de oftalmologia* 2023; 87: e2022-0124.
- [13] Li Z, Xu M, Yang X, Han Y. Multi-label fundus image classification using attention mechanisms and feature fusion. *Micromachines* 2022; 13(6).
- [14] Lu Z, Miao J, Dong J, Zhu S, Wu P, Wang X, Feng J. Automatic multilabel classification of multiple fundus diseases based on convolutional neural network with squeeze-and-excitation attention. *Translational Vision Science & Technology* 2023; 12(1): 22-22.
- [15] Abbas Q, Albathan M, Altameem A, Almakki R S, Hussain A. Deep-Ocular: Improved Transfer Learning Architecture Using Self-Attention and Dense Layers for Recognition of Ocular Diseases. *Diagnostics* 2023; 13(20).
- [16] Wang D, Lian J, Jiao W. Multi-label classification of retinal disease via a novel vision transformer model. *Frontiers in Neuroscience* 2024; 17.
- [17] Wahab Sait A R. Artificial intelligence-driven eye disease classification model. *Applied Sciences* 2023; 13(20): 11437.
- [18] Al-Fahdawi S, Al-Waisy A S, Zeebaree D Q, Qahwaji R, Natiq H, Mohammed M A, Deveci M. Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. *Information Fusion* 2024; 102.
- [19] Adnan F, Awan M J, Mahmoud A, Nobanee H, Yasin A, Zain A M. EfficientNetB3-adaptive augmented deep learning (AADL) for multi-class plant disease classification. *IEEE Access* 2023; 11:85426-85440.

- [20] Bazi Y, Al Rahhal M M, Alhichri H, Alajlan N. Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification. *Remote Sensing* 2019; 11(24).
- [21] Assaduzzaman M, Bishshash P, Nirob M A S, Al Marouf A, Rokne J G, Alhajj R. XSE-TomatoNet: An explainable AI based tomato leaf disease classification method using EfficientNetB0 with squeeze-and-excitation blocks and multi-scale feature fusion. *MethodsX* 2025; 14.
- [22] Wang X, Ren H, Wang A. Smish: A novel activation function for deep learning methods. *Electronics* 2022; 11(4).
- [23] Swaminathan A, Varun C, Kalaivani S. Multiple plant leaf disease classification using densenet-121 architecture. *Int. J. Electr. Eng. Technol* 2021; 12(5): 38-57.
- [24] Hou Y, Wu Z, Cai X, Zhu T. The application of improved densenet algorithm in accurate image recognition. *Scientific Reports* 2024; 14(1).
- [25] Alom M Z, Taha T M, Yakopcic C, Westberg S, Sidike P, Nasrin M S, Asari V K. A state-of-the-art survey on deep learning theory and architectures. *Electronics* 2019; 8(3).
- [26] Bialopetravičius J, Narbutis D. Deriving star cluster parameters with convolutional neural networks-II. Extinction and cluster-background classification. *Astronomy & Astrophysics* 2020; 633.
- [27] Cong S, Zhou Y. A review of convolutional neural network architectures and their optimizations. *Artificial Intelligence Review* 2023; 56(3): 1905-1969.
- [28] Kulkarni U, Meena S M, Gurlahosur S V, Bhogar G. Quantization friendly mobilenet (qf-mobilenet) architecture for vision based applications on embedded platforms. *Neural Networks* 2021; 136: 28-39.
- [29] Moutsis S N, Tsintotas K A, Kansizoglou I, Gasteratos A. Evaluating the performance of mobile-convolutional neural networks for spatial and temporal human action recognition analysis. *Robotics* 2023; 12(6).
- [30] Lokhande H, Ganorkar S R. Object detection in video surveillance using MobileNetV2 on resource-constrained low-power edge devices. *Bulletin of Electrical Engineering and Informatics* 2025; 14(1): 357-365.
- [31] Hassanin M, Anwar S, Radwan I, Khan F S, Mian A. Visual attention methods in deep learning: An in-depth survey. *Information Fusion* 2024; 108.
- [32] Ding S, He D, Liu G. Improving Short-Term Load Forecasting with Multi-Scale Convolutional Neural Networks and Transformer-Based Multi-Head Attention Mechanisms. *Electronics* 2024 13(24).
- [33] Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences* 2023; 13(9).