Power System Protection and Control
ISSN:1674-3415

OCULAR DISEASE CLASSIFICATION BY USING GLOBAL-LOCAL MULTI-LABEL CLASSIFICATION NETWORK

Hamed Tayebi[1], Mehdi Jafari Shahbazzadeh*[1], Mahdiyeh Eslami[1]

1Department of Electrical Engineering, Ke.C., Islamic Azad University, Kerman, Iran.
Corresponding Author Email: mjafari@iau.ac.ir

Received:26 September 2025 **Revised**:19 October 2025 **Accepted**: 22 November 2025

***ABSTRACT*:**

The timely detection of ocular fundus disease has emerged as a crucial concern in recent decades due to the lack of early-stage symptoms and the potential for severe effects such as blindness. In this context, it is imperative for physicians to accurately diagnose ocular fundus disease during its first phases, a task that is both labor-intensive and demanding. Various machine learning algorithms are employed to provide assistance to doctors through the implementation of helpful actions. Therefore, this study presents an innovative methodology for the classification of fundus pictures, encompassing eight distinct groups that include various pathologies as well as normal cases. The proposed model, referred to as the Global-Local Multi-Label Classification Network, incorporates both global and local perspectives in the learning process. This is achieved by leveraging a vision transformer for global understanding and employing convolutional layers for local analysis. The performance of this model is assessed using the ODIR-5K dataset, and the findings indicate that it outperforms existing state-of-the-art approaches across several measures.

***Keywords**: Classification, Fundus images, Vision Transformers, Deep learning, Transfer learning.*

INTRODUCTION

A large number of people around the world suffer from vision problems or blindness; among them, many could have been prevented or need to be addressed before it becomes too late [1]. The most common vision impairment is a refractive error, such as myopia, distinguished by the need for glasses [2]. However, the primary reason for vision loss or blindness is fundus diseases. They must be diagnosed in the early stages because their development to the late stage makes the treatment expensive and very hard or even impossible [3]. A *cataract* is a common fundus disease, a clouding of the eye's lens, sometimes by patches. Cataracts can occur when the optic disc, fovea, and other eye parts become hazy. It is treated by surgery to implant intraocular lenses. Unfortunately, this surgery is not available in some countries. Two other fundus diseases are diabetic retinopathy (DR) and age-related macular degeneration (AMD). Today, diabetes is widespread in the world, and DR is the most common complication of it. DR influences the tissues inside the retina. It has no obvious abnormal symptoms in the early stages, and patients may have only minor issues related to clear vision, but this will eventually lead to blindness. DR is one of the four major causes of loss of eye vision [4]. In aging communities, AMD is becoming the major blinding eye disease [5]. The macula is liable for clear central vision, and the distortion in vision starts if liquids collect in it. AMD can be distinguished by noticing fresh blood vessel growth or dead retinal cells, known as neovascularization and geographic atrophy, respectively [6]. It will rapidly cause irreversible visual impairment and has no timely and effective treatment. Glaucoma is another eye condition that harms the optic nerve due to unusually high pressure in the eye. It can be distinguished by noticing changes in the proportion of the optic disc cup and neuro-retinal edge surface region, known as the cup-to-disc ratio. Patients are often unaware of the progression of glaucoma because they get accustomed to their reduced range of view [7]. If left untreated, it can lead to blindness. Hypertension is a silent illness. This disease changes the biological shapes of veins, like length and thickness, and results in cardiovascular disease, stroke, and respiratory failures over the long run [8]. Fundus diseases such as AMD, DR, and glaucoma cause blindness to more than 10 million people worldwide each year [8]. Indeed, glaucoma is the second most common cause of blindness in developed countries [9], while AMD is the most common cause of blindness in people above 50 years old [10]. Moreover, the projection of change in vision loss from the year 2020 to 2050 shows that there could be a nearly 55% increase in the number of visual impairments [11] [12]. In Figure 1, some samples of simulated images are shown. The National Eye Institute, an institute of the National Institute of Health of the USA, has created these images to understand eye disease perception better [13].

DOI: 10.46121/pspc.53.4.13

Power System Protection and Control
ISSN:1674-3415

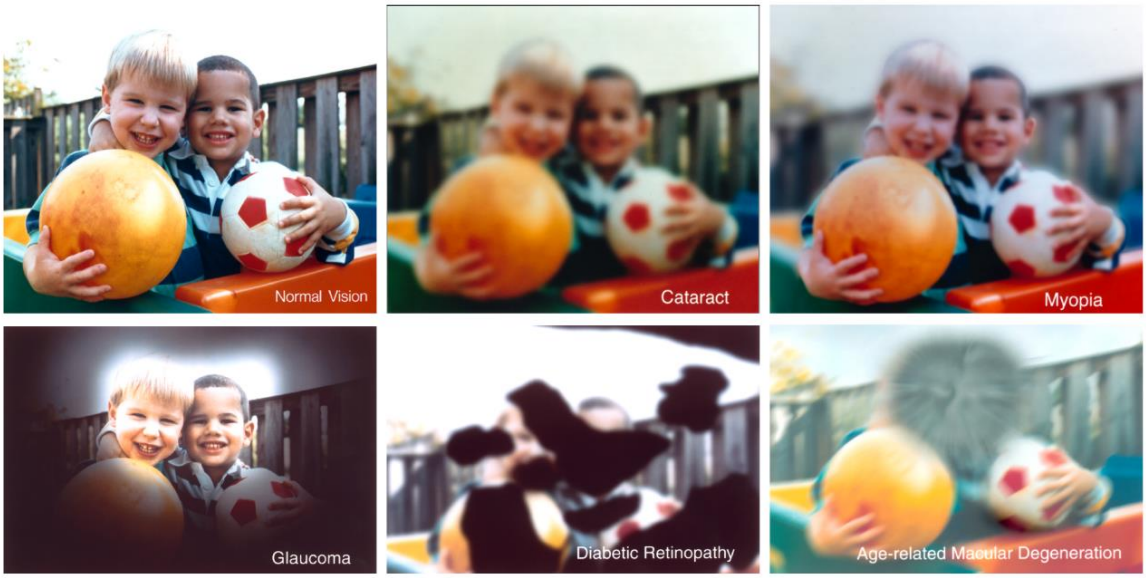


Figure 1: Samples of simulation of eye diseases perception [13]

Before any symptoms, early diagnosis of fundus diseases may be performed by regular examination of the retina. The preliminary method for retinal examination is color fundus photography. These images are medical images of the retina acquired using an ophthalmoscope. They can directly observe retinal vasculature, macula, and optic disc [12]. Fundus images provide effective measures for accurate diagnosis and treatment of eye diseases. Using artificial intelligence (AI) for medical image analysis can assist the use of fundus images for diagnosing and treating eye diseases. Using these technics reduces the time to process large datasets and minimizes variability in image interpretation. Moreover, compared to manual inspection, AI is more rapid, cost-effective, objective, and reliable and does not need trained specialists to grade images [14]. AI can help ophthalmologists make accurate diagnoses based on comprehensive medical data and provide new strategies to improve the diagnoses and treatments of eye diseases.

RELATED WORKS

Many previous fundus image classification studies have concentrated on identifying a single disease without considering other eye disorders, such as glaucoma [15], diabetic retinopathy [16], age-related macular degeneration [17], and myopia [18]. However, several researchers have made significant efforts to resolve the multi-label classification problem of ophthalmic diseases when a patient has more than one disorder, and these different disorders may influence each other. These proposed approaches have been evaluated on the ODIR-5K dataset.

In this section, studies of multi-label eye disease classification have been investigated. For instance, a shallow CNN-based model is proposed by Islam et al. [19] that has converted the multi-label classification problem into multiclass classification. They applied a CNN network to classify fundus images of the ODIR-5K dataset. The input of the CNN model is left and right eye fundus images separately, and the label is assigned according to each image. Although this approach made the disease classification model simpler, the model cannot distinguish multiple diseases. Wang et al. [20] have proposed an ensemble model for fundus image classification, which contains two parallel EfficientNet models, where feature concatenation is done at the last layer for final classification. They also apply a preprocessing stage on fundus images, using gray and color histogram equalization. VGG-16, ResNet, Inception-v4, and Densenet architectures [21] are applied in [14] for the classification task. Li et al. have performed sum, multiply, and concatenate operations on features extracted from the pre-train baseline model. It is found that element-wise sum operation on feature maps gives better disorder identification than the other techniques. Hasan et al. [22] used four different neural networks to detect cataracts, including InceptionV3, InceptionResNetV2, Xception, and Densenet121. They used the ODIR dataset and implemented binary classification for images labeled with cataracts. Kumar et al. [23] developed a structure using ensemble pre-trained CNN models to perform multi-label classification. Some research, such as [24] [25], found

DOI: 10.46121/pspc.53.4.13

that applying CNNs yields high accuracy for eye disorders classification. For example, they showed that transfer learning with ImageNet pre-trained models, such as the Inception network, was very effective. Another approach is an ensemble Approach. The main idea of this method is to apply two or more models learned from different data sets. Finally, each model predicts test data and the eventual output is obtained from voting among different predictions. The main disadvantage of this approach is that it requires substantial time and resources for training [26]. For instance, Yanga et al. [27] automatically adopted ensemble learning to detect cataracts. In another research, Sandhya et al. [28] proposed an approach based on ensemble learning for diabetic retinopathy detection.

In recent years, Convolutional Neural Networks (CNNs) have demonstrated superiority in various tasks such as classification, segmentation, and other vision-based works. They could concentrate on spatial features and apply them to their tasks. CNNs have proven their superior performance on various benchmarks. However, they have some disadvantages regarding the learning process and model architecture for image processing. One of these disadvantages is the need for a global understanding of the images and focus on local features due to the restricted receptive field. Recently, researchers have worked on a novel architecture to overcome these drawbacks, originating from Natural Language Processing (NLP) research.

Research on NLP led to the development of the transformer network in 2017 by Vaswani et al. [29]. Investigations on this novel network showed a considerable improvement in NLP by using it, making the transformer network one of the main network architectures in this field. Other approaches were developed based on further research on this network. For example, the BERT transformer network was proposed by Devlin et al. in 2018 [30]. Studies on these networks encouraged other researchers to use them in image analysis and machine vision, believing they can improve in these fields. In 2020, a paper was published titled: "An image is worth 16x16 words" [31]. By presenting their results in this seminal publication, Dosovitskiy et al. showed that the application of transformer networks could yield significant achievements in computer vision fields [31] [32].

Wassel et al. [33] have proposed a vision transformer-based ensemble for fundus image classification. To do this, they used one large merged dataset comprising six available fundus image datasets for Glaucoma detection.
Kamran et al. [34] have introduced VTGAN, a semi-supervised conditional GAN. It can produce the retinal vascular structure from fundus images and distinguish between healthy and abnormal retinas.

A novel lesion-aware transformer (LAT) has been introduced by Sun et al. [35] for joint grading of DR and lesion discovery. It is a deep model using an encoder-decoder structure, where the encoder is based on pixel relation and the decoder on the lesion filter. Gu et al. [36] proposed an intelligent DR classification model for fundus images. It can denote all the five stages of DR, including no DR, mild, moderate, severe, and proliferative. This model consists of two key modules: the feature extraction block, or FEB, mainly used for feature extraction, and the grading prediction block, or GPB, for classifying the five stages of DR. The FEB has a transformer with more fine-grained attention, to concentrate more on retinal hemorrhage and exudate areas.

This paper proposes an approach based on a local and global learning strategy for fundus disease classification. This method, called Global-Local Multi-Label Classification Network, contains two modules, a global module and a local module. A version of the vision transformer has been applied in the global module, and a shallow feature extractor convolutional network has been used as a local branch. Combining global and local learning is one of the innovations in the high-performance approach for classifying eye diseases. The results show that the simultaneous application of global and local modules can improve the learning process and that the classification metrics are higher than other state-of-the-art works in this area.

PROPOSED METHOD

In this study, a framework is proposed to classify eye fundus images. This framework consists of two parts. The first one is the preprocessing of fundus data, considering using transformers. The second part proposes a network based on the vision transformer and CNNs. In this regard, in this section, some explanations have been given about the preprocessing stage first, and next, the Global-Local Multi-Label Classification Network has been introduced.

DOI: 10.46121/pspc.53.4.13

3.1 Preprocessing

The preprocessing stage consists of three steps. Before all, some image processing techniques have been applied to the original dataset. Firstly, for more concentration on the region of interest of the fundus images, the surrounding black region of the fundus has been removed from them and then cropped.

One of the shortcomings of using the transformer approach is that input dimensions are limited. Since in the first stage of feeding the images into the transformer, they are converted to patches with constant same sizes, not every image having arbitrary dimensions can be processed by this approach. This is a major constraint on the dimensions of input images. The preprocessing stage considers this issue. Owing to the different dimensions of images, reducing the computational effort, and considering this issue, cropped images were resized to a 224 × 224 resolution.

Next, resized images are fed into an image enhancement step. Contrast limited histogram equalization (CLAHE) [37] has been used in this step. In this work, only the green channel is considered input among the three channels for RGB colors. This was done because the red channel has low contrast, and the blue channel might be noisier than other channels, so these are undesirable to be considered input [38]. Furthermore, the computational load is reduced due to using one channel instead of three channels for each image.

Transformer networks have numerous parameters and a unique learning process. Therefore, using them needs a large amount of data; otherwise, it provides the "data hungry" problem. Preparing a dataset with hundreds of samples is expensive, and medical image analysis is sometimes even impossible. Furthermore, well-equipped hardware is needed to process such a large amount of data, another problem for researchers in this field. In the medical image analysis field, the preparation of learning data and tagging them (for classification of images) or grand truth (for segmentation of images) needs experienced experts to spend a lot of time and money. In addition, only a few are usable among millions of samples for some cases and specific illnesses. Numerous works that have been done in this field show that learning transformer networks through a limited data set yields much weaker results compared to convolution networks, which use the deep learning approach and, in some cases is, even uncompetitive with them. Therefore, the data hungry problem is a major limitation for transformer networks.

In addition, one of the common problems in multi-label classification tasks is imbalancing the distribution of images with different labels. The data augmentation method cannot only deal with the data-hungry problem but also solve misbalancing issues. Hence, data augmentation has been used as the last preprocessing stage. In this work, the number of training data images has increased based on their distribution of labels in the dataset by applying different classical image transformations, and the distribution of different labels in the training data has become approximately equal. The transformations applied on training images for augmentation are random re-scaling, random rotation, random horizontal and vertical flip, and brightness, contrast, hue, and saturation change. Details of the augmentation step are explained in **Error! Not a valid bookmark self-reference.**. Eventually, the preprocessed data was randomly split into the datasets (75% and 25%) as training and validation data.

Considering this, the number of images in the set has increased from 7000 to 249,264 due to imbalanced data. In other words, by balancing the classes, approximately 33 to 40 images are produced from every image in the set.

Table 1: Augmentation methods that have been applied for training the dataset

Augmentation method	Description	Probability %
Rotation	Rotates the image by an angle in range: -10 to 10	70
Re-scaling	scaling ratio (0.4, 0.6, 0.8)	60
Random Contrast	changes the Contrast of fundus image	40
Random brightness	changes the brightness of fundus image	30
Random vertical flip	Flips the fundus image vertically	70
Random horizontal flip	Flips the fundus image horizontally	70
Random Hue and saturation	Changing the hue, saturation of the fundus image	60

DOI: 10.46121/pspc.53.4.13

Power System Protection and Control
ISSN:1674-3415

In this study, both left and right eye fundus images are used as the input simultaneously, and the label is assigned according to each person beside each image. Hence, the input consists of three layers: the green channel of left and right eye fundus images and the summation of these two as the third channel of the input.

3.2 Global-Local Multi-Label Classification Network

An approach proposed to classify fundus images based on diseases and abnormalities is called the Global-Local Multi-Label Classification Network, which is based on CNNs and transformer models. Figure 2 shows that this approach has global and local modules. The proposed approach can improve classification performance by simultaneously learning both globally and locally. In CNNs, the structure of blocks used for feature extraction concentrates only on a local subdivision of pixels, so it investigates the image locally. On the other hand, transformer models consider images globally and investigate and process the whole set of pixels every time. The simultaneous application of these two makes a model that investigates images both locally and globally, therefore performing a better classification. Since the ViT models need three channels for input and in this work, only two green channels are used, the summation of these two is fed into the network as the third channel.

As shown in Figure 2, the input of the proposed method is the right and left images of one person that are patched for the global module, and a whole image has been fed for the local module. Simultaneous consideration of right and left images for input is one of the contributions of this research. In previous works that use a vision transformer [31] to classify fundus images, only one eye's image is considered input. Considering both left and right images simultaneously, the proposed model can extract more features for every patient, making the learning process better. To describe the architecture of the proposed model, first, the global module is investigated, and then the local module is described.

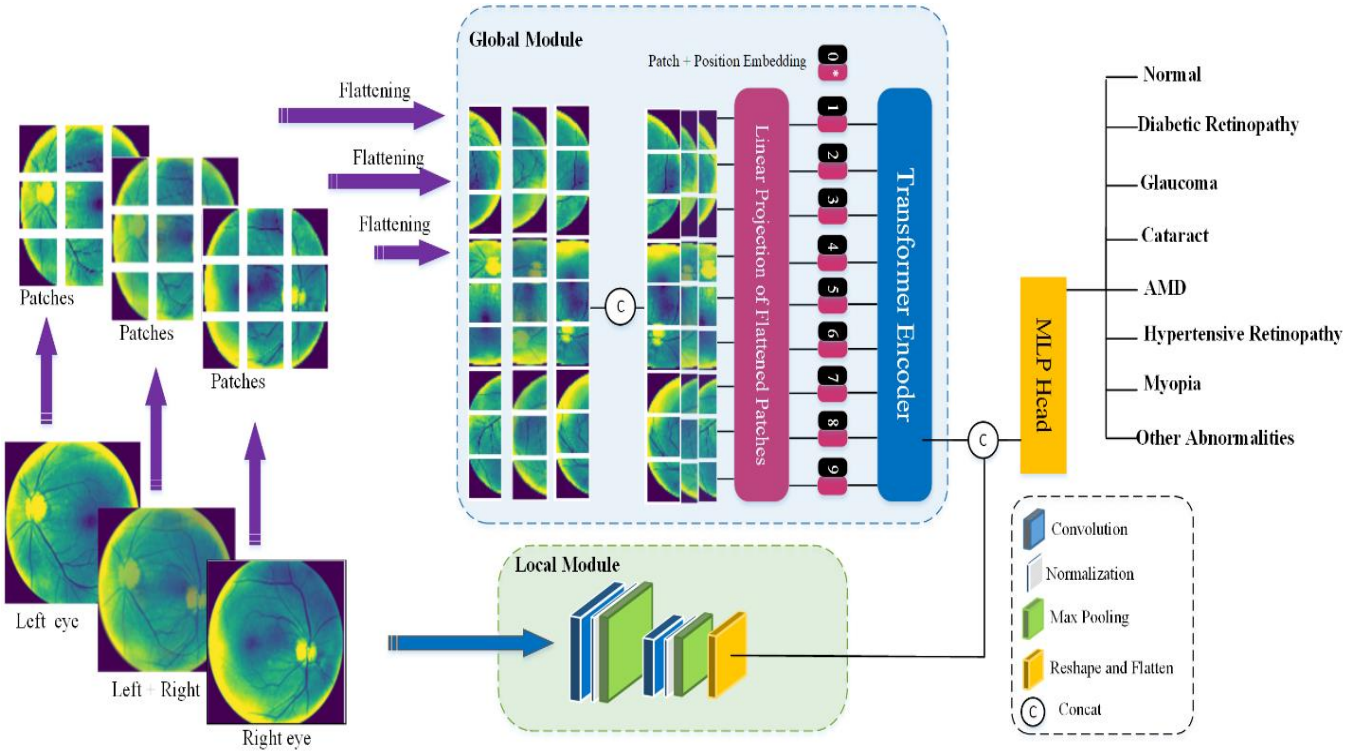


Figure 2: The structure of proposed model that is called Global-Local Multi-Label Classification Network

3.2.1 Global Module:

As shown in Figure 2, the global module, implemented based on the vision transformer (ViT) model, formulates the classification task as a sequence prediction task, capturing dependencies between patches. The first input images of the network are usually divided into same-size patches to use ViT. In this way, input image I with dimensions $R^{H\times W\times C}$ (where H, W, and C represent height, width, and number of channels, respectively) is divided into N patches with Dimensions $\times P \times C$, so $N = \frac{H\times W}{P^2}$. In this work, input image dimensions are

DOI: 10.46121/pspc.53.4.13

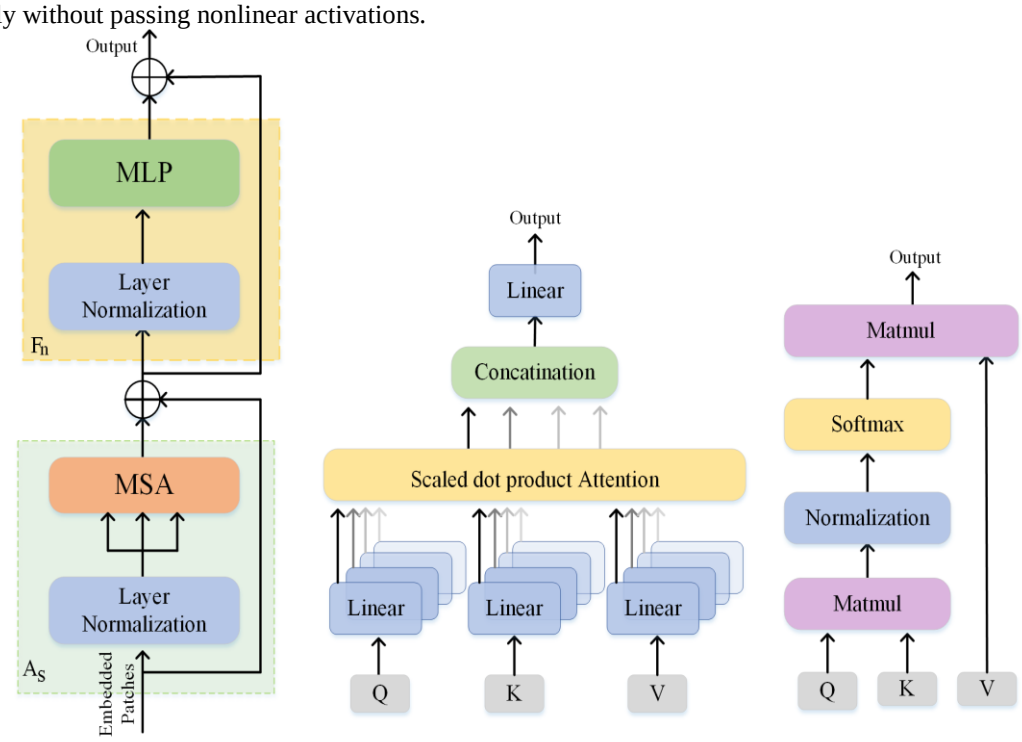
224*224*3, and patch dimensions are empirically taken 72*72 pixels, so the number of the patches obtains 9 (N=$\left\lfloor\frac{224\times224}{72^2}\right\rfloor$=9). In the next step, patches are vectorized using a flattening operation.

Output vectors indicate the amount of data for each part of the image concerning the whole image, which is very similar to transformers network in language processing. The similarity of each patch is compared to each of the input patches, and the most possible information is extracted. The outcome of these comparisons is that the network concentrates more on patches with more information and is more related to the final purpose, paying less attention to other patches. After patching input images, the PE matrix, which contains the similarity of patches to each other and is learnable, is concatenated to input patches. In summary, linear embedding is calculated using a trainable linear layer from vectorized patches. In the next step, positional encoding is added to linear embedding, and this sequence is entered into the ViT encoder using a trainable linear projection:

$$z_0 = \left[x_{class}\,;\, x_p^1 E\,;\, x_p^2 E\,;\,\ldots;\, x_p^{n_p} E\right] + E_{pos} \quad (1)$$

where, $x_{class} = z_0^0$ is a learnable embedding, E is the patch embedding projection, and E_{pos} is the position embedding that keeps the positional information of patches and is added to the patch [39]. Based on [31] mapping patch images converted to the embedding space with positional information, transformer encoders have been used in the next step.

Figure 3 shows the structure of the ViT encoder in more detail. The encoder consists of two blocks: multi-head self-attention (MSA) [40] and MLP. First, embedded patches are entered into the MSA layer, which concatenates multiple attention outputs. MSA helps learn the local and global dependencies of images. Next, a Multi-Layer Perceptron (MLP) is located, which consists of four layers with Gaussian error linear unit (GELU) activation functions and a normalization layer that is imposed before each block. The normalization layer improves the learning time and the generalization procedure. Different parts in the encoder are connected using residual connection links. These links are located after each block, letting gradients flow through the network directly without passing nonlinear activations.



(a) (b) (c)

Figure 3: (a) Structure of the transformer encoder with multi-head self-attention (b) An illustration of multi-head self-attention structure. (c) Overview of self-attention, Matmul represents matrix product of two arrays

This process is shown in equation 2, where L represents the number of transformer blocks. The output of the final transformer encoder layer is z_L^0, which is normalized by LN as described in Equation (2):

$$z_l' = \text{MSA}\,(\text{LN}\,(z_{l-1}) + z_{l-1}\,,\ l = 1, 2, 3, \ldots, L \quad (2)$$

DOI: 10.46121/pspc.53.4.13

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l \quad , l = 1, 2, 3, \dots, L$$
$$y = \text{LN}(z_L^0)$$

The main part of the transformer encoder is MSA and its heads. In this architecture, each head contains Scaled dot product Attention (SA) [40]. Each head of MSA is *applied* for calculating three different vectors, which are query matrix Q, key matrix K, and value matrix V:

$$Q^i = XW_Q^i, \quad K^i = XW_K^i, \quad V^i = XW_V^i \qquad (3)$$
$$i \in \{1,2,3, \dots, h\}$$

where, X is the input and W_Q W_K, and $W_V$represent the weight matrixes that are used in the linear transformers. The tuple of (Q, K, V) is entered into SA, and the attention of input image patches is calculated as below:

$$\text{SA}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{D_h}}\right).V \qquad (4)$$
$$D_h = \frac{D}{h}$$

where $D_h$is the dimension of key vector k, and applying square root provides a suitable normalization to make the gradient more stable. The consequences of SAs from heads are merged in MSA based on the formula as follows:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^k, VW_i^V)$$
$$\text{MSA}(Q, K, V) = \text{concatination}[\text{head}_1; \text{head}_2; \dots; \text{head}_i].W_L \qquad (5)$$

where, W_i^Q,W_i^k,W_i^V and W_L are weight trainable matrices. The final output of MSA is obtained by passing the concatenation of all self-attention heads through a linear layer.

Since transformer models have many parameters, usually for the learning process of this module, the transfer learning approach is used to improve the classification accuracy. One of the main reasons that has made the transformers models successful is using the transfer learning approach. In this approach, the main idea is to train the model on a large-scale dataset and save the resulting weights. Then, to learn the model on the intended dataset, instead of randomly assigning the initial model weight values, pre-trained weights can be used, and learning of the new task can be completed by using the information obtained from the learning stage of the pre-trained model. The more the new task is similar to the previous task that the model is trained by, the more benefit is gained from the application of transfer learning [41].

3.2.2 Local Module

Although learning transformer models can be faster on patches, using them alone is insufficient because of their global view. Being patch-wise limits the learning of these models to inter-patch pixels and makes it dependent only on them. Therefore, in addition to patch-wise learning, global learning is considered, and classification and analysis of medical images can be performed more effectively. As mentioned before, considering this issue, the proposed approach consists of two modules: the global module that performs patch-wise learning based on transformer models and the local module that extracts the effective features from the whole image, which helps to learn and makes the model view more locally. To do this, as shown in Figure 2, the local module gets the input image that includes three channels (extracted green layer of the right and left eye images and the sum of them) and passes it through several convolution and max pooling layers. After extracting local features from the input image, the flattening layer makes output feature maps flatten. The next stage is concatenated with the global module's output, and a set of locally and globally extracted features are entered into the MLP. In the local module, only local feature maps are extracted, and classification is performed after passing concatenated feature maps through MLP that consists of four layers. In addition, the multi-head number is taken as eight, mentions the number of classes.

3.2.3 Training procedure

Binary cross-entropy (BCE) loss has been used to train the proposed model. The equation of this loss function, which is minimized through the learning process, is as follows, which is minimized through the learning process is as follows:

DOI: 10.46121/pspc.53.4.13

$$\text{Loss}_{\text{CE}(p,p')} = -\left(\frac{1}{wh}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(p(x,y)\log(p'(x,y))) + \big(1-p(x,y)\big)\log(1-p'(x,y))\right) \quad (6)$$

Here, w and h are image dimensions, $p(x,y)$ is related to the pixel in the image and $p'(x,y)$ shows the anticipation of output in pixel (x, y).

In this study, the ViT-B/16 model is applied in the global module. As mentioned before, a large number of parameters and the problem of being data hungry cause that for application of transformers usually transfer learning be used. To improve the learning process of this module and make it faster, model weights pre-trained on the ImageNet [42] dataset are used, and after that learning of the model is done with fundus images to classify them based on the disease type.

This model used an Adam optimizer with a learning rate of 10^{-4} and processed input images with a batch size 32. In addition, the dropout approach is considered to learn better, with different rates from 0.1 to 0.3 to decrease the probability of overfitting. The proposed model has been implemented in TensorFlow 2.8.0. A Tesla V100-SXM2-32GB graphic card is used for the experiments and simulations in this research.

EXPERIMENTS

In this study, the proposed model is evaluated by a dataset of fundus images. This section describes the experiments performed to obtain the proposed model and related results. A comparison is made with state-of-the-art models evaluated on the ODIR 2019 dataset to show the superiority of the proposed model.

4.1 Dataset

The dataset used in this research is the Ocular Disease Intelligent Recognition (ODIR-5K) database [42]. It corresponds to the international competition on the ocular disease intelligence recognition challenge held by Pekin University and is accessible on the grand challenge website. This ophthalmic dataset consists of 7000 organized color fundus images of the left and right eyes of 1000 patients with single or multiple abnormalities in each image. 1000 color fundus images were considered as test data. Skilled human readers have labeled the images of the ODIR dataset based on physicians' diagnostic keywords, and each pair included the patient's age and gender information. Then, they have been checked by quality control. Shanggong Medical Technology Co., Ltd. collects the dataset from different hospitals and medical centers in China. Based on the labels, each pair of images is classified into one of these eight groups: normal (N), diabetes (D), glaucoma (G), cataract (C), AMD (A), hypertension (H), myopia (M) and other diseases/abnormalities (O). The dataset is highly imbalanced, considering the number of images in the eight mentioned groups. Images are taken by Canon, Zeiss, and Kowa cameras in different illumination conditions, resulting in different sizes and resolutions. In most cases, there is also a black area surrounding the central part of each image. This dataset uses patient-level diagnostics, which means that in the image analysis, images related to one patient are analyzed at the same time. In Figure 4 and Figure 5, some samples of dataset images are shown. In Figure 4, the variety of fundus images and their labels are seen, and in Figure 5, the fundus images are given that each contains more than one disease.

DOI: 10.46121/pspc.53.4.13

Power System Protection and Control
ISSN:1674-3415

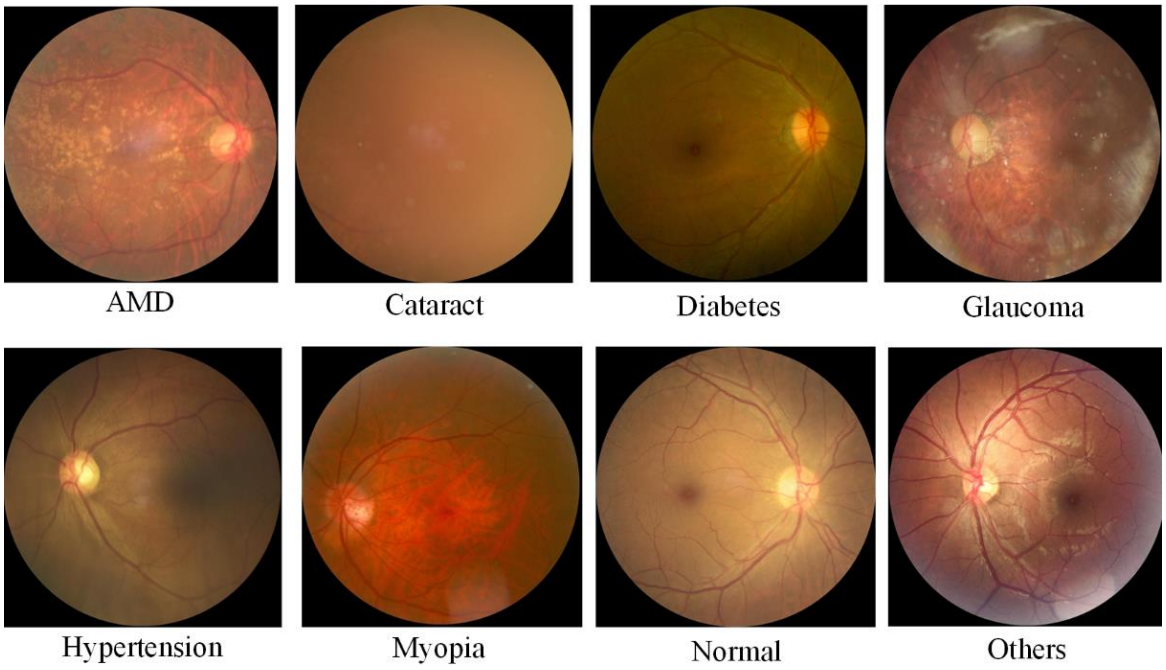


Figure 4: Samples of ODIR-5K dataset and their lables *[42].*

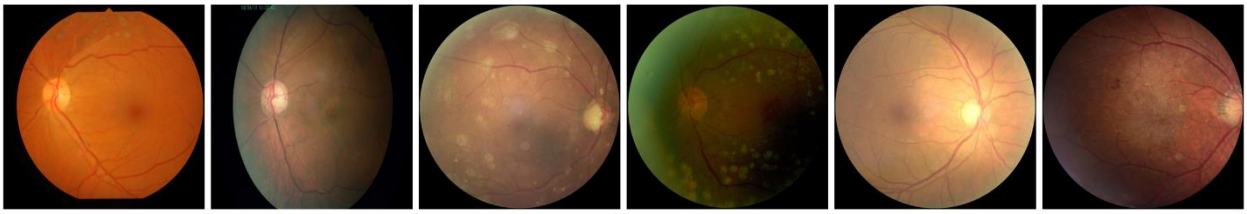


Figure 5: samples of fundus images that contains more than one disease *[42].*

4.2 Metrics

As described in section 4.1, the proposed approach is learned from ODIR 2019 data and evaluated by the test set. The test set of this data includes 1000 images, which are 500 pairs of fundus images of patients' left and right eyes. The learned model investigates the test images, and desired labels are gained to evaluate the proposed model. Since the labels of the test dataset are not provided for the researchers, the proposed model evaluation results are gained online at the Grand Challenge. To do this, outputs of the proposed model are sent to the ODIR challenge site and are investigated there. Criteria used here are area under the curve (AUC), F1, and Kappa score. In addition, the final score is also calculated by the challenge site to evaluate the proposed models.

Generally, one common and primitive assessment criterion for classification is accuracy, which reports data that are classified correctly, as in eq. 7. The recall metric refers to the number of true positive predictions performed, divided by all positive predictions that could made, while the precision criterion gives only positive true predictions among all positive predictions. The missed positive predictions of these two criteria are also represented in equations 8 and 9. The Kappa coefficient is a criterion in statistics for the evaluation of consistency. In most cases, a Kappa coefficient over 0.7 is considered very good. The AUC metric is the area under the ROC curve. The more the AUC metric is close to one, the better the model classification is performed. The final score refers to the average F1 score, Kappa, and AUC value..

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

DOI: 10.46121/pspc.53.4.13

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}} \tag{9}$$

$$\text{F1score} = \left[\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right] \tag{10}$$

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{11}$$

$$\text{Final Score} = \frac{\text{F1}_{\text{Score}} + \text{Kappa} + \text{AUC}}{3} \tag{12}$$

Accuracy (Eq. (7)), Recall (Eq. (8)), and Precision (Eq. (9)) criteria are defined based on true positive (TP), true negative (FP), false positive (FP), and false negative (FN); while F1 score (Eq. (5)) is defined as the harmonic mean of the precision and recall parameters.

4.3 Results and discussion

This section describes the implementation procedure of the global-local multi-label classification method, and the effect of the application of different innovations in it is investigated. As seen in Table 2, the ViT model is used as the baseline model in this study. In this way, the input image is the image of one eye, independently, and includes all three red, green, and blue channels. In this part, all images are modified by flip transformation similar to the image of the left eye to make the model act on them similarly. Next, the model's input is modified so that each image enters the model with its identity. This means that left and right images are combined as input. In addition, for better extraction of features and reduction of the computational load of the model, only the green channel of images is input to the network. Because of transfer learning and the ImageNet dataset, input images must include three channels. For the third channel, one of the green channels of the left or right eye is repeated randomly. Applying this approach to the model's input improved the precision of all criteria (third row in Table 2). Next, to simultaneously consider the information from the left and right eyes, the summation of two green channels, left and right eyes is considered the third input channel and is fed to the model. As shown in Table 2, the change of inputs made a great change in the performance of the model, such that in the final score criterion, the amount of 78.80% that was the result of classification of the model with color fundus images without considering left and right eyes, was increased to 83.28%, that shows the effect of paying attention to the input and using the green channel of left and right eye images separately. In addition, considering the sum of two green channels affected the model's performance, too.

In the next stage, local learning is applied to propose an effective approach for multi-labeled classification, in addition to the global learning that transformer models do, and a global-local multi-label classification approach is developed. In this regard, the ViT approach, described as the global and local views, is added to the model by applying convolutional layers. As seen in the last row of Table 2, using local and global approaches simultaneously, the performance is improved in all criteria dramatically. In all parts, the image dimensions were 224×224, the number of patches was 9, and their dimensions were 72×72.

Table 2: The proposed approach results in different stages

Used method	Input	Final Score%	Kappa score%	AUC value %	F1 score %
Vision Transformer model	Enter all images separately and as left eye	78.80	55.76	91.94	88.72
Vision Transformer model	Enter concatenation of left and right green channel	81.51	61.76	92.74	90.0025
Vision Transformer model	Enter concatenation of left and right green channel and their sum	83.28	64.60	93.78	91.45

DOI: 10.46121/pspc.53.4.13

global-local multi label classification	Enter concatenation of left and right green channel and their sum	**88.79**	**77.17**	**97.12**	**92.26**

For comparison of the proposed approach with other models evaluated on the ODIR 2019 dataset, results are given in Table 3 based on the mentioned criterion. The proposed global-local multi-label classification performed better in most cases than state-of-the-art works on classifying ODIR 2019 eye fundus images. For example, Islam et al. [19] used the shallow CNN approach for this challenge to classify with fewer parameters but did not get a high precision. Jordi et al. [44] solved the multi-label classification by changing it to a multiclass classification. Wang et al. [20] proposed the EfficientB3 model, that is composed of two models. The first model is EfficientNet, which analyses images with gray histogram equalization; the second is EfficientNet, which analyses images with color histogram equalization. The voting technique gains results. In addition, left and right images are concatenated and investigated in this work. Li et al. [45], using the ResNet101 model, performed the classification process, and in [9], four CNN pre-learned models, including ResNet, InceptionV3, MobileNet, and VGG16 were used, and their best precision result is reported. In another study [10] different versions of pre-trained Resnet (ResNet18 ،ResNet34 ،ResNet50، ResNet101) were used. Proposed models in [44] - [46] have used pre-trained CNNs. These proposed models are convolutional models that, despite having many parameters and using transfer learning, cannot perform classification with high accuracy because of their local view. Smitha et al. [47] proposed an approach based on a semi-supervised GAN model. As shown, this approach has improved the results compared to CNNs. However, in the proposed method, by considering local and global learning, the proposed method has achieved better results than previous works.

Table 3: comparison of the proposed approach results with other models evaluated on the ODIR 2019 dataset

State of the art works	Used method	Final Score%	Kappa score%	AUC value %	F1 score %
Islam et al. **[19]**	Shallow CNN	-	31	80.05	0.85
Jordi et al **[43]**	VGG16	-	-	88.71	81.76
Bali et al. **[45]**	Transfer learning VGG16	-	-	0.87	0.91
Wang et al. **[20]**	EffifinetB3	70	49	73.00	89.00
Li et al. **[44]**	ResNet101	-	-	93.00	91.30
Gour and Khanna **[47]**	Two I/P VGG16	-	-	84.93	85.57
He et al. **[48]**	ResNet models	-	-	92.70	90.70
Li et al. **[49]**	different popular deep neural network models	82.67	76	78	94
Smitha et al. **[46]**	Semi-supervised GAN model	0.8333	81	84	85
Demir and Burak **[50]**	(RCNN+LSTM) +NCAR+SVM	-	-	97.00	89.97
Our proposed method	**global-local multi label classification**	**88.79**	77.17	**97.12**	**92.26**

CONCLUSION

In this paper, to investigate eye diseases on fundus images, the global-local multi-label classification approach consists of two modules: local and global. The investigated dataset is ODIR 2019, which includes eight classes of different diseases and is patient-based. Being patient-based means that images of two eyes of a patient are investigated simultaneously. The number of diseases related to each image is not the same; in other words, a patient may have several diseases. Considering these, an approach needed to be proposed capable of properly analyzing images. The proposed approach not only has a local view of images like CNNs but also analyzes images globally by using transformer models. Transfer learning is applied to reduce the effect of model parameters and being data hungry. Images of left and right eyes are input to the model simultaneously; green channels of these two images and some of them as the third channel are fed to the model to be analyzed. In addition to fundus

DOI: 10.46121/pspc.53.4.13

images, other medical images can be analyzed by this approach. In future works, besides improving this model for better performance on fundus images, research will be done on applying global-local multi-label classification in the classification and segmentation of other medical images such as ultrasound [52], MRI [53], CT, and other medical imaging.

Attention:
The data used in this research is public.
The dataset used in this research is the Ocular Disease Intelligent Recognition (ODIR-5K) database. It corresponds to the international competition on the ocular disease intelligence recognition challenge held by Pekin University and is accessible on the grand challenge website(https://grand-challenge.org).

Data Availability: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Funding Information: This research was not funded by any organization or university.

Authors' Contributions
M. Jafarishahbazzadeh: Conceptualization, Methodology, Software, Writing–original draft, validation, Project Administration; M. Eslami: Methodology, Software, Supervision, Writing– review & editing, data curation, funding Acqisition; **H. Tayebi**: Software, Investigation, Validation, validation, Resources, Simulations, Formal Analysis.

REFERENCES

1. Md. Khaled Hassan, et al., "A survey on an intelligent system for persons with visual disabilities," Aust. J. Eng. Innov. Technol , vol. 3, no. 6, pp. 97-118, 2021.
2. Foreman, J., Dirani, M., Taylor, H, "Refractive error, through the lens of the patient.," Clinical & Experimental Ophthalmology, vol. 45, no. 7, p. 673–674, 2017.
3. Jonas, Jost B., et al., "Visual impairment and blindness due to macular diseases globally: A systematic review and meta-analysis," American journal of ophthalmology, vol. 158, no. 4, pp. 808-815, 2014.
4. Walton OB, Garoon RB, et al., "Evaluation of automated teleretinal screening program for diabetic retinopathy," JAMA ophthalmology , vol. 134, no. 2, pp. 204-209, 2016.
5. Antonio PR, Marta PS, et al., "Factors associated with changes in retinal microcirculation after antihypertensive treatment," Journal of human hypertension , vol. 28, no. 5, pp. 310-315, 2014.
6. Luca Giancardo, et al., "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," Medical image analysis, vol. 16, no. 1, p. 216–226, 2012.
7. Tham YC, Li X, Wong, et al., "Global prevalence of glaucoma and projections of glaucoma burden through 2040 a systematic review and meta-analysis," Ophthalmology, vol. 121, no. 11, pp. 2081-2090., 2014.
8. P-R Antonio, et al., "Factors associated with changes in retinal microcirculation after antihypertensive treatment," Journal of human hypertension, vol. 28, no. 5, 2014.
9. Antonio PR, Marta PS, et al., "Computerized retinal image analysis - a survey.," Multimedia Tools and Applications, vol. 79, no. 31–32, p. 22389–22421, Aug 2020.
10. Badar M, Haris M, et al., "Application of deep learning for retinal image analysis: A review.," Computer Science Review, vol. 1, no. 35, 2020.

DOI: 10.46121/pspc.53.4.13

11. Gondal WM, et al., "Weakly-supervised localization of diabetic retinopathy lesions in retinalfundus images," in IEEE international conference on image processing (ICIP)., Beijing, China, 2017.
12. Yin P, Wu Q, Xu Y,, "PM-Net: Pyramid multi-label network for joint optic disc and cup segmentation," Proc. Med. Image Comput. Comput.-Assisted Intervention, p. 129–137, 2019.
13. "National Eye Institute, NIH: Eye disease simulations (2020)," [Online]. Available: https://medialibrary.nei.nih.gov/search?keywords=&category=Eye%20Disease%20Simulation.
14. Li N, Li T, Hu C, Wang K, Kang H, "A benchmark of ocular disease intelligent recognition: one shot for multi-disease detection," Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, p. 177–193, 2021.
15. Deepak RanjanNayak, Dibyasundar Das, et al., "ECNet: An evolutionary convolutional network for automated glaucoma detection using fundus images," Biomedical Signal Processing and Control , vol. 67, p. 102559, 2021.
16. Arvind Sai Krishnan, Derik Clive R, et al., "A transfer learning approach for diabetic retinopathy classification using deep convolutional neural networks," in 15th IEEE India Council International Conference (INDICON), 2018.
17. FelixGrassmann, Judith Mengel kamp, et al, "A Deep Learning Algorithm for Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular," Ophthalmology, vol. 125, no. 9, pp. 1410-1420, 2018.
18. Zahra Sobhaninia, Hajar Danesh, Rahele Kafieh, J Jothi Balaji, Vasudevan Lakshminarayanan, "Determination of foveal avascular zone parameters using a new location-aware deep-learning method," in Applications of Machine Learning, 2021.
19. Md. Tariqul Islam; Sheikh Asif Imran, et al., "Source and Camera Independent Ophthalmic Disease Recognition from Fundus Image Using Neural Network," in IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), 2019.
20. Jing Wang; Liu Yang; et al., "Multi-label classification of fundus images with efficientnet," IEEE Access,, vol. 8, pp. 212499-212508, 2020.
21. Gao Huang, Zhuang Liu, et al., "Densely connected convolutional networks," Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700-4708, 2017.
22. Md Kamrul Hasan, et al., "Cataract Disease Detection by Using Transfer Learning-Based Intelligent Methods," Computational and Mathematical Methods in Medicine , 2021.
23. E. Sudheer Kumar and C. Shoba Bindu, "MDCF: Multi-Disease Classification Framework On Fundus Image Using Ensemble Cnn Models," Journal of Jilin University, vol. 40, no. 09, pp. 35-45, 2021.
24. S. P. K. Karri, Debjani Chakraborty, and Jyotirmoy Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," Biomedical optics express,, vol. 8, no. 2, pp. 579-592, 2017.
25. Daniel S.Kermany, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell , vol. 172, no. 5, pp. 1122-1131, 2018.

DOI: 10.46121/pspc.53.4.13

26. Seyed Vahid Moravvej, Roohallah Alizadehsani, Sadia Khanam, et al., "RLMD-PA: a reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights," Contrast Media & Molecular Imaging, 2022.
27. Ji-Jiang Yanga, Jianqiang Lib, Ruifang Shena, et al., "Exploiting ensemble learning for automatic cataract detection and grading," Computer methods and programs in biomedicine, vol. 124, pp. 45-57, 2016.
28. Mulagala Sandhya, et al., "Detection of Diabetic Retinopathy (DR) Severity from Fundus Photographs: An Ensemble Approach Using Weighted Average," Arabian Journal for Science and Engineering, pp. 1-8, 2021.
29. Ashish Vaswani, Noam Shazeer, et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
30. Devlin J, Chang MW, Lee K, Toutanova K. , "Bert: Pre-training of deep bidirectional transformers for language understanding.," [Online]. Available: arXiv preprint arXiv:1810.04805. 2018 Oct 11..
31. Dosovitskiy A, et al., "An image is worth 16x16 words: Transformers for image recognition at scale.," 2020.
32. Zahra Sobhaninia, Nasrin Abharian, Nader Karimi, Shahram Shirani, Shadrokh Samavi, "Endoscopy Classification Model Using Swin Transformer and Saliency Map," 2023.
33. Moustafa Wassel, et al., "Vision Transformers Based Classification for Glaucomatous Eye Condition".
34. Sharif Amit Kamran, et al., "VTGAN: Semi-supervised Retinal Image Synthesis and Disease Prediction using Vision Transformers," Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
35. Rui Sun, Yihao Li, et al., "Lesion-Aware Transformers for Diabetic Retinopathy Grading," in In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
36. Zongyun Gu, et al., "Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention," Computational Intelligence and Neuroscience , 2023.
37. Zahra Sobhaninia, Nader Karimi, Pejman Khadivi, Shadrokh Samavi, "Brain tumor segmentation by cascaded multiscale multitask learning framework based on feature aggregation," Biomedical Signal Processing and Control, vol. 85, p. 104834, 2023.
38. Thomas Walter, Pascale Massin, et al., "Automatic detection of microaneurysms in color fundus images," Medical Image Analysis, vol. 11, no. 6, pp. 555-566, 2007.
39. Devlin, Jacob, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in arXiv preprint arXiv:1810.04805 , 2018.
40. Dosovitskiy, Alexey, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in arXiv preprint arXiv:2010.11929, 2020.
41. Seyed Vahid Moravvej, Seyed Jalaleddin Mousavirad, et al., "An Improved DE Algorithm to Optimise the Learning Process of a BERT-based Plagiarism Detection Model," in IEEE Congress on Evolutionary Computation (CEC), 2022.
42. J Deng, et al, "Imagenet: A large-scale hierarchical image database.," in IEEE conference on computer vision and pattern recognition, 2009.

DOI: 10.46121/pspc.53.4.13

43. "Ocular Disease Recognition (2021) Retrieved from," https://www.kaggle.com/andrewmvd/ocular-disease-recognition-odir5k. [Online].
44. Jordi et al., "Ocular disease intelligent recognition through deep learning architectures," in Universitat Oberta de Catalunya: Barcelona, Spain , 2019.
45. Cheng Li; et al., "Dense Correlation Network for Automated Multi-Label Ocular Disease Detection with Paired Color Fundus Photographs," in International Symposium on Biomedical Imaging (ISBI), 2020.
46. Akanksha Bali, Vibhakar Mansotra, "Transfer Learning-based One Versus Rest Classifier for Multiclass Multi-Label Ophthalmological Disease Prediction," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 12, no. 12, 2021.
47. A. Smitha, P. Jidesh, "Classifcation of Multiple Retinal Disorders from Enhanced Fundus Images Using Semi supervised GAN," SN Computer Science, vol. 3, no. 1, pp. 1-11, 2021.
48. Gour, Neha, and Pritee Khanna, "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network," Biomedical Signal Processing and Control , vol. 66, p. 102329, 2021.
49. Junjun He, et al., "Self-speculation of clinical features based on knowledge distillation for accurate ocular disease classification," Biomedical Signal Processing and Control, vol. 67, p. 102491, 2021.
50. Li et al., "A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection," Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Springer International Publishing, pp. 177-193, 2021.
51. Fatih Demir and Burak Ta¸scı, "An Effective and Robust Approach Based on R-CNN+LSTM Model and NCAR Feature Selection for Ophthalmological Disease Detection from Fundus Images," Journal of Personalized Medicine , vol. 11, no. 12, 2021.
52. Zahra Sobhaninia, Ali Emami, Nader Karimi, Shadrokh Samavi, "Localization of fetal head in ultrasound images by multiscale view and deep neural networks," in 2020 25th International Computer Conference, Computer Society of Iran (CSICC), 2020.
53. Zahra Sobhaninia, Nader Karimi, et al., "Medial Residual Encoder Layers for Classification of Brain Tumors in Magnetic Resonance Images," in 30th International Conference on Electrical Engineering (ICEE), 2022.

DOI: 10.46121/pspc.53.4.13