DÜZCE ÜNİVERSİTESİ
BİLİM ve TEKNOLOJİ DERGİSİ
Düzce University Journal of Science & Technology
Cilt Volume 13
Sayı Issue 1
2025

Düzce University
Journal of Science & Technology

Research Article

Detection of Eye Pressure Disease Using ViT-Based Hybrid Learning Methods

Mahmut KAYA [a,*], Yusuf BİLGEN [b]

[a,*] *Department of Artificial Intelligence and Data Engineering, Faculty of Engineering, Elazığ, TURKEY*
[b] *Department of Computer Engineering, Faculty of Engineering, Siirt University, Siirt, TURKEY*
** Corresponding author's e-mail address: mahmutkaya@firat.edu.tr*

DOI: 10.29130/dubited.1494138

ABSTRACT

Glaucoma is a disease that occurs after a certain age due to damage to the optic nerves. Today, machine learning methods can be successfully applied to detect such diseases. Instead of using the image data directly, the classification process is carried to a new representation space, positively affecting the classification performance. In this study, principal component analysis (PCA), linear discriminant analysis (LDA), and vision transformation (ViT) methods are used for feature extraction. In addition, the CLAHE filtering technique before ViT B16 was used in one of the proposed models. The classification process was performed with six different models using these methods alone or in combination, and the results are presented comparatively. Fine Tuned ViT-PCA-SVM and Fine Tuned ViT-LDA-SVM models achieved 92% classification success. As a result, the combination of ViT, which is a deep learning method, and PCA or LDA, which are machine learning methods, as feature extraction methods increased the classification success.

Keywords: *Glaucoma, Classification, Vision transformer, Principal Component Analysis, Linear discriminant analysis*

ViT Tabanlı Hibrit Öğrenme Yöntemleri ile Göz Tansiyonu Hastalığının Tespiti

ÖZET

Glokom, belirli bir yaştan sonra görme sinirleri üzerinde oluşan bir hasardan dolayı ortaya çıkan bir rahatsızlıktır. Bu tür hastalıkların tespitini yapmada günümüzde makine öğrenmesi yöntemleri başarıyla uygulanabilmektedir. Görüntü verilerinin doğrudan kullanımı yerine yeni bir temsili uzaya taşınarak sınıflandırma işleminin gerçekleştirilmesi sınıflandırma performansını olumlu etkilemektedir. Bu çalışmada öznitelik çıkartmada temel bileşen analizi (PCA), doğrusal ayırım analizi (LDA) ve görü dönüştürücü (ViT) yönteminden yararlanılmıştır. Ayrıca önerilen modellerden birinde ViT B16 öncesi CLAHE filtremele tekniği kullanılmıştır. Bu yöntemlerin tek başına veya bir araya getirildiği altı farklı model ile sınıflandırma işlemi gerçekleştirilmiş olup sonuçlar karşılaştırmalı olarak verilmiştir. Fine Tuned ViT-PCA-SVM ve Fine Tuned ViT-LDA-SVM modelleri %92 oranında sınıflandırma başarısı elde etmiştir. Sonuç olarak derin öğrenme yöntemi olan ViT ve makine öğrenmesi yöntemlerinden olan PCA veya LDA'nın öznitelik çıkartma olarak bir arada kullanıldığı yöntemler sınıflandırma başarısını arttırmıştır.

Received: 07/06/2024, Revised: 20/09/2024, Accepted: 02/10/2024

Anahtar Kelimeler: *Glokom, Sınıflandırma, Görü dönüştürücü, Temel bileşen analizi, Doğrusal ayırma analizi*

I. INTRODUCTION

Glaucoma is a disorder caused by damage to the main visual nerve (optic nerve), which has a significant proportion among the causes of vision loss in middle age [1]. The main factor in glaucoma is long-term increased intraocular pressure [2]. However, there are also factors, such as lack of blood supply to the optic nerve head, that cause this condition [3, 4, 5]. When the diagnosis and treatment of glaucoma are delayed, it can result in blindness [1]. In this respect, Glaucoma is an important condition that should be taken seriously. According to estimates in the literature, approximately 64.3 million people in the 40-80 age group were diagnosed with glaucoma worldwide in 2013, and this number is expected to reach 112 million by 2040 [14].

Fundus is a term used in ophthalmology and refers to the posterior region of the inner part of the eye. This area includes the retina, optic disc (the beginning of the optic nerve), and blood vessels on the posterior wall of the eye. The first fundus imaging was performed with devices available in the 1960s. Fundus examination is used for disorders related to eye vessel structures, neurological problems, and retinal-optic nerve problems that cause vision loss. Fundus imaging provides images of structures such as the optic nerve, retina, and eye vessels [6].

Using machine learning and deep learning methods in medical imaging and disease diagnosis is becoming increasingly important [7, 8]. Deep learning segmentation applications for muscle pathology image analysis [9], tumor growth prediction using convolutional neural networks [10], automatic classification and reporting of multiple common thoracic diseases using chest radiographs [11], and early detection of epidemics are just a few of the applications.

Many successful machine learning methods exist for both classification and clustering problems in obtaining meaningful results from medical data [12]. However, various data transformation methods can be applied to these data before using classification and clustering methods. Data transformation can significantly improve classification performance by extracting more meaningful information from the data [13]. In this sense, various machine learning methods have proven their success in the literature. These transformations include principal component analysis (PCA) and linear discriminant analysis (LDA). Recently, deep learning methods have been used on large datasets, going beyond the disease diagnosis experiments of computer technologies and providing results with high accuracy rates. Vision transformer models also achieve very successful results for image data in the literature.

In this study, vision transformer b16 model (ViT), principal component analysis (PCA), and linear discriminant analysis (LDA) methods were used together and separately to obtain new features for the detection of glaucoma disease using a large dataset consisting of 17,292 fundus images from different hospitals. SoftMax or SVM was used in the proposed models to perform the classification process after feature extraction.

In our study, we used a hybrid model that we believe has not adequately detected glaucoma disease from model fundus images. The hybrid model combines the strengths of classical image processing techniques and deep learning algorithms, allowing us to analyze glaucoma findings in fundus images more successfully.

II. LITERATURE REVIEW

A summary of the studies in the literature is given below, including different approaches to detect Glaucoma. Some studies classified fundus images using transfer learning methods in convolutional neural networks. In the second group of studies, other clinical data were used instead of fundus images, and classification studies were performed using machine learning models such as SVM. In the third group of studies, the authors developed deep learning models and trained them with fundus images. In the fourth group of studies, segmentation data alone or with fundus images were used for classification.

In their study, Şatır et al. performed data reduction with the rough sets method using the data of 168 individuals, followed by classification using decision trees and artificial neural networks. The study used parameters such as intraocular pressure, central corneal thickness, disc area, and age. When decision trees were used in the study, a success rate of 93.45% was achieved. In the classification using artificial neural networks, a success rate of 91.67% was achieved [1].

Yıldırım and Ozbay worked with AlexNet, ResNet-18, VGG16, SqueezeNet, and GoogleNet convolutional neural network models using 1000 images of people obtained from the open-source Origa (-light) dataset. In the study, the results were compared with many metrics. In terms of accuracy metric, the best value was obtained from GoogleNet with 97.98%, and the lowest value was obtained from the SqueezeNet model with 93.43% [14].

Uçar used a dataset of 4854 fundus images in his study. He used a transfer learning method with VGG16, Inception-V3, EfficientNet, DenseNet, ResNet50, and MobileNet architectures. He compared model performance with different validation metrics and cross-validation. As a result of the study, the models gave similar results. The highest accuracy value was obtained from the DenseNet architecture, with 96.19%. The lowest value was obtained from the MobileNet architecture, with 92.79% [15].

Dey et al. first applied some preprocessing to 100 fundus images, and feature extraction was performed using principal component analysis (PCA). The outputs were subjected to binary classification using support vector machine (SVM). The RBF kernel was used when applying SVM, and the gamma parameter was set to 1.5. The study, 10-fold cross-validation was used, and success was evaluated using different metrics. As a result of the study, 86% accuracy was achieved [16].

Karrothu et al. preprocessed the fundus images obtained from the Rim-One-R2 dataset, consisting of 248 images, by applying a CLAHE filter and then trained the VIT model to detect disease. Depending on whether the fundus images are left-eye or right-eye images, cropping was applied from the appropriate position and left the parts with the optic disc. As a result of the study, the success rate of 95.7% was obtained in CLAHE-applied images and 91.4% in unapplied images [17].

In the Wu et al. study, an SVM model was trained with 10-fold cross-validation using 114 OCT features and three clinical features (age, gender, and refraction) divided into nine groups in which values such as average thickness of the retina and nerve fiber layer were determined region by region from 752 Spectralis optical coherence tomography (OCT) data obtained from a hospital in Taiwan and performance evaluation was performed with various metrics. In this study, each OCT feature and three clinical features (age, gender, refraction) were evaluated for their association with glaucoma using the mutual information method. From these features, the top 20 features with the highest performance according to their mutual information values were selected and trained with the SVM model. The best-performing features were continuously added, and this process continued iteratively until ten features were selected. Only features that improved the model's performance compared to the previous steps were considered for the final subset. As a result of the study, with 96% accuracy in the advanced disease group and 92% in the intermediate disease group, the highest success rate was obtained, while a 73% success score was obtained in the early stage. [18].

In Phasuk et al.'s study, the network considers disease image information at two levels: the global image and the local disk region. In this paper, the DenseNet-121 model was used to learn the global image, and then a second part was used to segment the disc and cupping region using a residual deconvolution neural network. The ResNet-50 architecture was utilized in the second part. Horizontal translation and rotation were used for data augmentation to ensure generalization. In some neural networks, a maximum filter was applied to remove the blood vessel image. In addition, the contrast-limited adaptive histogram equalization (CLAHE) method was also used to increase the contrast in the fundus image and prevent over-amplification noise. The results show that the data augmentation used affects the training time. Two streams were used to train the network in the optical disk region. After the local disk region is cropped, the local disk and cup region are labeled using segmentation. These data, which have the same structure as the network in the global image, were then imported into the Densenet-121 model. The second flow focuses on the polar coordinates of the optical and cup structures. Data augmentation was achieved with polar center shift and polar radius parameters in this part. The final classification was performed with a classical artificial neural network. As a result of the study, a 94.0% success rate was obtained [19].

Li et al. used 39,745 image data for classification in their study. After dividing the dataset into training and test data, the data was normalized and resized to 299x299, and local area averaged color subtraction was applied to ensure color stability. Horizontal shifting and random rotation operations were performed for data augmentation. Inception-v3 architecture was utilized in the study. The images misclassified as a result of the training were also examined by experts, and it was found that many of the patients had different eye problems. A high AUC value of 0.986 was obtained in the study [20].

Fu et al. derived their M-Net model from the U-Net architecture. In this architecture, an image is taken, and a polar transformation is applied to the optical disk region. The output is sent to a series of convolution layers. The model outputs maps segmenting the optic disc and optic cup. M-Net is a network designed explicitly for multi-label segmentation in fundus disease images. An automated method is used to localize the optic disc region. Then, the original fundus image is transformed into a polar coordinate system based on the detected disc center. The M-Net model has side outputs, and combining these outputs provides the final result. The study was performed using ORIGA and SCES datasets containing 2326 images, including 214 glaucoma-positive images, and an AUC value of 0.89 was achieved [21].

The DENet model, designed in Fu et al.'s study, performs classification by considering the fundus image's global and local disk region. This architecture includes four streams. These are the global image stream that produces the result based on the whole fundus image, the segmentation guidance network that detects the optic disc region and provides the disc-segmentation representation by estimating a probability, the disc region stream that works on the disc region cropped by the disc-segmentation map obtained from the segmentation guidance network, and the disc polar stream that transfers the disc region image to the polar coordinate system. The outputs of these four streams are then combined into the final glaucoma screening result. ResNet and U-Net architectures were used to realize these streams. The study used three datasets, ORIGA, SCES, and SINDI, containing 8109 images, 327 of which were glaucoma-positive, resulting in an AUC of 0.91 [22].

When the studies in the literature are examined, it is seen that very successful models have been obtained. However, the success rates of these models have not yet reached the success and stability that can be used in medical applications. This shows the necessity of artificial intelligence studies on glaucoma disease. In our study, we have made significant contributions to the subject based on this need identified in the literature and the health sector, and a successful model has been put forward with a hybrid model we have developed with a different approach to solving the problem. The multi-layered data analysis in fundus images of our hybrid artificial intelligence model brings an important innovation to the knowledge in the field by using the hybrid artificial intelligence model approach, which has not been sufficiently studied in the literature in diagnosing glaucoma.

III. MACHINE LEARNING AND DEEP LEARNING METHODS

A. VISION TRANSFORMER (VIT)

Vision Transformer is one of the deep learning architectures used in image processing. While the Transformer architecture was originally developed for natural language processing (NLP) tasks, ViT is one of the first major works to apply this architecture to image processing tasks [23]. Figure 1 shows the basic structure of a vision transformer model.

Vision Transformer divides images into small patches and processes these patches as an array. Each patch is treated like words in natural language processing. The attention mechanisms of the Transformer architecture then process the patches. The attention mechanisms allow it to focus on the relationships between different image regions, allowing the model to understand which regions are more important. While CNNs focus on local features, ViT can model the overall connections between all image regions. When trained on large amounts of data, ViT goes beyond the limitations of Convolutional Neural Networks (CNNs), which focus on local features, by breaking images into small chunks, processing these chunks as a sequence, and modeling the overall connections between regions through attentional mechanisms; therefore, ViT can achieve more impressive results on large datasets.

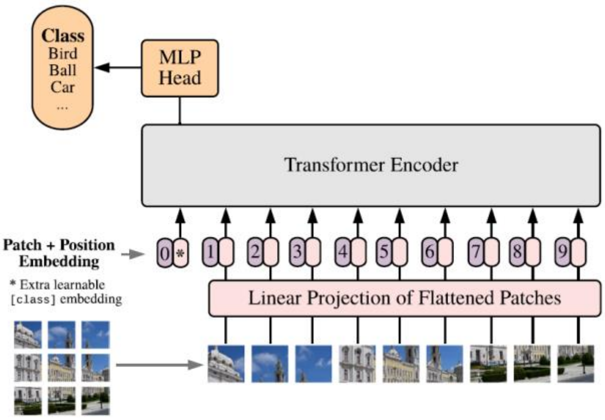

Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

***Figure 1.** Model Overview of Vision Transformer [23]*

B. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a statistical method that simplifies data by identifying the most important features in data sets [24]. Its main purpose is to find the principal components that maximize the variance in the data set and reduce its size by multiplying the data by a matrix of these components. The resulting new data set is usually smaller than the original dataset, discarding redundant information and noise while retaining the most important features. The implementation of PCA involves first centering the data around the mean value, then calculating the covariance matrix, finding the eigenvalues and eigenvectors of this matrix, and finally projecting the data onto these eigenvectors. This method reveals hidden structures and relationships, especially in large data sets, to facilitate visualization and make the data more understandable. PCA has been applied in many fields, such as face recognition, image processing, genetics, and financial analysis. In the PCA method, the calculation steps are done in 5 steps as follows;

First, the data is centralized by extracting the mean of each feature. If our data matrix X is of size n×p, where n is the number of samples and p is the number of features, we calculate the mean $\tilde{x}$ for each feature and subtract this mean from the data matrix

$$\mathrm{\bar{x}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} x_i$$

$$X' = X - 1_n\bar{x}$$

Here, (1_n) is a unit vector of length n, and Xl denotes the centralized data matrix. In the second step, we calculate the covariance matrix of the centralized data matrix. Denote the covariance matrix by it is calculated as:

$$\Sigma = \frac{1}{n-1}(X')^T X'$$

In the third step, we need to find the eigenvalues and eigenvectors of the covariance matrix. With eigenvalues λ and eigenvectors v and Σ being the covariance matrix:

$$\Sigma v = \lambda v$$

To find the eigenvalues and corresponding eigenvectors of the covariance matrix. The fourth step is to select the Principal Components. We sort the eigenvalues from largest to smallest and select the eigenvectors usually associated with the largest eigenvalues. These selected eigenvectors transform the data into a less dimensional space.

In the final stage, we transform the data by multiplying it by the selected eigenvectors. Denote the transformed data matrix by X_{new}:

$$X_{new} = X'W$$

W is a size p x k matrix formed by the selected eigenvectors, and k is the number of principal components selected.

C. LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA is a dimensionality reduction and classification technique used to classify samples from a dataset into classes [25]. Its main goal is to find a transformation that maximizes the separation between different classes. This process aims to maximize the distance between classes while minimizing the distance between instances belonging to the same class. LDA is beneficial for classification and data analysis. This is because it emphasizes the separation between classes, helping to achieve more accurate classification results. Unlike PCA, it is a supervised method and uses class labels. LDA analyzes how the data is distributed to distinguish specific classes. Therefore, it is more suitable for classification processes.

In LDA, the intra-class and inter-class scatter matrices are first calculated. The intra-class scatter matrix shows how the samples of the same class are distributed concerning the class average, while the inter-class scatter matrix shows how different classes differ from the overall average. These matrices are then used to find the discriminant axes that maximize class separation. Dimensionality reduction is performed by projecting the data onto these axes. LDA is used in supervised learning scenarios and has applications in fields as diverse as face recognition, biological data analysis, and finance.

The computation in the LDA method can be divided into several stages. Let X represent the features in a dataset. Let each sample contain an array denoted by x_i divided into c classes.

Let x_i be the sequence of each sample and, c be the classes, μ_k be the mean vector of each class is calculated as;

$$\mu_k = \frac{1}{N_k}\sum_{i \in C_k} x_i$$

N_k denotes the number of samples in the kth class, and C_k is the set of all data samples belonging to the kth class.

The mean vector μ for all data is calculated as follows.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where N is the total number of samples, the within-class variance matrix S_W is calculated as follows.

$$S_W = \sum_{k=1}^{c}\sum_{i\in C_k}(x_i - \mu_k)(x_i - \mu_k)^T$$

In this way, the variance of the samples within each class is found. The between-class variance matrix S_B is calculated as follows.

$$S_B = \sum_{k=1}^{c} N_k(\mu_k - \mu)(\mu_k - \mu)^T$$

Thus, it finds the separation between classes. LDA finds the best linear projection to separate classes.

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

This ratio represents the ratio of between-class variance to within-class variance. The w that maximizes this ratio represents the best linear discriminator. To solve this problem, we solve the eigenvalue problem, usually expressed as follows.

$$S_W^{-1} S_B w = \lambda w$$

Here λ is the eigenvalue, and w is the corresponding eigenvector. These eigenvectors are used as separating lines for the linear projection of the data. After this step, the eigenvectors corresponding to the m largest eigenvalues are selected. A projection matrix $W = [u_1, u_2, \dots, u_m]$ is created with the selected eigenvectors. The input vectors are projected into a lower dimensional space by applying $X' = XW$ to this projection matrix. The projection reduces the dataset size while increasing the separation between classes. X is the original data matrix, and X' is the projected data matrix.

D. SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines (SVM) is a powerful and flexible machine learning algorithm that is one of the supervised learning models used for classification tasks. It performs particularly well in high-dimensional data spaces and marginally separable data sets. The basic principle of SVM is to find the best decision boundary (hyperplane) to classify the examples in the dataset into classes. This decision boundary is chosen to maximize the marginal distance between instances belonging to different classes. The examples that are closest to the boundary and determine the position of this boundary are called “support vectors.” Figure 2 shows a simple SVM implementation in two dimensions. The figure shows the optimal hyperplane, the optimal margin, and the data points in red and blue. The red and blue data points represent different classes, and the optimal hyperplane is positioned to maximize the separation between these two classes.

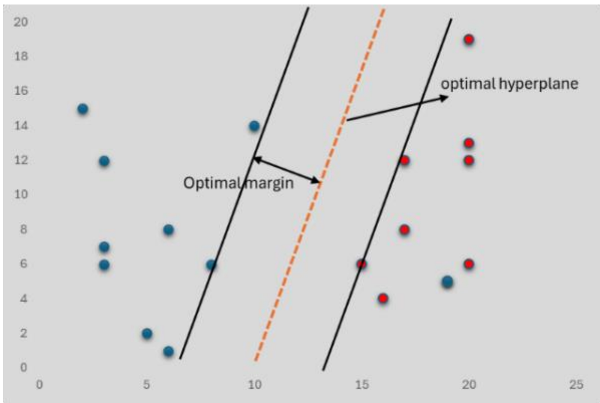


Figure 2. A simple two-dimensional SVM example.

SVM uses the kernel trick technique to process data sets that cannot be linearly separated. This technique transforms the data points from the original feature space into a higher-dimensional space, allowing linear separation in the new space. Standard kernel functions include polynomial, radial basis function (RBF), and sigmoid functions. The SVM algorithm is known for its high accuracy in many situations and its ability to create highly generalizable models. Therefore, it is successfully used in many fields.

E. CONTRAST LIMITED ADAPTIVE HISTOGRAM EQUALIZATION (CLAHE)

CLAHE is a method for improving contrast in images. It is particularly effective in low-contrast images and is widely used in medical imaging, satellite imagery, and photography.

CLAHE is based on the Histogram Equalization method. Histogram Equalization increases the contrast of an image by expanding its histogram, i.e., its color distribution, over the entire image. However, in some cases, it can result in excessive contrast enhancement and the appearance of noise. To solve this problem, CLAHE operates in a contrast-limited and adaptive way. CLAHE limits contrast increases above a certain threshold to prevent too high contrast enhancement. This prevents excessive increase of noise in the image. In addition, the image is divided into small regions to enable adaptive processing. Histogram equalization is applied separately for each region. Thus, a more balanced contrast enhancement is applied by considering the regional contrast differences of the image. Figure 3 shows an original fundus image from the dataset after the CLAHE application. Figure 3 clearly shows the positive effect of CLAHE application on contrast.

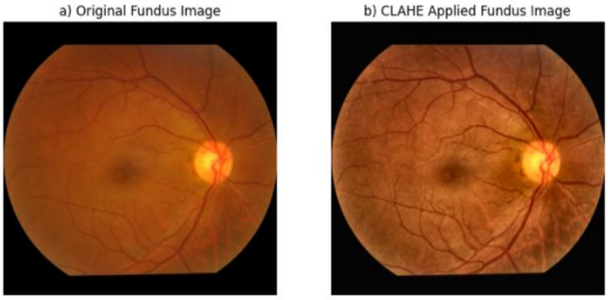


Figure 3. *CLAHE filter applied to RGB fundus image. a) Original image, b) CLAHE applied fundus image*

IV. MATERIAL AND PROPOSED MODELS

In this study, glaucoma disease was detected by classification with hybrid models. For this purpose, ViT, PCA, LDA, CLAHE, and SVM methods were utilized.

A. DATASET

SMDG-19 open-source dataset was used in our study. Information about this dataset is shown in Table 1.

Tablo 1. *SMDG-19 dataset*

Data	**Number of Glaucoma Positive Images**	**Number of Glaucoma Negative Images**	**Total**
Number of training samples	3328	5293	8621
Number of validation samples	2208	3539	5747
Number of test samples	1120	1754	2874
Total	6656	10586	17242

Depending on the dataset's size, the model's requirements, and the results of the preliminary studies, the dataset was divided into 50% training, 33% validation, and 17% test data. The original images in the dataset are at a standardized scale of 512x512. The images were obtained by merging and standardizing 19 different open-source datasets. Figure 5 shows a sample of glaucoma-positive and glaucoma-negative images from the dataset.

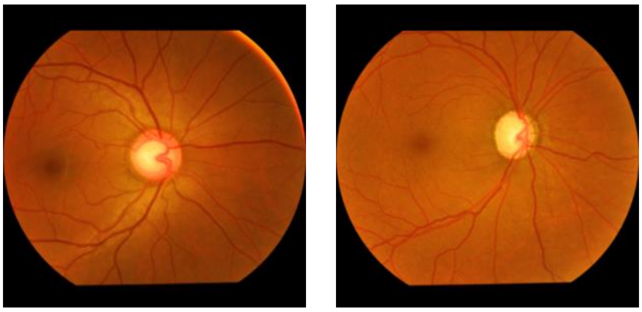


Figure 5. *Sample Glaucoma negative (left) and Glaucoma positive (right) images from the SMDG-19 dataset.*

B. DATA PREPROCESSING

One of the most important steps to increasing the success of deep neural network methods is to train the models with as much data as possible. Increasing data provides better learning and increases the model's reliability. For this reason, data augmentation was performed on the dataset before training the data. The data augmentation process can have different perspectives depending on the models applied, and each model, and the data augmentation techniques used in these models are given in Table 2.

Table 2. *Preprocessing and data augmentation for models.*

Model	Image size	Preprocessing and Data Augmentation
Fine Tuned ViT	224x224	Random horizontal and vertical flip
CLAHE-Fine Tuned Vit	224x224	CLAHE filter on the green channel, Random horizontal and vertical flip
Fine Tuned ViT-PCA-SVM	224x224	Random horizontal and vertical flip
Fine Tuned ViT-LDA-SVM	224x224	Random horizontal and vertical flip
ViT-PCA-SVM	224x224	Random horizontal and vertical flip
Incremental PCA - SVM	224x224	-

The SMDG-19 dataset used in the study increases the generalizability of the model. It is expected to provide a high degree of relevance to clinical applications thanks to its high number of images and the application of various preprocessing techniques. The large scope and standardization of the dataset increases the likelihood of adequately representing variations in real-world conditions, which aims to strengthen the expectation of supporting the clinical validity of the model.

C. VIT B16 MODEL

In our study, transfer learning was mainly performed on the pre-trained weights of the ViT B16 model using the fundus images in our dataset. To see the difference, we directly used the pre-training weights of the ViT B16 model without transfer learning in our ViT-PCA-SVM model. In our classification and transfer learning studies, the final layer of the model was set to have a softmax activation function, and a dense layer with two outputs was utilized. The images were first preprocessed and then trained with the specified parameters. A summary of our ViT model is given in Table 3.

Table 3. *ViT Model Summary*

Layer (type)	Output Shape	Parameter Count
Input Layer	(224, 224, 3)	0
Vit-b16 (Functional)	768	85798656
Flatten	768	0
***Dense (Feature)**	64	49216
Dense	32	2080
Dense	2	66

D. PROPOSED MODELS

In this study, we first focus on feature extraction and then classification of fundus images. Feature transformation methods have provided outstanding solutions for moving the data to a better representation space. After feature extraction, SoftMax or SVM methods were used to predict the data category.

This study tried six different hybrid models, with the aim of obtaining the best result by combining different models. Figure 4 shows the structure of the hybrid models used.

In the first model, Fine-Tuned ViT and PCA methods were used for feature extraction, while SVM was used for classification. In the second model, Fine-Tuned ViT was first used for feature extraction. Then, a new transformation was made with LDA, and in the last stage, SVM was used for classification. In the third model, the CLAHE approach was first used on fundus images, and then the Fine-Tuned ViT model was preferred.

The data was trained and classified in the fourth model with Fine Tuned Vit B16. In the fifth model, data transformation was performed with Incremental PCA, and then SVM was used for classification. Finally, in the sixth model, feature extraction was performed on the fundus images first with ViT, then with PCA, and then with SVM classification.

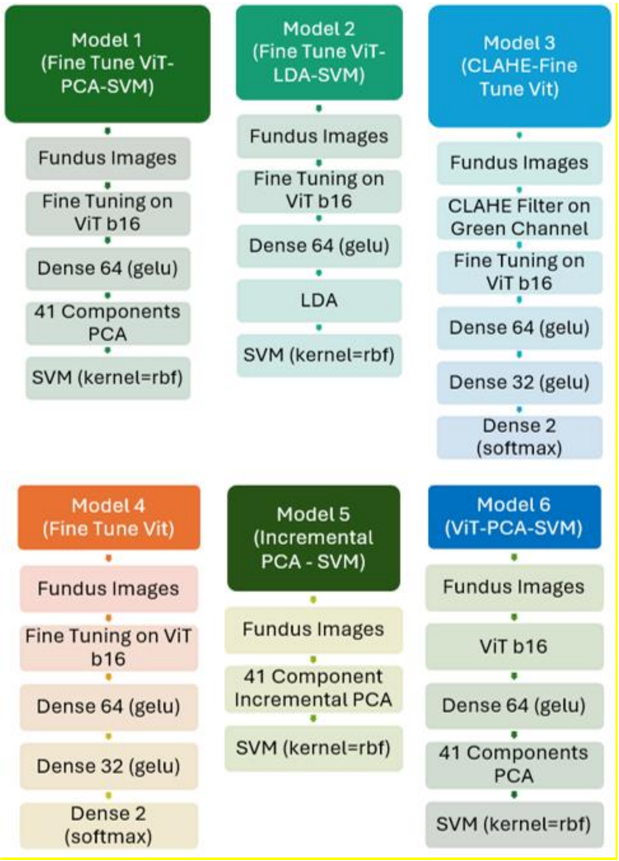


Figure 4. *Used model structures.*

E. PARAMETER SETTINGS

Parameters suitable for the structure of the model were selected for the models we used in our study. In the training of ViT models, Sparse Categorical Crossentropy was used as the loss function, and Adam with Weight Decay (AdamW) was used as the optimization algorithm. In the classification layer of ViT models, softmax was used as the activation function. The ViT models' batch size was 16, and the models were trained in 5 epochs. In the model with Incremental PCA, 256 was used as a batch size. Since the PCA method is based on statistical calculations, large batch sizes are preferred to analyze high-dimensional data together and increase efficiency. In contrast, smaller batch sizes are used because ViT models provide better generalization and computational efficiency with smaller batches. Radial Basis Function Kernel (RBF) was used as Kernel in all SVM classifier models. The RBF kernel was chosen because it can learn efficient separation boundaries in high-dimensional data space and provides flexibility and robust performance for modeling complex relationships between data points. The number of components was chosen as 41. In cases where LDA was used as a classifier model, the number of components was set to “1” since the number of classes was 2.

F. PERFORMANCE EVALUATION METRICS

Our study used accuracy, precision, recall, and F1-score metrics to evaluate the classification performance. In addition, model performance was evaluated with the complexity matrix. Accuracy refers to the overall performance of the classification model. It shows how much of the total predictions are made correctly. It is usually used when the class distribution is balanced, and all classes are equally important. Mathematically, it is calculated as the ratio of correct predictions to total predictions, as in Equation 1.

Precision measures how many of the positively predicted samples are positive. In other words, it shows how “accurate” the model is. It is preferred when false positives (FP) are significant. Precision is preferred when correct positive predictions are more important than errors in the negative class. Its mathematical expression is as in Equation 2.

Recall measures how much of the samples that are actually positive are correctly predicted to be positive. It shows how well the model “remembers” positive cases. It is calculated using the expression in Equation 3.

The F1-Score is a metric that balances precision and sensitivity. It is calculated as the harmonic mean of both values and indicates the overall balance of the model. It is used when a trade-off between precision and sensitivity needs to be considered and is calculated using the expression in Equation 4.

The confusion matrix shows the performance of a classification model in detail. It presents the model's correct and incorrect predictions for each class in matrix form. It is used to examine the model's performance for each class in detail. Usually, the matrix rows represent the actual classes, and the columns represent the predicted classes. Four basic terms are usually used in the matrix.

True Positive (TP, True Positive): Situations that are positive and correctly predicted as positive.

False Positive (FP, False Positive): Situations that are actually negative but incorrectly predicted as positive.

True Negative (TN, True Negative): Situations that are negative and correctly predicted as negative.

False Negative (FN, False Negative): Cases that are actually positive but incorrectly predicted as negative. The representation of the complexity matrix is as in Table 4.

Table 4. *Confusion matrix.*

	Actual	
Predicted	TP	FP
	FN	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

V. EXPERIMENTAL RESULTS AND DISCUSSION

The plots of the accuracy and loss graphs at each epoch for the train and validation data of the 6 different models applied in our study are shown in Figure 6.

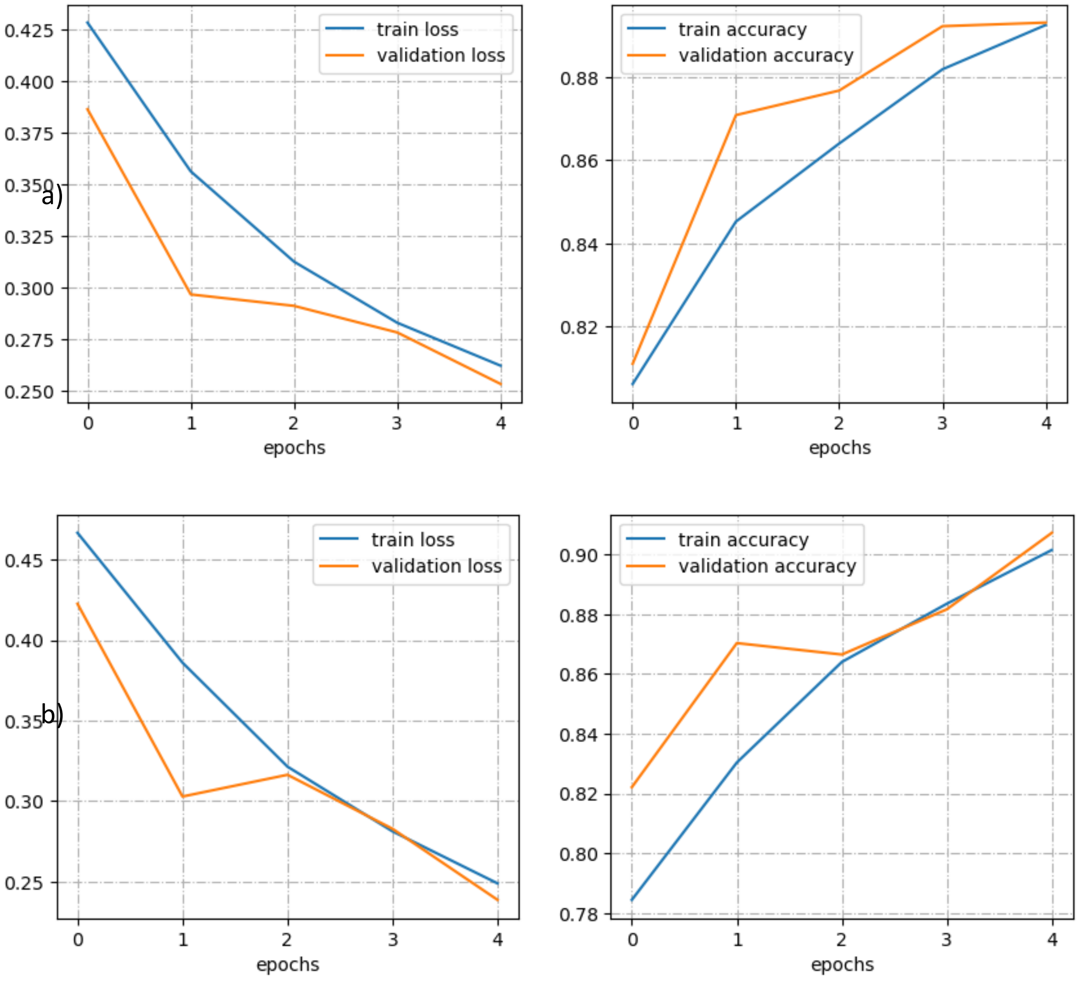
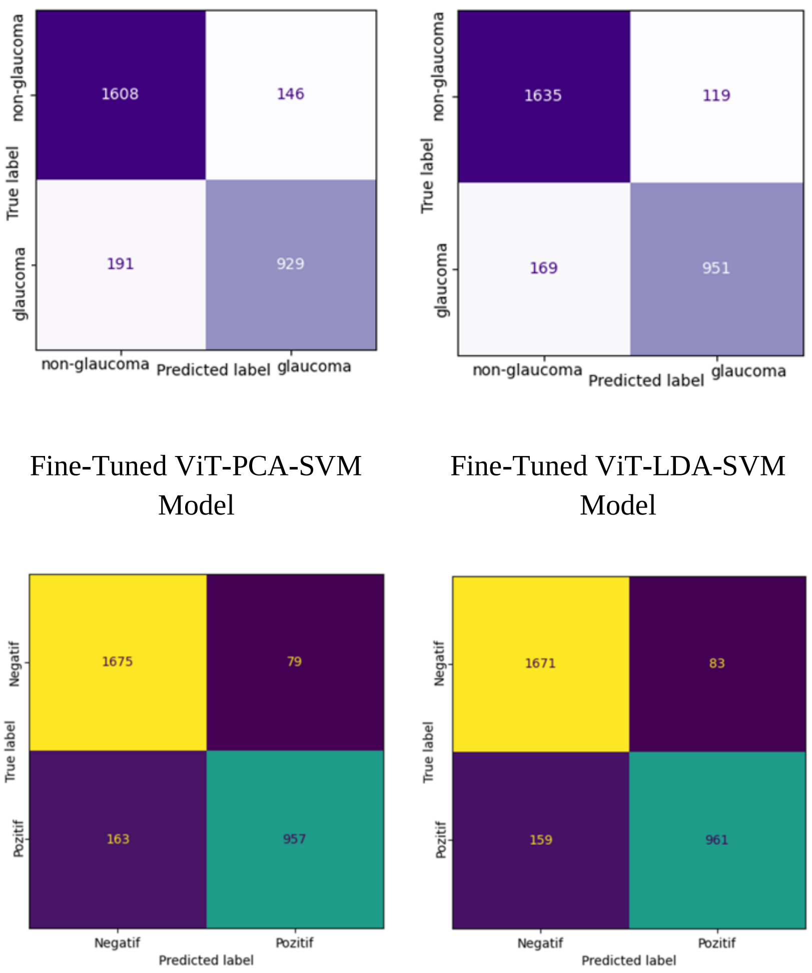


Figure 6. *Accuracy-epochs, Loss-epoch graphs of train and validation data. a) ViT fine tuning, b) ViT fine tuning after CLAHE filter.*

When the loss and accuracy graphs are analyzed, it is seen that the loss value decreases significantly for the train data and the accuracy value increases as expected as the epoch increases. However, this situation is not as clear for validation data. Since we train on pre-trained weights, the accuracy value rises above 84% after the first epoch. When Accuracy starts with a high value from the beginning, the changes in the subsequent iterations are not very significant. Still, there is a decrease in loss and an increase in accuracy between the initial and final values, as desired. The confusion matrixes of our models are given in Figure 7.

Fine Tuned Vit Model

CLAHE-Fine Tuned Vit Model



Fine-Tuned ViT-PCA-SVM Model

Fine-Tuned ViT-LDA-SVM Model

ViT-PCA-SVM Model

Incremental PCA - SVM Model

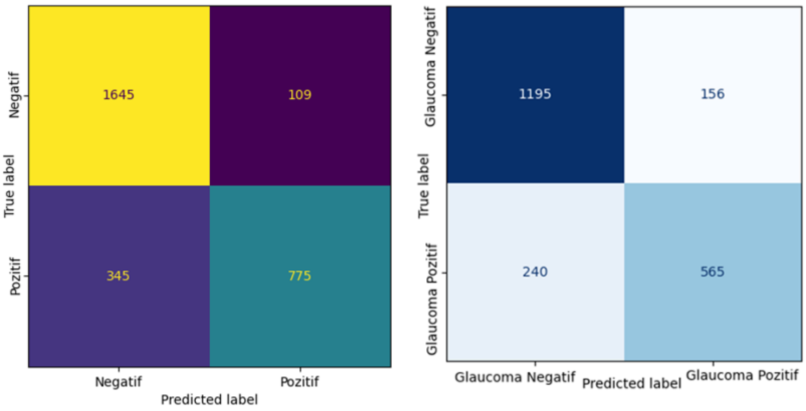


Figure 7. *Confusion matrices of the models used.*

The performance metrics calculated using the confusion matrixes are given in Table 5.

Table 5. *Classification performance of models.*

Method	Accuracy	Precision	Recall	F1- Score
Fine Tuned Vit	0.88	0.88	0.88	0.88
CLAHE-Fine Tuned Vit	0.90	0.90	0.90	0.90
Fine Tuned ViT-PCA-SVM	0.92	0.92	0.92	0.92
Fine Tuned ViT-LDA-SVM	0.92	0.92	0.92	0.92
ViT-PCA-SVM	0.84	0.85	0.84	0.84
Incremental PCA - SVM	0.81	0.83	0.88	0.86

When the results are examined, it is seen that the best score of 0.92 is obtained in the models that follow the process steps of feature extraction from ViT models trained with the transfer learning method, reduction of these features with LDA or PCA, and then classification with SVM. In the "Fine Tuned Vit" and "CLAHE-Fine Tuned Vit" models, where classification is performed with dense layers after transfer learning is applied, the scores increase when we apply the CLAHE filter. In the ViT-PCA-SVM model, where we classify directly with the results of the pre-trained model without applying the transfer learning stage, the scores are lower than the other ViT models, as expected. It is important to see the benefit of transfer learning.

In the case where the images are directly classified with SVM after data reduction with PCA without using deep learning methods, although the results are lower than the other methods, it is seen that it is close to the values in the case where transfer learning is not applied and even surpasses this model in sensitivity and F1 score. This indicates that using deep learning without transfer learning is not very beneficial. In the Incremental PCA SVM model, direct PCA could not be applied due to memory problems. Incremental PCA, where the data is processed piece by piece, had to be applied.

When other studies in the literature are examined, it is seen that most of the studies have worked with smaller data sets. In addition, it is seen that many studies use data from patients in a specific hospital [16], [17], [18], [19]. This will negatively affect the generalizability of the results obtained. In the dataset we used, images from different hospitals in different countries were used. In this respect, it can be said that the models we obtained are more generalizable. Our study obtained an 81.0% accuracy when we applied Incremental PCA and SVM without ViT. Dey et al. [16] applied a similar method in their study and achieved 86% accuracy on a smaller dataset.

In our study, the success scores increase slightly when the CLAHE filter is applied, which is in line with Karrothu et al.'s study [17], which shows the study's reliability. With a large dataset, this study achieved a high success rate of 92% in detecting glaucoma disease, providing strong potential for earlier and more accurate patient diagnosis in clinical applications. Future studies with larger data sets are expected to pave the way for the direct use of the model in clinical studies.

VI. CONCLUSION

This study used six models to detect glaucoma disease using the SMDG-16 dataset of 17242 fundus images. Different approaches were applied to our models, and the results were analyzed. These approaches show the effects of transfer learning and using the CLAHE filter in ViT models. We also show the effects of classification with a dense layer or using combinations of PCA/LDA-SVM on the performance. Finally, it is shown how success scores would change in the absence of deep learning.

The study's results showed that the highest scores were 92% for the Fine Tuned ViT-PCA-SVM and Fine Tuned ViT-LDA-SVM models. The lowest scores were seen in the Incremental PCA-SVM model without deep learning and the ViT-PCA-SVM model without pre-trained weights.

REFERENCES

[1] E. ŞATIR, F. AZBOY, A. AYDIN, H. ARSLAN, and Ş. HACIEFENDİOĞLU, “Veri İndirgeme ve Sınıflandırma Teknikleri ile Glokom Hastalığı Teşhisi,” El-Cezeri Fen ve Mühendislik Dergisi, vol. 3, no. 3, pp. 485–497, Sep. 2016, doi: 10.31202/ecjse.258576.

[2] Y. Yılmaz, A. Koytak, K. Erol, Y. Çınar, and Y. Özertürk, “Primer Açık Açılı Glokomda Fundus Floresein Anjiyografi,” Kartal Eğitim ve Araştırma Hastanesi Tıp Dergisi, vol. 17, no. 1, pp. 1–5, 2006.

[3] S. S. Hayreh, “Evaluation of Optic Nerve Head Circulation: Review of the Methods Used,” J Glaucoma, vol. 6, no. 5, pp. 319–330, 1997.

[4] J. E. Grunwald, J. Piltz, S. M. Hariprasad, J. Dupont, and M. G. Maguire, “Optic nerve blood flow in glaucoma: effect of systemic hypertension,” Am J Ophthalmol, vol. 127, no. 5, pp. 516–522, May 1999, doi: 10.1016/S0002-9394(99)00028-8.

[5] J. E. Grunwald, J. Piltz, S. M. Hariprasad, and J. DuPont, “Optic nerve and choroidal circulation in glaucoma.,” Invest Ophthalmol Vis Sci, vol. 39, no. 12, pp. 2329–2336, Nov. 1998.

[6] L. A. Yannuzzi et al., “Ophthalmic fundus imaging: today and beyond,” Am J Ophthalmol, vol. 137, no. 3, pp. 511–524, Mar. 2004, doi: 10.1016/j.ajo.2003.12.035.

[7] Kaya, Y., Yiner, Z., Kaya, M., & Kuncan, F. (2022). A new approach to COVID-19 detection from X-ray images using angle transformation with GoogleNet and LSTM. Measurement Science and Technology, 33(12), 124011.

[8] Şenol, A., & Kaya, M. (2024). An Investigation on the Use of Clustering Algorithms for Data Preprocessing in Breast Cancer Diagnosis. Türk Doğa ve Fen Dergisi, 13(1), 70-77.

[9] Z. Wang and M. Lemmon, “Stability analysis of weak rural electrification microgrids with droop-controlled rotational and electronic distributed generators,” in 2015 IEEE Power & Energy Society General Meeting, IEEE, Jul. 2015, pp. 1–5. doi: 10.1109/PESGM.2015.7286507.

[10] T. M. Lehmann, C. Gonner, and K. Spitzer, “Survey: interpolation methods in medical image processing,” IEEE Trans Med Imaging, vol. 18, no. 11, pp. 1049–1075, 1999, doi: 10.1109/42.816070.

[11] G. J. Grevera and J. K. Udupa, “An objective comparison of 3-D image interpolation methods,” IEEE Trans Med Imaging, vol. 17, no. 4, pp. 642–652, 1998, doi: 10.1109/42.730408.

[12] Şenol, A., Canbay, Y., & Kaya, M. (2021). Makine Öğrenmesi Yaklaşımlarını Kullanarak Salgınları Erken Evrede Tespit Etme Alanındaki Eğilimler. Bilişim Teknolojileri Dergisi, 14(4), 355-366.

[13] Utku, A., & Akcayol, M. A. (2024). Spread patterns of COVID-19 in European countries: hybrid deep learning model for prediction and transmission analysis. Neural Computing and Applications, 1-17.

[14] Ö. YILDIRIM and F. ALTUNBEY ÖZBAY, “Fundus Görüntülerinden Derin Öğrenme Teknikleri ile Glokom Hastalığının Tespiti,” European Journal of Science and Technology, vol. 44, pp. 1–6, Dec. 2022, doi: 10.31590/ejosat.1216404.

[15] M. UÇAR, “Glokom Hastalığının Evrişimli Sinir Ağı Mimarileri ile Tespiti,” Deu Muhendislik Fakultesi Fen ve Muhendislik, vol. 23, no. 68, pp. 521–529, May 2021, doi: 10.21205/deufmd.2021236815.

[16] A. Dey and S. Bandyopadhyay, “Automated Glaucoma Detection Using Support Vector Machine Classification Method,” Br J Med Med Res, vol. 11, no. 12, pp. 1–12, Jan. 2016, doi: 10.9734/BJMMR/2016/19617.

[17] A. Karrothu and A. Chunduru, “Glaucoma Detection Using Computer Vision and Vision Transformers,” International Journal of Computing and Digital Systems, vol. 14, no. 1, pp. 1–13, 2023.

[18] C.-W. Wu, H.-Y. Chen, J.-Y. Chen, and C.-H. Lee, “Glaucoma Detection Using Support Vector Machine Method Based on Spectralis OCT,” Diagnostics, vol. 12, no. 2, pp. 391–406, Feb. 2022, doi: 10.3390/diagnostics12020391.

[19] S. Phasuk et al., “Automated Glaucoma Screening from Retinal Fundus Image Using Deep Learning,” in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Jul. 2019, pp. 904–907. doi: 10.1109/EMBC.2019.8857136.

[20] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, and M. He, “Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs,” Ophthalmology, vol. 125, no. 8, pp. 1199–1206, Aug. 2018, doi: 10.1016/j.ophtha.2018.01.023.

[21] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, “Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation,” IEEE Trans Med Imaging, vol. 37, no. 7, pp. 1597–1605, Jul. 2018, doi: 10.1109/TMI.2018.2791488.

[22] H. Fu et al., “Disc-Aware Ensemble Network for Glaucoma Screening From Fundus Image,” IEEE Trans Med Imaging, vol. 37, no. 11, pp. 2493–2501, Nov. 2018, doi: 10.1109/TMI.2018.2837012.

[23] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv, vol. 11929, Oct. 2020.

[24] M. Turk and A. Pentland, “Eigenfaces for Recognition,” J Cogn Neurosci, vol. 3, no. 1, pp. 71–86, Jan. 1991, doi: 10.1162/jocn.1991.3.1.71.

[25] K. Etemad and R. Chellappa, “Discriminant analysis for recognition of human face images,” Journal of the Optical Society of America A, vol. 14, no. 8, p. 1724, Aug. 1997, doi: 10.1364/JOSAA.14.001724.