IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 October 2023, accepted 15 October 2023, date of publication 23 October 2023, date of current version 27 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326528

RESEARCH ARTICLE

Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images

WALEED NAZIH[1,2], AHMAD O. ASEERI[1], OSAMA YOUSSEF ATALLAH[3], AND SHAKER EL-SAPPAGH[4,5]

[1]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[2]Department of Computer Science, Cairo Higher Institute for Engineering, Computer Science and Management, Cairo 11865, Egypt
[3]Department of Biomedical Engineering, Medical Research Institute, Alexandria University, El-Hadra Bahry, Alexandria 21561, Egypt
[4]Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt
[5]Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha 13518, Egypt

Corresponding author: Ahmad O. Aseeri (a.aseeri@psau.edu.sa)

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number (IF-PSAU-2022/01/19574).

ABSTRACT Diabetic Retinopathy (DR) is a result of prolonged diabetes with poor blood sugar management. It causes vision problems and blindness due to the deformation of the human retina. Recently, DR has become a crucial medical problem that affects the health and life of people. Diagnosis of DR can be done manually by ophthalmologists, but this is cumbersome and time consuming especially in the current overloaded physician's environment. The early detection and prevention of DR, a severe complication of diabetes that can lead to blindness, require an automatic, accurate, and personalized machine learning-based method. Various deep learning algorithms, particularly convolutional neural networks (CNNs), have been investigated for detecting different stages of DR. Recently, transformers have proved their capabilities in natural language processing. Vision transformers (ViTs) are extensions of these models to capture long-range dependencies in images, which achieved better results than CNN models. However, ViT always needs huge datasets to learn properly, and this condition reduced its applicability in DR domain. Recently, a new real-world and large fundus image dataset called fine-grained annotated diabetic retinopathy (FGADR) has been released which supported the application of ViT in DR diagnosis domain. The literature has not explored FGADR to optimize ViT models. In this paper, we propose a novel ViT based deep learning pipeline for detecting the severity stages of DR based on fundus photography-based retina images. The model has been built using FGADR dataset. The model has been optimized using a new optimizer called AdamW to detect the global context of images. Because FGADR is an imbalanced dataset, we combine several techniques for handling this issue including the usage of F1-score as the optimization metric, data augmentation, class weights, label smoothing, and focal loss. Extensive experiments have been conducted to explore the role of ViT with different data balancing techniques to detect DR. In addition, the proposed model has been compared with the state-of-the-art CNN algorithms such as ResNet50, Incep-tionV3, and VGG19. The adopted model was able to capture the crucial features of retinal images to understand DR severity better. It achieved superior results compared to other CNN and baseline ViT models (i.e., 0.825, 0.825, 0.826, 0.964, 0.825, 0.825, and 0.956 for F1-score, accuracy, balanced accuracy, AUC, precision, recall, specificity, respectively). The results of the proposed ViT model were quite encouraging to be applied in real medical environment for assisting physicians to make accurate, personalized, and timely decisions.

INDEX TERMS Deep learning, vision transformer, diabetic retinopathy, machine learning, disease diagnosis.

The associate editor coordinating the review of this manuscript and approving it for publication was Berdakh Abibullaev.

I. INTRODUCTION

The global prevalence of diabetes has seen a significant increase in recent years, with estimates suggesting a rise

from 9.3% (463 million) in 2019 to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045 [1]. The associated healthcare expenditure has also risen by 316% over the last 15 years, reaching a staggering 966 billion dollars [2]. Diabetes is a major cause of several debilitating conditions, including blindness, heart attacks, and kidney failure [3]. One of the complications that arise from prolonged diabetes is DR, which involves damage to the blood vessels behind the retina and may cause vision loss if not detected early [4], [5]. It affects approximately 30 to 40% of individuals with diabetes worldwide, with over 100 million people living with DR [6]. The condition becomes more prevalent in patients who have had diabetes for over 20 years, accounting for 80% of such individuals and contributing to 12% of new blindness cases, and a major cause of vision loss and blindness for those aged between 20 and 74 years [7].

By 2040, IDF Diabetes Atlas is expecting that one in three diabetes patients will develop DR [8]. Worldwide, patients with DR are expected to grow from 126.6 million in 2010 to 191.0 million by 2030, and the number of patients with vision-threatening DR is expected to increase if no suitable action is taken [9]. According to WHO, DR is estimated to account for 4.8% of the number of cases of blindness worldwide [10]. To avoid the complications of DR, the early detection of the disease is crucial to sustain the patient's vision effectively [11]. Currently, physicians manually investigate the fundus images of eyes to measure the severity of DR. However, these diagnostic procedures are difficult, error-prone, expensive, and time consuming [5], [11], [12]. In addition, because it needs much time of already overloaded physicians, many patients do not receive medical care in a timely manner which results in a severe DR state for many cases. Fortunately, much of the visual loss from DR is preventable if it is coupled with an early assessment and screening [6]. Hence, it is crucial to address this issue by providing an automatic, accurate, accessible, and cheap methodology for the early detection and grading of the disease [13].

Machine learning (ML) and deep learning (DL) techniques have been used to handle this medical problem based mainly on the fundus images which visually record the present ophthalmic appearance of a person's retina [4]. Retinal blood vessel segmentation, lesion segmentation, and DR classification are the regular steps for DR detection. This classification problem is mainly formulated as a binary classification task (i.e., DR or normal retina) which is called DR detection. The grading of DR stage consists of annotating the infected parts and determining the types of infection: mild, moderate, or severe. This task is usually formulated as a multiclass classification task [14]. Using ML models, several studies have been conducted for DR classification and grading. In [15], authors proposed the so called ''tetragonal local octa pattern (T-LOP) features'', which is a new method for representing features of the fundus images. Gayathri et al., [16] utilized support vector machine, random forest, and decision tree to classify DR, and Washburn [17] used Gabor wavelet method with AdaBoost classifier to grade DR. Recently, Selvachandran et al., [7] provided a comprehensive survey about DR detection techniques.

DL has more advanced architectures such as convolutional neural networks (CNNs) that are able to automatically extract deep and spatial features from images [18]. No independent features extraction and feature selection steps are needed with DL algorithms because these techniques can learn deep representations from images. Xu et al. [19] employed a CNN model using the Kaggle EyePACS dataset for classifying retinal fundus images. Stochastic gradient descent was used as the optimizer, and data augmentation techniques such as image resizing, rotation, flipping, shearing, and translation were applied to increase the diversity of the images. The dataset was split into training and testing sets, with 800 and 200 images, respectively. Kazakh-British et al., [20] combined a simple CNN model with an anisotropic diffusion filter to automatically classify DR. Following the same method of building simple DL models, in [21] the authors combined the CNN architectures with the Wiener filter and OTSU for the segmentation to perform binary classification for DR detection. On the other way, transfer learning has been used to use pretrained deep model for better feature extraction. In [22], Rego used InceptionV3 to detect DR using RGB and textures features. Pamadi et al., [23] built both binomial and multinomial classification of fundus images by utilizing MobileNetV2, Saranya et al. [24] used DenseNet-121 model to detect DR from fundus images, and different architectures of EfficientNet were investigated in [25]. Some studies proposed hybrid models by using DL model for feature extraction and regular ML model for classification. For example, Boral and Thorat [26] utilized InceptionV3 for representation learning and support vector machine for DR classification. Ensemble models, also called Hybrid CNN architecture, are well-known techniques to build robust and accurate classifiers. Jiang et al. [27] proposed an ensemble model for DR detection. The ensemble utilized three CNN models. In addition, the Adaboost was used for efficient integration of these DL models' outputs using learned weights. The image preprocessing steps were resizing (i.e., 520, 520, 3), rotation, translation, mirroring, contrast, sharpness, and brightness. In [28], Kaushik et al., built a stacking ensemble model consisting of three CNN models for DR classification. Zhang et al., [29] proposed an ensemble model consisting of three CNN architectures (i.e., InceptionV3, Xception, and InceptionResNetV2) to detect DR, and an ensemble of ResNet50 + DenseNet169 + DenseNet201 to grade DR. This study consedered a set of preprocess steps to improve the quality of the images. On the other hand, Bellemo et al., [30] proposed an ensemble with VGGNet + ResNet, and Xie et al., [31] proposed an ensmeble with VGGNet + ResNet + DenseNet for DR grading without doing any preprocessing steps.

DR is graded into five stages based on the International Clinical Diabetic Retinopathy (ICDR) scale [32]: no apparent retinopathy, mild nonproliferative diabetic retinopathy

FIGURE 1. Examples of diabetic retinopathy retina. The left image is a normal retina, and right is a DR-4 retina.

(NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR) [33]. Figure 1 provides visual examples comparing a normal retina with one affected by diabetic retinopathy, highlighting multiple lesions. In [34], a simple CNN model has been explored to grade DR after applying a green channel filter on fundus images. Luo et al., [35] proposed MVDRNet by combining a VGG-16 model with attention mechanisms. The study utilized a dataset containing multi-view fundus images. Araujo et al., [36] proposed a CNN model cased GRADUATE. It was an uncertainty-aware DL model that could give a pathologically explainable description to support its decisions. Gayathri et al., [16] proposed hybrid multipath CNN model for deep representation learning for DR grading and combine the features extractor with three different ML classifiers including random forest, SVM, and J48. Shaik and Cherukuri [37] proposed "Hinge Attention Network (HA-Net)" which combined VGG-16 for spatial representation learning and multiple attention stages for DR severity grading. Li et al., [38] proposed a semi-supervised auto-encoder graph network (SAGN) where the autoencoder was used for feature extraction, radial bases function was used to calculate neighbor correlations, and the graph CNN was used to grade DR. Wang et al. [39] compared InceptionV3, AlexNet, and VGG-16 for DR grading, and found that InceptionV3 achieved the best accuracy. ML and DL models are trained using publicly available datasets [4] such as DRIVE [40], EyePACS [41], APTOS [42], STARE [43], DIARETDB [44], HEIMED [45], ROC [46], Messidor [47], e-ophtha [48], DDR [49], FGADR [33], DeepDRiD [50], IDRiD [51], and RFMiD [52]. The main problems of these data are the severe data imbalance and poor image quality. Han et al. [53] proposed a CNN model called a category weighted network to address data imbalance at the model level. The study employed relation weighted labels instead of one-hot labels to preserve the distance relationship between labels. Using the DDR and APTOS datasets, the model achieved high kappa scores and accuracy for DR grading. Tummala et al., [12] proposed an EfficientNetV2-based deep ensemble model for automated estimation of fundus images quality. The model achieved a test accuracy of 75% for the quality enstimation using the DeepDRiD dataset. For more information about the role of deep learning in detection and staging of DR, readers are guided to these survey studies [5], [7], [11], [13], [54], [55], [56], [57], [58].

Most studies in computer vision rely on CNN architectures due to their ability to encode spatial equivariance through convolutional layers, enabling deep visual representations. However, transformer-based architectures utilizing self-attention mechanisms, such as vision transformers (ViTs), have shown superior performance to standard convolutions in various computer vision applications [55], despite their increased computational requirements [59]. ViT offers several advantages over CNNs including: (1) ViT can capture longer-range dependencies among pixels than CNNs, (2) ViT has a built-in saliency mechanism which supports model to concentrate more on specific focus points during processing, and (3) ViT defends better against adversarial attacks compared to CNNs. It's worth noting that the flexible CNN architectures support the scaling of depth from a few to hundreds of layers, a flexibility that is not present in ViT. The ViT model was first introduced by Dosovitskiy et al. in 2020 [60], adapting the attention mechanism from text-based data to images. ViT converts an input image into a sequence of patches, which are then processed by a vision encoder and fed into a multilayer perceptron (MLP) for classification. Each patch is mapped to a latent vector using transformer layers, incorporating positional embeddings. Transformer encoders, consisting of multi-head self-attention (MSA) and MLP modules, are employed to aggregate global information and propagate features across layers. The resulting image representation is used for classification. ViT has been explored in many domains and it outperformed CNN [61], [62]. In addition, it has been successfully applied to detect and grade DR in various studies [63], [64], [65], [66]. Kumar et al., [67] compared three DL architectures including Transformer-based network (i.e., Swin-Transformer and Vision-Transformer (ViT)), CNN (i.e., EfficientNet and ResNet), and multi-layered perceptron

TABLE 1. A description of the dataset.

DR Scale	Description	Number of Images
0	no retinopathy	101
1	mild non-proliferative DR (NPDR)	212
2	moderate NPDR	595
3	severe NPDR	647
4	proliferative DR	287

(MLPMixer) for DR detection. They discovered that models based on transformer architectures had better accuracy than these. Yu et al., [65] proposed multiple instances learning VIT (MIL-ViT) DL model. The training has been done in two stages: first, pretraining on a large fundus image dataset, and second fine-tuning on the downstream task for DR detection. Performance has been evaluated using APTOS2019 and RFMiD2020 datasets which show that MIL-ViT achieved better results than CNN. Zhang et al., [68] proposed the TC-Net image segmentation framework by combining both CNN and ViT. For locality-aware perspective, an encoder-decoder CNN model is used to dig out local information using the convolution operations. For long-range dependencies, a ViT model is used to focus on the global context. In addition, dynamic cyclical focal loss was used to address the class imbalance. The TC-Net achieved mean pixel accuracy of 0.517 and 0.699 on the DDR and IDRiD, respectively. Adak et al., [69] proposed an ensemble of transformers models where four models have been integrated to determine the degree of DR severity. Gu et al., [70] proposed a DR grading model based on an integrated model of vision transformer and residual attention. The proposed model has two main blocks (1) feature extraction block based on transformer that can pay more attention to retinal hemorrhage and exudate areas, and (2) grading prediction block based on residual attention that can capture different spatial regions for different classes. For more studies about transformers in medical images, readers are guided to these recent studies [71], [72]. Most of the current DR diagnosis systems do not achieve satisfactory performance, and there is room for improvements in the literature. The output of the proposed model can be translated into a set of new knowledge that can be translated into medical practices for physicians in hospitals and medical centers. In [82] and [83], the authors discussed the role of translational medicine concept which is the ''effective translation of the new knowledge, mechanisms, and techniques generated by advances in basic science research into new approaches for prevention, diagnosis, and treatment of disease.'' These studies highlighted the crucial role of AI in applying translational medicine which improves healthcare decisions. The main contributions of this work can be summarized as follows.

- Adaptation of Vision Transformer (ViT) for the automated classification of diabetic retinopathy, tailoring the model specifically for this purpose.
- Utilizes a dataset comprising high-quality resolution images, annotated by three ophthalmologists, and containing a substantial number of images (1842 in total).
- Addresses the challenge of imbalanced datasets by employing image augmentation techniques and assigning appropriate loss weights.
- The proposed model is validated on unseen test data and compared against other state-of-the-art models, establishing its performance and effectiveness in comparison to existing approaches.

The remainder of the paper is organized as follows. Section II presents the methodology of the study and the proposed model. Section III represents the results of the study and discussion about findings. Finally, conclusion and future work are drawn in section IV.

II. MATERIALS AND METHODS

In this section, we describe the dataset and the proposed model architecture. We explain the resulting model's hyperparameters, and we specify the used evaluation metrics.

A. DATA DESCRIPTION

The development of computer-aided diagnosis systems for DR faces significant challenges related to the availability and quality of training data. While there are some publicly accessible DR databases, most of them lack detailed annotations and only provide image-level labels, which are often unreliable. To address these limitations, it is essential to have datasets that include both pixel-level lesion masks and image-level severity grading. Among the available datasets, the fine-grained annotated diabetic retinopathy FGADR and IDRiD datasets fulfill these requirements. The IDRiD dataset, although containing only 81 images with segmented lesions, does not provide a sufficiently large dataset for our purposes. Therefore, the FGADR dataset, consisting of 1,842 images with six segmented lesions, is considered the most suitable choice. The FGADR dataset includes both pixel-level lesion annotations and image-level grading labels. The severity of DR is graded according to the international protocol [32] as shown in Table 1. In addition, the annotations in the FGADR dataset were performed by three ophthalmologists, ensuring reliable and consistent grading labels. Additionally, the images and masks in the FGADR dataset are already cropped, focusing on the retina area, eliminating unnecessary black areas in the image margins. Moreover, all images in the dataset have the same dimensions of 1280×1280, eliminating the need for padding and ensuring consistency in the dataset. Zhou [33] performed a benchmark study based on this dataset, but they provided just a preliminary result which could be improved using other hybrid techniques, new DL architectures like ViT, and extra data preparation steps. During experimentation, we followed the data usage agreement of Zhou et al., [33] and all the experiments were carried out in accordance with relevant guidelines and regulations. FGADR is available from the corresponding author on reasonable request [33]. Note that all experiments

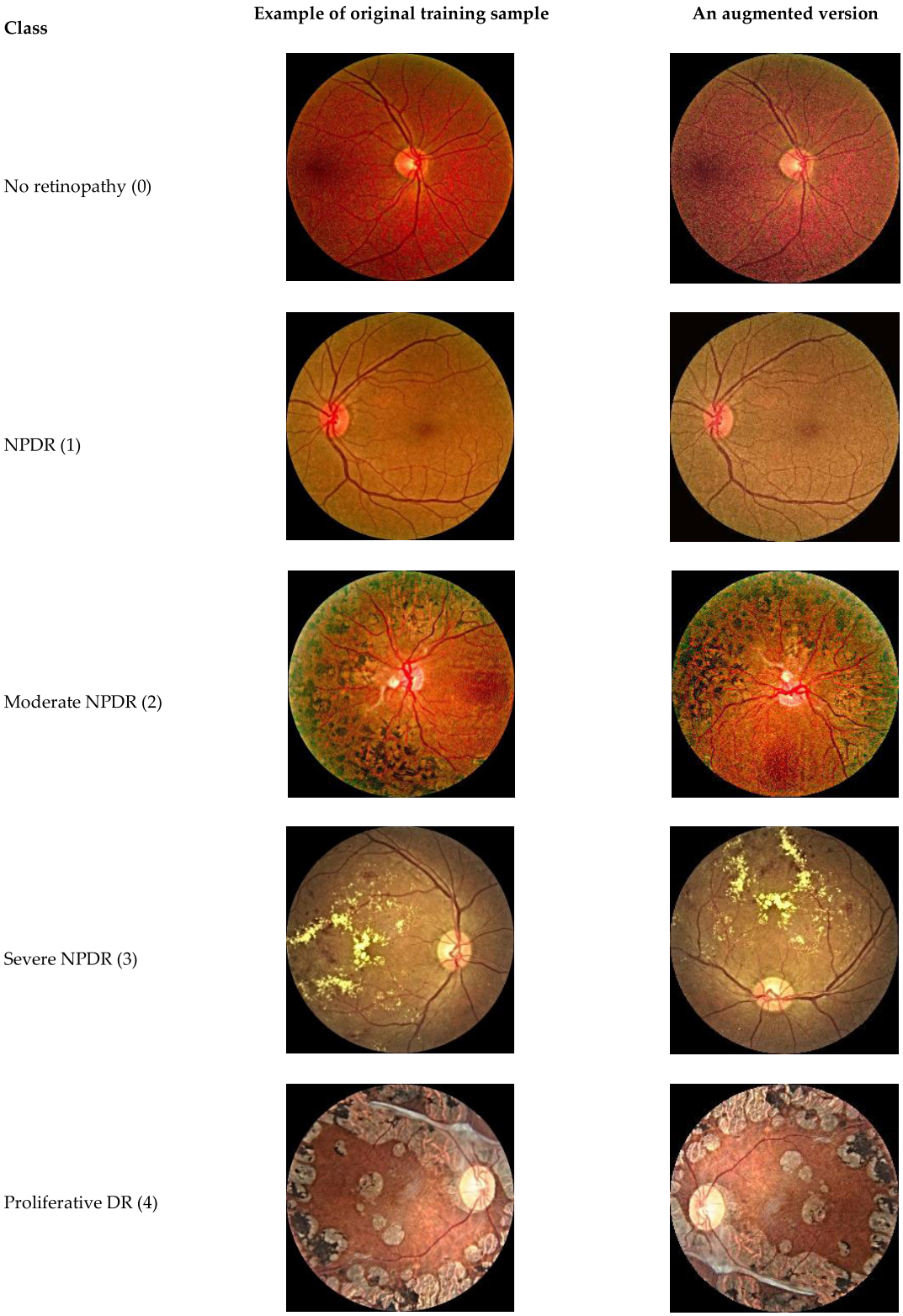


FIGURE 2. Examples from train set for images from every class and example of its augmented version.

were carried out based on the relevant guidelines and regulations, and we followed the protocols mentioned by the data releasing organization in their respective licenses.

B. PROPOSED ViT FRAMEWORK

In this section, we discuss the proposed ViT framework for DR grading. The proposed framework has the four main steps

of (1) image augmentation, (2) image normalization, (3) data splitting, (4) data balancing, and (5) ViT model training and validation.

1) IMAGE AUGMENTATION

Six augmentation steps have been taken to increase the size of the dataset. Figure 2 shows examples of the classes and augmented versions from these images.

Flip vertical and horizontal: With probabilities p_hf and p_vf.

a: TRANSPOSE

By swapping the height and width dimension.

b: RANDOM ROTATION

This technique is consistently applied, where the rotation angle α is randomly determined within the specified range defined by θ, i.e., $\alpha \in [-\theta, \theta]$, with $\alpha = \{90,180,270\}$.

c: RANDOM BRIGHTNESS ADJUSTMENT

This is pixel-level adjustment. We applied this transformation 40% of the time with a random factor $\beta \in [\beta_{\min}, \beta_{\max}]$ for $\beta = 0$ results in a complete black image, $\beta = 1$ results in the original image unchanged, and $\beta > 1$ increases the brightness by the β factor.

d: RANDOM CONTRAST ADJUSTMENT

This is pixel-level adjustment. We applied this transformation with a probability of 40% based on a factor $k \in [k_{\min}, k_{\max}]$ for $k = 0$ results in a solid gray image, $k = 1$ results in the original image unchanged, and $k > 1$ increases the contrast by the k factor.

e: RANDOM SATURATION ADJUSTMENT

This is pixel-level adjustment with probability 40%. It adjusts the saturation of RGB images by a random factor β randomly picked in the interval $\beta \in [\beta_{\min}, \beta_{\max}]$. This process involves converting RGB images into a floating-point representation, followed by a conversion to the HSV color space. An offset is applied to the saturation channel, and then the image is converted back to the RGB color space. Finally, the image is restored to its original data type.

2) IMAGE NORMALIZATION

Data normalization is an important pre-processing step. It ensures that the pixel values come from the Gaussian distribution. Neural networks rely on gradient calculations and try to learn how weighty a feature or a pixel should be in determining the class of an image. Normalized pixel values help the gradient calculations to stay consistent and not get so large which slows down or prevents the network conversion. As a result, normalization of pixel values (intensity) of images is recommended to avoid the influence of high frequency noise and low noise, accelerate the training process, make the training more stable, and to give the same weight to different features (pixels) which results in more accurate model weights. Normalization is performed by subtracting the mean μ from each pixel. The resulting values are then divided by the standard deviation σ, which is computed from all pixel values. More formally, the normalized value $z(x) = ((x\text{-}\mu)) / \sigma$. The resulting distribution resembles the Gaussian function centered at zero. Because pixel numbers must be positive, the scale of the data is [0, 1].

3) DATA SPLITTING

The dataset has been split into 80%/10%/10% for training/validation/testing. Training dataset is used to optimize our proposed ViT model, model validation is based on a separate 10% of the data, and testing or generalization performance is based on 10% of unseen dataset.

4) DATA BALANCING AND BALANCED TRAINING

As shown in Table 1, FGADR is an imbalance dataset. Training an DL model based on imbalanced dataset results in a biased model towards the majority class. The model could consider the minority class examples as outliers and completely concentrate on the majority class. However, the minority class is always the positive class that needs full attention in the model's learning process. Handling this problem involves many techniques including dataset or model approaches. The most common dataset-based techniques are sampling-based techniques and data augmentation. In our study, we apply different data augmentation techniques, see Section II-B1. Model based approaches include several techniques [53]. In our paper, we combine several techniques including the usage of F1-score as the optimization metric, class weights, label smoothing, and focal loss. The F1-score was chosen over accuracy for two main reasons [73]. The F1-score is particularly suitable for datasets with imbalanced distributions of classes. Imbalanced datasets often have minority classes, such as the one at hand. Accuracy may be biased towards the more frequent classes, resulting in lower classification accuracy for the less frequent classes. This score mitigates this bias and provides a more reliable evaluation. By considering the F1-score, the evaluation of the ViT model's detection performance considers the model's ability to balance precision and recall, as well as its overall discriminative power in differentiating between different classes in the dataset. The default weighting for classes in a classification task 1 for every class. This uniform mechanism is not suitable for the imbalanced dataset. Class weighting is a technique to give high weight to the minority class (i.e., positive class). In our study, we give a weight to every class such that it is proportional to the number of examples in that class, i.e., weight of class x = total # samples/(# classes $*$# samples in class x). The model loss is calculated by comparing the model output $\hat{y}$ and the label y of the corresponding image, and y is usually encoded by one-hot code. Conventional one-hot code contains only two values, 0 and 1, which assumes that the distances between all classes are equal. This creates a problem, especially in the multi-classification tasks because it ignores the logical relationship

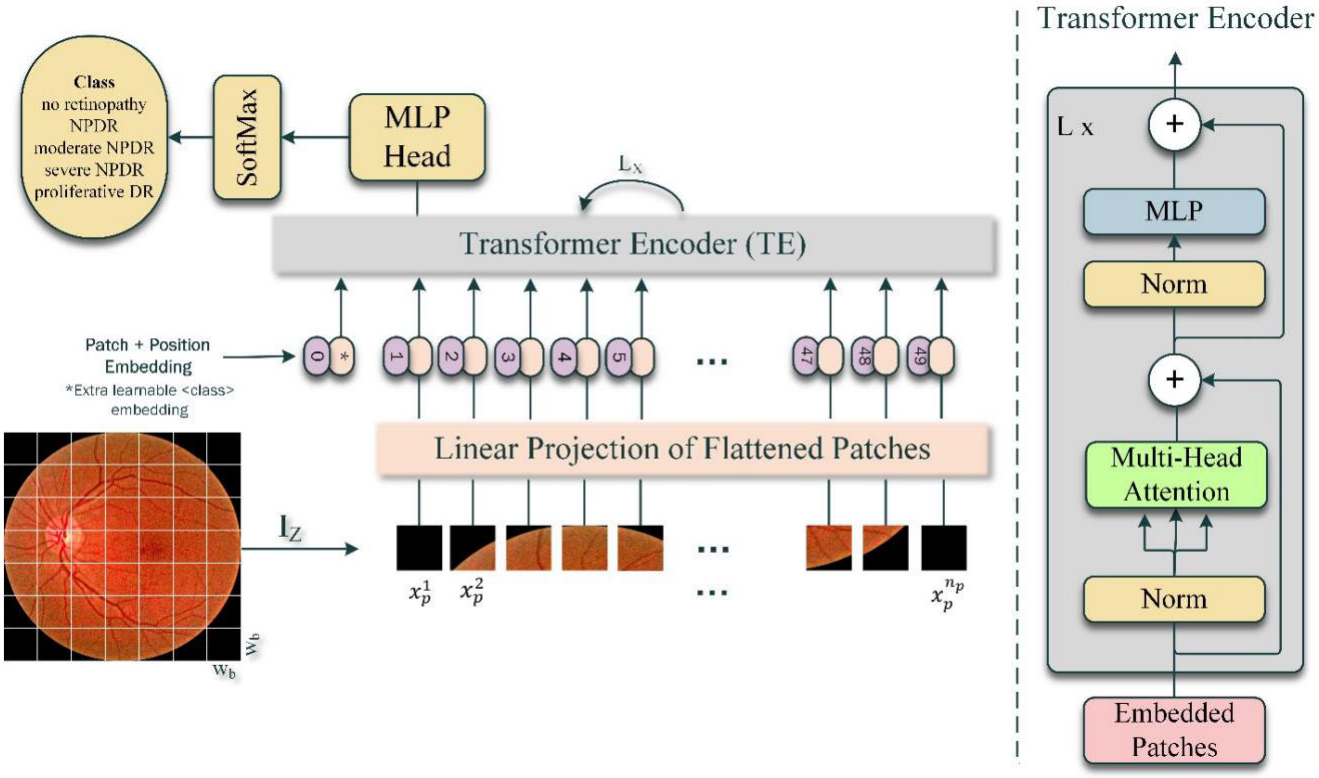


FIGURE 3. Proposed ViT architecture; MLP is for multilayer perceptron.

between the classes. For example, in the five-class DR classification task, the distance between No retinopathy and NPDR, and between Moderate NPDR and Severe NPDR differed considerably. Label smoothing [74] is a technique used in classification tasks to prevent the model from becoming overconfident. Instead of assigning a hard label of 0 or 1 to the true class, label smoothing assigns a smoothed label value between 0 and 1 for each class.

Lin et al., [2] proposed the focal loss to force customized balance among classes in imbalanced datasets, which demonstrates superior performance [3]. Focal loss is a better alternative for cross-entropy. Focal loss concentrates on the examples that the model gets wrong predictions rather than the ones that it can confidently predict. This makes the model gradually enhance its predictions on hard examples using down weighting. Down weighting is an approach that reduces the influence of easy examples of the majority class on the loss function which results in a trained model that gives more attention to hard examples. A modulating factor is added to the cross-entropy loss to implement the focal loss, $-\sum_{i=1}^{n} \alpha_i \, (i - p_i)^{\gamma} \log_p (p_i)$ and γ is the focusing and α is the weighing hyperparameters.

5) VIT MODEL TRAINING

ViT, a network architecture with high capacity, shares similarities with transformers employed in language processing. However, ViT adopts self-attention instead of convolution to gather information across different locations. Within ViT, two essential components play crucial roles: (1) multiheaded self-attention, which facilitates the early aggregation of global information, and (2) residual connections, which effectively propagate features from lower layers to higher layers. The local multiheaded self-attention concept in convolutional neural networks (CNNs), derived from the structure of convolutional receptive fields, combines CNNs with self-attention or applies Transformers to smaller-size images [4]. The fundamental concept behind ViT is to transform the input image into a sequence of image patches, akin to textual words, and extract features using a vision encoder. These features are then fed into a multilayer perceptron (MLP), as illustrated in Figure 3.

- The input image I_z of size 224×224 is converted into a sequence of flattened patches x_p^i, $i = 1, 2, \ldots, n_p$. All patches have the same size of $w_p \times w_p \times c_p$, c_p is the number of channels in I_z, $w_p = 64$ is imperially chosen and $n_p = \left(\frac{256}{64}\right)^2 = 16$, and as fundus images are RGB, $c_p = 3$.
- Each patch x_p^i is flattened and mapped to a D-dimensional latent vector (patch embedding z_0) by transformer layers using a trainable linear projection. $z_0 = \left[x_{class}; x_p^1\mathbb{E}; x_p^2\mathbb{E}; x_p^3\mathbb{E}; \ldots; x_p^{n_p}\mathbb{E}\right] + \mathbb{E}_{pos}$, $\mathbb{E} \in \mathbb{R}^{w_p \times w_p \times c_p \times D}$ is the patch embedding projection, $\mathbb{E}_{pos} \in \mathbb{R}^{(n_p+1)\times D}$ is the position embeddings added to patch embeddings to preserve the positional information of

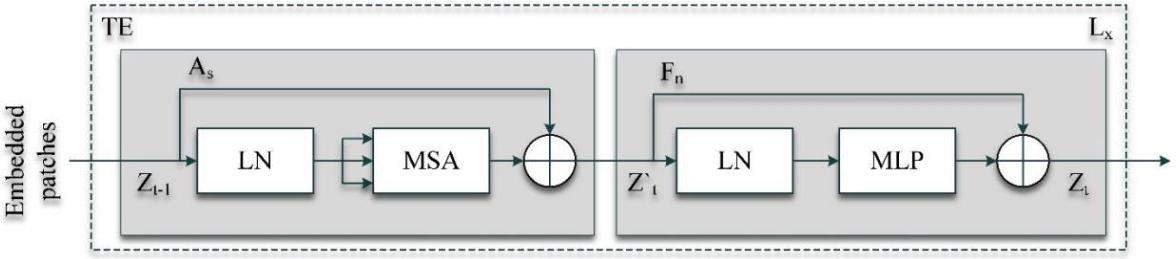


FIGURE 4. Internal architecture of a transformer encoder (TE) [69].

patches, $x_{class} = z_0^0$ is a learnable embedding [5]. At this stage, the patch images are mapped to the embedding space with positional information.

- Add a sequence of transformer encoders [6]. Transformer encoder (TE) has a well-known architecture, see Figure 4, which includes two blocks A_s and F_n containing MSA (Multi-head Self-Attention) and MLP modules, respectively [6]. Layer normalization (LN) and residual connection are used before and after each of these modules, respectively. Formally, $z'_l = MSA\,(LN\,(z_{l-1})) + z_{l-1}$, $z_l = MLP\left(LN\left(z'_l\right)\right) + z'_l$, $l = 1, 2, \ldots, L$, and L is the total number of transformer blocks. In this context, the MLP module consists of two layers with 4D and D neurons, respectively, employing the GELU (Gaussian Error Linear Unit) activation function.
- MSA with h heads is the core component of the transformer encoder. Each head $i \in \{1, 2, \ldots, h\}$ of MSA includes a scaled dot-product attention [6]. A head i calculates a tuple comprising query Q^i, key K^i, and value V^i [6], as follows. $Q^i = XW_Q^i$, $K^i = XW_K^i$, and $V^i = XW_V^i$, X is the input embedding, and W_Q, W_K, W_V are the weight matrices used in the linear transformation process.
- The resulting tuple (Q, K, V) is used as input to the scaled dot-product attention which calculates the attention required to pay to the input image patches. Formally, $SA\,(Q, K, V) = \psi\left(\frac{QK^T}{\sqrt{D_h}}\right)V$, ψ is SoftMax function, and $D_h = \frac{D}{h}$.
- The outcomes of scaled dot-product attentions across all heads are concatenated in MSA, $MSA\,(Q, K, V) = \left[SA^1; SA^2; \ldots; SA^h\right]W_L$, W_L is a weight matrix.
- Additional transformer encoder blocks can be incorporated into the system. Following the application of multiple blocks, the $<$ $class$ $>$ token accumulates contextual information. The resulting state of the learnable embedding from the Transformer encoder, denoted as (z_L^0), serves as the image representation. This image representation, denoted as y, is obtained by applying layer normalization to representation $y = LN\left(z_L^0\right)$.

As illustrated in Figure 3, an MLP head is introduced, comprising a hidden layer consisting of 128 neurons. The output layer of the MLP consists of five neurons, utilizing the SoftMax function to generate a probability distribution for determining the severity level of DR.

C. PERFORMANCE EVALUATION

Four standard metrics were used to evaluate the resulting classifiers including accuracy, balanced accuracy, specificity, precision, recall, AUC, and F1-score, where TP is the true positive, TN is the true negative, FP is the false positive, FN is the false negative, and TPR (truepositiverate) $=$ TP$/$ (TP + FN)(see Equations 1–7).

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

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{2}$$

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

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

$$\text{Specificity (truenegativerate)} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{Balancedaccuracy} = \frac{TPR + TTNR}{2} \tag{6}$$

$$\text{AUC} = \frac{TP}{2 * (TP + FN)} + \frac{TN}{2 * (FP + TN)} \tag{7}$$

III. RESULTS AND DISCUSSION

This section focuses on conducting a series of experiments to optimize the architecture of the proposed ViT model and attain optimal performance. Furthermore, a comparison is made between the performance of the proposed model and several state-of-the-art classification models. Various experiments were performed, involving different training optimizers and hyperparameters, aiming to identify the most suitable configuration.

A. EXPERIMENTAL SETUP

All experiments were carried out on the Google Colab platform, utilizing a Tesla V100 with 16 GB of memory and 12 GB RAM. The implementation of the experiments involved the utilization of Keras and TensorFlow libraries [7]. Images of FGADR dataset are downsized into 224 $\times$ 224 and the dataset was split into three parts, 80% for training

TABLE 2. Two architectures of ViT model with different sizes.

Model	Layers	Hidden Size	MLP size	# of Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M

TABLE 3. Comparing performances of ViT-Base and ViT-Large models with 16 and 32 image patches.

Model, patch size	F1-Score	Accuracy	Balanced Accuracy	AUC	Precision	Recall	Specificity
ViT-Base, 16	0.616	0.616	0.597	0.841	0.616	0.616	0.9
ViT-Base, 32	0.789	0.789	0.791	0.951	0.789	0.789	0.947
ViT-Large, 16	0.678	0.678	0.679	0.913	0.678	0.678	0.919
ViT-Large, 32	**0.804**	**0.804**	**0.805**	**0.959**	**0.804**	**0.804**	**0.951**

dataset, 10% for validation dataset to fine-tune the model's hyperparameters, and 10% for testing dataset to test the generalization performance of the proposed model. Results are reported in terms of the testing performance.

In the training process of the ViT models, several considerations were considered to optimize the trainable parameters and achieve the highest performance. To minimize the loss, the training process employed the minibatch optimization approach. This approach involves dividing the training data into smaller batches or subsets, in this case, a batch size of 32 training samples was used. The model's parameters were updated based on the average gradient calculated from each batch, which helps in reducing the computational burden and memory requirements compared to processing the entire dataset at once. Cross-entropy loss is used in the training process. It is commonly used in multi-class classification problems because it effectively measures the dissimilarity between the predicted probabilities and the true class labels. By minimizing this loss, the model learned to assign higher probabilities to the correct classes, improving its overall accuracy. The model's trainable parameters, such as weights and biases, are tuned using the RAdam optimizer [8]. The number of epochs was 100 epochs. Learning rates from 1e-4 to 4e-4 were tried, and the learning rate of 3e-4 resulted in the best performance.

B. EXPERIMENT 1: ViT MODEL AND PATCH SIZE SELECTION

The first experiment was conducted to choose between two variants of ViT architectures, i.e., ViT-Base and ViT-Large. Both model variants are pretrained using ImageNet21k and ImageNet2012 datasets. ViT-Base is a base model with 86 million parameters, and ViT-Large is a larger model with 307 million parameters. The architecture information of the two ViT variants is shown in Table 2. We added three dense layers to train the classifier (i.e., 128 units, 64 units, 32 units). Two batch normalization layers have been added, one before the first dense layer and another layer after the first dense layer. The GeLu activation function has been used in all internal layers and SoftMax has been used in the extra output layer.

Table 3 shows the performance of different ViT architectures using different batch sizes. We tested the performance of each model using 16×16 and 32×32 patches. The ViT-Large model with 32×32 patches exhibited the best performance, as indicated in Table 3 (i.e., 0.804, 0.804, 0.805, 0.959, 0.804, 0.804, and 0.951 for F1-score, accuracy, balanced accuracy, AUC, precision, recall, specificity, respectively). Because ViT-Large has a more complex architecture, larger batch size achieved better results.

Overfitting has been prevented using several regularization techniques including dropout and L2 regularization. Consequently, this model (i.e., ViT-Large) was selected for further analysis, and all subsequent experiments were conducted using this model. Several fine-tuning experiments of ViT-Large model, especially the optimizer using the Radam algorithm, were carried out in order to enhance the performance of ViT-Large model and to beat the reported performance of the same problem at hand (i.e., 0.81 F1-score) [9].

C. EXPERIMENT 2: WEIGHT DECAY SELECTION FOR THE LARGE ViT MODEL

Optimizer selection is a crucial step for training a deep learning model. Best optimizer finds the best weights quickly. Different optimizers have been explored. For example, Adam is a well-known optimizer which has been used to optimize the initial model. Based on the performance of the resulting model, another variant of Adam optimizer, i.e., AdamW, was explored and utilized in this experiment [10]. RAdam is known for its fast convergence rate and lower memory requirements. AdamW includes regularization which is a modified version of L2 regularization. In the formulation of L2 regularization, the weight-decay is added to the gradient update term, whilst in the formulation of AdamW, the weight-decay is decoupled from the gradient update term. Several different values of weight-decay parameter were explored in this experiment, as shown in Table 4. In addition, the learning rate determines the step size at which the model's parameters are updated during optimization. Learning rates from 1e-4 to 4e-4 were explored, where the learning rate of 3e-4 resulted in the best performance.

TABLE 4. Performances of ViT-Large model with AdamW optimizer.

Weight-Decay	F1-Score	Accuracy	Balanced Accuracy	AUC	Precision	Recall	Specificity
1e-3	0.675	0.675	0.676	0.923	0.675	0.675	0.919
2e-3	0.620	0.620	0.620	0.884	0.620	0.620	0.905
1e-4	**0.783**	**0.783**	**0.785**	**0.958**	**0.783**	**0.783**	**0.946**
5e-5	0.777	0.777	0.779	0.957	0.777	0.777	0.944
2e-4	0.723	0.723	0.724	0.940	0.723	0.723	0.930
3e-4	0.747	0.747	0.748	0.951	0.747	0.747	0.937
4e-4	0.735	0.735	0.737	0.943	0.735	0.735	0.934

TABLE 5. Comparing performances of ViT-Large model with RAdam optimizer and dropout.

# layers, dropout	F1-Score	Accuracy	Balanced Accuracy	AUC	Precision	Recall	Specificity
3, 0.3	0.669	0.669	0.671	0.92	0.669	0.669	0.92
2, 0.1	0.744	0.744	0.746	0.946	0.744	0.744	0.936
2, 0.2	**0.759**	**0.759**	**0.761**	**0.946**	**0.759**	**0.759**	**0.94**
2, 0.5	0.627	0.627	0.629	0.898	0.627	0.627	0.907

TABLE 6. Comparing performances of ViT-Large model with RAdam optimizer and regularization.

L2 Regularization	F1-Score	Accuracy	Balanced Accuracy	AUC	Precision	Recall	Specificity
0.1	0.624	0.624	0.625	0.907	0.624	0.624	0.906
0.001	**0.726**	**0.726**	**0.728**	**0.942**	**0.726**	**0.726**	**0.931**

The highest achieved result was an F1-Score of 0.783 with a weight decay value of 1e-4. None of the experiments surpassed the best outcome reported in experiment 1, where the F1-Score was 0.804. Therefore, RAdam was employed in all subsequent experiments due to its superior performance.

D. EXPERIMENT 3: SELECTION OF THE DROPOUT AND REGULARIZATION

Large models such as ViT are prone to overfitting. In the next two experiments, dropout and L2 regularization techniques were utilized to improve the model performance and ensure overfitting prevention. Three and two layers of dropout were added to the ViT model with different dropout values as shown Table 5. In addition, Table 6 shows the results of an experiment to select the best regularization parameter. L2 regularization was utilized with two different values. Introducing two layers of dropout with a value of 0.2 yielded the best F1-Score of 0.759. As it can be noticed, because the proposed model is based on the transfer learning where the layers of the ViT are frozen, a small probability is needed for the dropout layer to provide the best results. L2 regularization with a value of 0.001 resulted in an F1-Score of 0.726, see Table 6. However, none of these techniques improved upon the model performance observed in experiment 1 (i.e., F1-Score = 0.804). The exploration of the above hyperparameters indicates that the ViT model is well trained to solve our medical problem and indicates that the architecture of the added extra dense layers has minor impact on the performance of the overall model. In other words, the extracted deep representation from the ViT model is good enough to train simple classifier.

E. EXPERIMENT 4: ROBUSTNESS IMPROVEMENT USING LABEL SMOOTHING

Label smoothing [11] is a regularization technique commonly used in classification tasks to prevent the model from becoming overconfident in its predictions. Instead of assigning a hard label of 0 or 1 to the true class, label smoothing assigns a smoothed label value between 0 and 1. This encourages the model to learn more robust and generalizable representations. Different values of label smoothing were tried in this experiment as shown in Table 7.

When employing label smoothing with a value of 0.1, the best F1-score achieved in this experiment was 0.738. This result was inferior to the best outcome reported in experiment 1, which utilized label smoothing with a value of 0.2. Previous experiments did not achieve better results than the best outcome obtained in experiment 1, which reported an F1 score of 0.804. Consequently, in subsequent experiments, different techniques were employed to address the class imbalance within the FGADR dataset, as illustrated in Table 1. This class imbalance presents challenges during training, as the model may exhibit bias towards the majority classes and encounter difficulties in accurately predicting the minority classes. Furthermore, we made modifications to the training process with the

TABLE 7. Comparing performances of ViT-Large model with label smoothing.

Label Smoothing	F1-Score	Accuracy	Balanced Accuracy	AUC	Precision	Recall	Specificity
0.0	0.732	0.732	0.734	0.948	0.732	0.732	0.933
0.1	**0.738**	**0.738**	**0.740**	**0.944**	**0.738**	**0.738**	**0.935**
0.4	0.729	0.729	0.731	0.938	0.729	0.729	0.932

TABLE 8. Comparing performances of ViT-Large model with class weights.

Class Weights	F1-Score	Acc	Balanced Accuracy	AUC	Precision	Recall	Specificity
2.0, 1.5, 1.0, 1.0, 1.5	0.783	0.783	0.784	0.963	0.783	0.783	0.946
2.0, 2.0, 1.0, 1.0, 2.0	0.765	0.765	0.765	0.953	0.765	0.765	0.941
3.64, 1.73, 0.62, 0.57, 1.28	0.783	0.783	0.783	0.956	0.783	0.783	0.946
3.64, 1.73, 0.91, 0.86, 1.28	**0.825**	**0.825**	**0.826**	**0.964**	**0.825**	**0.825**	**0.956**
3.65, 1.75, 1.0, 0.95, 1.25	0.798	0.798	0.799	0.961	0.798	0.798	0.950
3.64, 1.73, 1.0, 0.86, 1.28	0.801	0.801	0.802	0.962	0.801	0.801	0.950

objective of maximizing the F1 score of validation datasets, as opposed to minimizing validation loss in previous experiments.

F. EXPERIMENT 5: CLASS WEIGHT FOR DATASET BALANCING

One approach to address class imbalance is using class weighting. Class weights assign different weights to each class based on their prevalence in the dataset. By assigning higher weights to the minority class and lower weights to the majority class, the model is encouraged to pay more attention to the minority class during training, thus mitigating the impact of class imbalance. Manual weights were tried in the beginning which assign every class a weight value between 1 and 2. Then, automatic weights were calculated according the equation of $classweight = totalsamples/(numclasses*classsamples)$. The best result was achieved using a manually fine-tuned version of the automatic weights which achieved F1-score of 0.825 which surpassed the best reported performance (i.e., 0.81 F1-Score), as shown in Table 8.

Two versions of manual weight adjustment (i.e., class weights ranging from 1 to 2) yielded F1-Scores of 0.783 and 0.765, respectively. However, automatic weights performed better.

Automatic weights were calculated based on the number of samples in each class (i.e., 3.64, 1.73, 0.62, 0.57, 1.28), and they achieved the same F1-Score as the manual weights. These weights were fine-tuned to pay more attention to the minority classes during training. One of these trials (weights: 3.64, 1.73, 0.91, 0.86, 1.28) surpassed the best F1-Score reported in experiment 1 (F1-Score = 0.804) and the accuracy reported in (accuracy = 0.810) [12]. The confusion matrix of the ViT model using these weights is depicted in Figure 5.

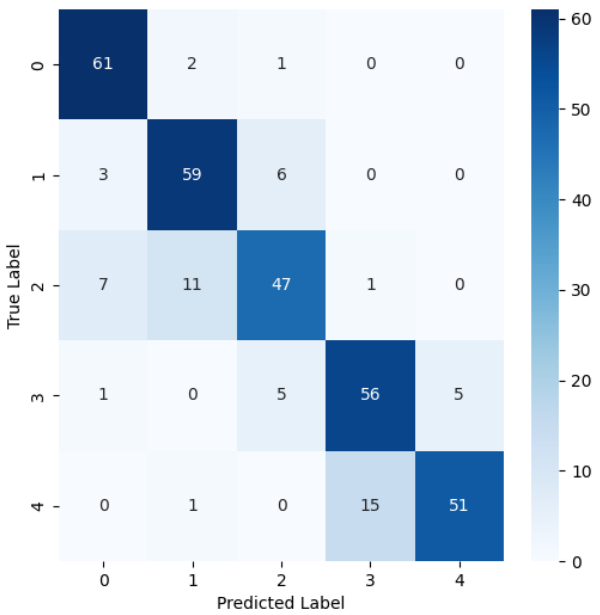


FIGURE 5. Confusion matrix of proposed ViT model with finetuned class weights on the test dataset.

G. EXPERIMENT 6: TWO-PHASE ViT PRE-TRAINING FOR DATASET BALANCING

The most recent experiment aimed to tackle the imbalanced dataset by employing a two-phase ViT pre-training approach [13]. In this approach, the ViT model was initially pre-trained using a large fundus database (i.e., transfer step), resulting in substantial improvements in model performance when fine-tuned for the downstream retinal disease classification task (i.e., adaptation step). For the initial step, a preprocessed version of the Kaggle EyePACS[1] dataset was utilized. This copy exclusively consisted of the training dataset, comprising

[1] https://www.kaggle.com/datasets/mariaherrerot/eyepacspreprocess

TABLE 9. Performance of ViT-Large model with RAdam optimizer and focal loss.

Alpha, Gamma	F1-score	Accuracy	Balanced Accuracy	AUC	Precision	Recall	Specificity
0.25, 2	**0.786**	**0.786**	**0.787**	**0.960**	**0.786**	**0.786**	**0.947**
0.5, 2	0.786	0.786	0.787	0.960	0.786	0.786	0.947
0.75, 3	0.750	0.750	0.752	0.953	0.750	0.750	0.938
0.05, 0.5	0.744	0.744	0.746	0.949	0.744	0.744	0.936

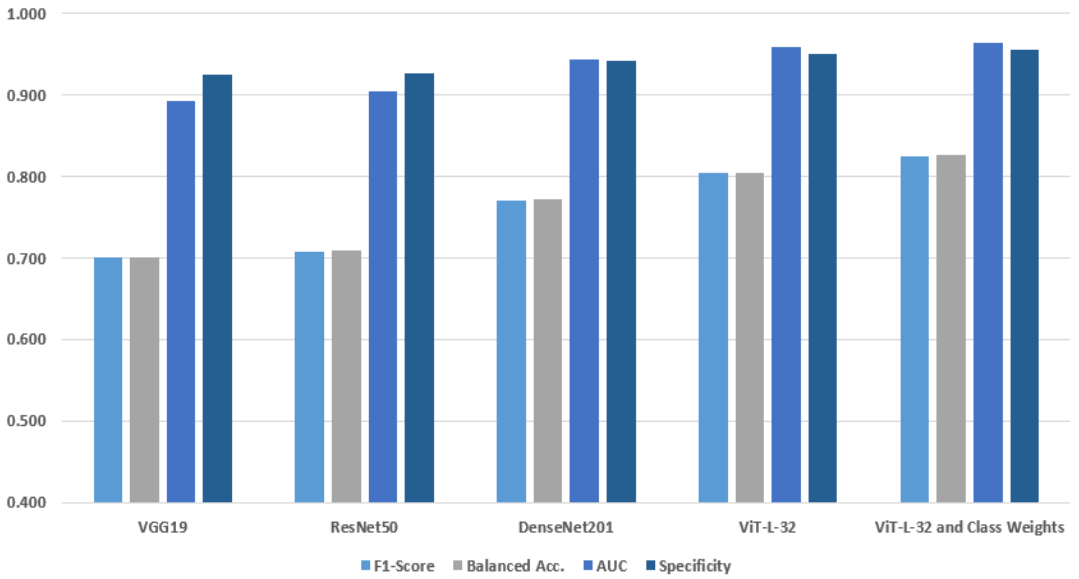


FIGURE 6. Comparison between different ViT pipelines.

35,126 labeled images. Moreover, all images were resized to 1024 × 1024 and cropped to remove significant amounts of black space. The images were sourced from EyePACS and captured under diverse conditions by various devices at multiple primary care sites across California and other locations. Each subject contributed two images of their left and right eyes, both with the same resolution. A clinician assessed each image for the presence of DR, employing a scale of 0–4. Constructing ViT models using this approach necessitates substantial computational resources, resulting in only a limited number of ViT models being built using the EyePACS dataset. These models were subsequently trained on the FGADR dataset to enhance classification performance. One of these models achieved an improved F1-Score of 0.828 on the FGADR dataset.

H. EXPERIMENT 7: FOCAL LOSS FOR DATASET BALANCING

Focal loss was tried in this experiment with different alpha and gamma combinations. By using focal loss, the model can effectively focus more on the minority class and mitigate the dominant influence of the majority class.

As illustrated in Table 9, the highest F1-Score obtained was 0.792 with an alpha of 0.25 and a gamma of 2. This suggests that utilizing focal loss with the ViT model did not result in a better F1-Score.

I. EXPERIMENT 8: COMPARISON WITH OTHER MODELS

The proposed models were compared with other models, as shown in Table 10. Additionally, a comparison was made with other published results that employed the FGADR dataset [12]. The proposed model demonstrated superior performance in terms of F1-Score, while models incorporating segmentation in addition to classification achieved better specificity.

A summary of the results from the most significant fine-tuned ViT models is presented in Figure 6. The large ViT model with a patch size of 32 achieved the highest results among the four ViT variants presented in experiment 1 and was used as the benchmark for sub-sequent experiments. The utilization of the AdamW optimizer and focal loss did not surpass the ViT model, while fine-tuned class weights outperformed the benchmark ViT, achieving an F1-Score of 0.825. As shown in Figure 6, the proposed ViT model outperformed the performance of the aforementioned CNN models. The proposed model outperformed VGG19 by 0.125, 0.125, 0.126, 0.071, 0.125, 0.125, and 0.031 for F1-score, accuracy, balanced accuracy, AUC, precision, recall, and specificity, respectively.

TABLE 10. Comparing of the proposed ViT model with other CNN models.

Model	F1-Score	Accuracy	Balanced Accuracy	AUC	Precision	Recall	Specificity
MobileNetV3Large	0.536	0.536	0.538	0.838	0.536	0.536	0.884
InceptionV3	0.680	0.680	0.680	0.887	0.680	0.680	0.920
ConvNeXtLarge	0.689	0.689	0.691	0.891	0.689	0.689	0.922
VGG19	0.700	0.700	0.700	0.893	0.700	0.700	0.925
ResNet50	0.708	0.708	0.710	0.904	0.708	0.708	0.927
DenseNet201	0.771	0.771	0.772	0.944	0.771	0.771	0.942
Segmentation and Densenet121 [12]	-	0.810	-	-	-	0.810	0.980
Segmentation and InceptionV3 [12]	-	0.810	-	-	-	0.842	0.990
Proposed ViT	**0.825**	**0.825**	**0.826**	**0.964**	**0.825**	**0.825**	**0.956**

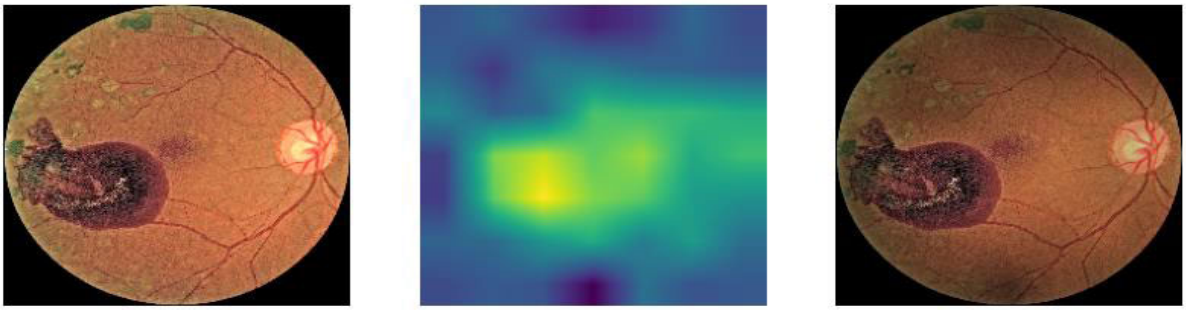


FIGURE 7. Visualization of attention map for an input image.

It outperformed ResNet50 by 0.117, 0.117, 0.116, 0.06, 0.117, 0.117, and 0.029 for F1-score, accuracy, balanced accuracy, AUC, precision, recall, and specificity, respectively. In addition, it outperformed DenseNet201 by 0.054, 0.054, 0.054, 0.02, 0.054, 0.054, and 0.014 for F1-score, accuracy, balanced accuracy, AUC, precision, recall, and specificity, respectively.

The resulting mode is more efficient than CNN and all the literature studies. We added an extra step in the application of vision transformers by applying it in this new domain and achieving state-of-the-art results. The resulting model can be applied in real medical environments to provide accurate and fast assistance to ophthalmologists. Note that our deep learning model is based directly on the ViT transformer network architecture. We did not retrain the model from scratch, but we used the transfer learning technique to reuse the pretrained ViT model with its pre-learned weights. We optimized a set of dense layers which have been used to train the classifier for predicting the severity of diabetic retinopathy. This process has been followed by huge literature studies in the domain of medical image analysis [81]. The current study proved the suitability of transformer models to analyze medical images.

At this point, we optimized a robust DL model, but model trustworthy needs an extra step of explainability. Decision visualization is an important technique to highlight the important regions in the image where the model has used to make its decision. The attention mechanism in ViT is a critical component that enables these models to process and understand visual data. It involves multi-head self-attention, where input image patches are transformed into Query, Key, and Value vectors. Attention scores are computed based on the relevance of patches to each other, and a weighted sum of Values captures informative relationships. This process is repeated across multiple attention heads and layers. Figure 7 shows, from left to right, an input image, ViT output attention as heatmap, and the input image masked by ViT attention. Lighter regions have more attention scores than darker regions and therefore they are more relevant to the final disease diagnosis. We have a survey for the recent advances in model explainability [84]. In the future, we will explore the application of these techniques to enhance the understandability of the deep learning models.

IV. CONCLUSION

In this study, we explored the problem of automated severity stage detection of DR from fundus images. We proposed a vision transformer deep learning pipeline to capture long-range dependencies in images. The study used the transfer learning technique to train a large vision model on a relatively small dataset. For experimentation, the model has been trained using the new real-world FGADR dataset. We checked the performance of the proposed model using the original imbalanced data. Then, we improved the model

performance with collection of data and algorithm-based balancing techniques. The proposed ViT model achieved superior results than other baseline ViT and state-of-the-art CNN architectures. To conclude, this study explored (1) the role of learning the long-term spatial dependencies in medical images using transformers to improve the performance of disease detection, (2) the role of data balancing techniques (i.e., data centric and algorithmic) to train more stable and accurate models, (3) the role of hyperparameter optimization to improve the performance of large models with frozen weights of early layers, and (3) the role of transfer learning to simplify the model optimization process. Our ViT model achieved superior results compared to the CNN and ViT baseline architectures (i.e., 0.825, 0.825, 0.826, 0.964, 0.825, 0.825, and 0.956 for F1-score, accuracy, balanced accuracy, AUC, precision, recall, specificity, respectively). It is important to study the tradeoff between the model performance and its complexity. In the future, we will investigate the role of model complexity to enhance the results. This will need the testing of ViT architectures in different experimental environments and with different datasets. We will explore alternative approaches for compressing transformers and developing lightweight models. We will improve the robustness of the model by testing its performance using external validation and measuring its uncertainty. Additionally, we will investigate the performance of ViT across different numbers of layers and heads. As our ViT model did not perform any lesion segmentation, we will explore differ segmentation techniques for fundus images to improve the detection of DR and to improve the understandability of the resulting decision by providing visual explainability of the model decision. Before AI model deployment, it must be tested regarding its fairness against any bias in data or algorithm, the robustness against adversarial attacks, the capability to protect the privacy of patient, the stability and robustness, and the ability to quantify and enhance model's uncertainty. The extension of our model to be trustworthy will be considered in a separate future study.

REFERENCES

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, Nov. 2019, Art. no. 107843, doi: 10.1016/j.diabres.2019.107843.

[2] I. D. Atlas. (2023). *IDF Diabetes Atlas*. [Online]. Available: https://diabetesatlas.org/

[3] WHO. (2023). *Diabetes*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100377.

[5] P. Uppamma and S. Bhattacharya, "Deep learning and medical image processing techniques for diabetic retinopathy: A survey of applications, challenges, and future trends," *J. Healthcare Eng.*, vol. 2023, Feb. 2023, Art. no. 2728719.

[6] T.-E. Tan and T. Y. Wong, "Diabetic retinopathy: Looking forward to 2030," *Frontiers Endocrinol.*, vol. 13, Jan. 2023, Art. no. 1077669.

[7] G. Selvachandran, S. G. Quek, R. Paramesran, W. Ding, and L. H. Son, "Developments in the detection of diabetic retinopathy: A state-of-the-art review of computer-aided diagnosis and machine learning methods," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 915–964, Feb. 2023.

[8] *IDF Diabetes Atlas*, 9th ed., 2023. [Online]. Available: https://diabetesatlas.org/atlas/ninth-edition/

[9] N. Congdon, Y. Zheng, and M. He, "The worldwide epidemic of diabetic retinopathy," *Indian J. Ophthalmol.*, vol. 60, no. 5, pp. 428–431, 2012.

[10] D. S. W. Ting, G. C. M. Cheung, and T. Y. Wong, "Diabetic retinopathy: Global prevalence, major risk factors, screening practices and public health challenges: A review," *Clin. Exp. Ophthalmol.*, vol. 44, no. 4, pp. 260–277, May 2016.

[11] M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep learning techniques for diabetic retinopathy classification: A survey," *IEEE Access*, vol. 10, pp. 28642–28655, 2022.

[12] S. Tummala, V. S. G. Thadikemalla, S. Kadry, M. Sharaf, and H. T. Rauf, "EfficientNetV2 based ensemble model for quality estimation of diabetic retinopathy images from DeepDRiD," *Diagnostics*, vol. 13, no. 4, p. 622, Feb. 2023.

[13] A. Sebastian, O. Elharrouss, S. Al-Maadeed, and N. Almaadeed, "A survey on deep-learning-based diabetic retinopathy classification," *Diagnostics*, vol. 13, no. 3, p. 345, Jan. 2023.

[14] C. Raja and L. Balaji, "An automatic detection of blood vessel in retinal images using convolution neural network for diabetic retinopathy detection," *Pattern Recognit. Image Anal.*, vol. 29, no. 3, pp. 533–545, Jul. 2019.

[15] T. Nazir, A. Irtaza, Z. Shabbir, A. Javed, U. Akram, and M. T. Mahmood, "Diabetic retinopathy detection through novel tetragonal local octa patterns and extreme learning machines," *Artif. Intell. Med.*, vol. 99, Aug. 2019, Art. no. 101695.

[16] S. Gayathri, V. P. Gopi, and P. Palanisamy, "Diabetic retinopathy classification based on multipath CNN and machine learning classifiers," *Phys. Eng. Sci. Med.*, vol. 44, no. 3, pp. 639–653, Sep. 2021.

[17] P. S. Washburn, Mahendran, Dhanasekharan, Periyasamy, and Murugeswari, "Investigation of severity level of diabetic retinopathy using Adaboost classifier algorithm," *Mater. Today, Proc.*, vol. 33, pp. 3037–3042, 2020.

[18] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[19] K. Xu, D. Feng, and H. Mi, "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image," *Molecules*, vol. 22, no. 12, p. 2054, Nov. 2017.

[20] N. S. P. Kazakh-British, A. A. Pak, and D. Abdullina, "Automatic detection of blood vessels and classification in retinal images for diabetic retinopathy diagnosis with application of convolution neural network," in *Proc. Int. Conf. Sensors, Signal Image Process.*, Oct. 2018, pp. 60–63.

[21] Padmanayana and B. K. Dr. Anoop, "Binary classification of diabetic retinopathy using convolutional neural networks with fundus color images," *Mater. Today, Proc.*, vol. 58, pp. 212–216, Jan. 2022.

[22] S. Rêgo, M. Dutra-Medeiros, F. Soares, and M. Monteiro-Soares, "Screening for diabetic retinopathy using an automated diagnostic system based on deep learning: Diagnostic accuracy assessment," *Ophthalmologica*, vol. 244, no. 3, pp. 250–257, 2021.

[23] A. M. Pamadi, A. Ravishankar, P. A. Nithya, G. Jahnavi, and S. Kathavate, "Diabetic retinopathy detection using MobileNetV2 architecture," in *Proc. Int. Conf. Smart Technol. Syst. Next Gener. Comput. (ICSTSN)*, Mar. 2022, pp. 1–5.

[24] P. Saranya, S. K. Devi, and B. Bharanidharan, "Detection of diabetic retinopathy in retinal fundus images using DenseNet based deep learning model," in *Proc. Int. Mobile Embedded Technol. Conf. (MECON)*, Mar. 2022, pp. 268–272.

[25] W. Mudaser, P. Padungweang, P. Mongkolnam, and P. Lavangnananda, "Diabetic retinopathy classification with pre-trained image enhancement model," in *Proc. IEEE 12th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Dec. 2021, pp. 0629–0632.

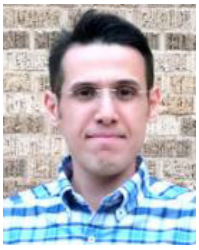
[26] Y. S. Boral and S. S. Thorat, "Classification of diabetic retinopathy based on hybrid neural network," in *Proc. 5th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Apr. 2021, pp. 1354–1358.

[27] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, and W. Qian, "An interpretable ensemble deep learning model for diabetic retinopathy disease classification," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2045–2048.

[28] H. Kaushik, D. Singh, M. Kaur, H. Alshazly, A. Zaguia, and H. Hamam, "Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models," *IEEE Access*, vol. 9, pp. 108276–108292, 2021.
[29] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, and Z. Yi, "Automated identification and grading system of diabetic retinopathy using deep neural networks," *Knowl.-Based Syst.*, vol. 175, pp. 12–25, Jul. 2019.
[30] V. Bellemo, Z. W. Lim, G. Lim, Q. D. Nguyen, Y. Xie, M. Y. T. Yip, H. Hamzah, J. Ho, X. Q. Lee, W. Hsu, M. L. Lee, L. Musonda, M. Chandran, G. Chipalo-Mutati, M. Muma, G. S. W. Tan, S. Sivaprasad, G. Menon, T. Y. Wong, and D. S. W. Ting, "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: A clinical validation study," *Lancet Digit. Health*, vol. 1, no. 1, pp. e35–e44, May 2019.
[31] Y. Xie, Q. D. Nguyen, H. Hamzah, G. Lim, V. Bellemo, D. V. Gunasekeran, M. Y. T. Yip, X. Qi Lee, W. Hsu, M. L. Lee, C. S. Tan, H. T. Wong, E. L. Lamoureux, G. S. W. Tan, T. Y. Wong, E. A. Finkelstein, and D. S. W. Ting, "Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: An economic analysis modelling study," *Lancet Digit. Health*, vol. 2, no. 5, pp. e240–e249, May 2020.
[32] C. P. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, Sep. 2003.
[33] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 818–828, Mar. 2021, doi: 10.1109/TMI.2020.3037771.
[34] C. Harshitha, A. Asha, J. L. S. Pushkala, R. N. S. Anogini, and C. Karthikeyan, "Predicting the stages of diabetic retinopathy using deep learning," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 1–6.
[35] X. Luo, Z. Pu, Y. Xu, W. K. Wong, J. Su, X. Dou, B. Ye, J. Hu, and L. Mou, "MVDRNet: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108104.
[36] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Â. Carneiro, A. M. Mendonça, and A. Campilho, "GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101715.
[37] N. S. Shaik and T. K. Cherukuri, "Hinge attention network: A joint model for diabetic retinopathy severity grading," *Int. J. Speech Technol.*, vol. 52, no. 13, pp. 15105–15121, Oct. 2022.
[38] Y. Li, Z. Song, S. Kang, S. Jung, and W. Kang, "Semi-supervised auto-encoder graph network for diabetic retinopathy grading," *IEEE Access*, vol. 9, pp. 140759–140767, 2021.
[39] X. Wang, Y. Lu, Y. Wang, and W.-B. Chen, "Diabetic retinopathy stage classification using convolutional neural networks," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 465–471.
[40] A. H. Asad, A. T. Azar, N. El-Bendary, and A. E. Hassaanien, "Ant colony based feature selection heuristics for retinal vessel segmentation," 2014, *arXiv:1403.1735*.
[41] E. Dugas, J. Jared, and W. Cukierski. (2015). Diabetic Retinopathy Detection. Kaggle. [Online]. Available: https://kaggle.com/competitions/diabetic-retinopathy-detection
[42] M. Karthik and S. Dane. (2019). APTOS 2019 Blindness Detection. Kaggle. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection
[43] *The STARE Project*, Shiley Eye Center, San Diego, CA, USA, 2004.
[44] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "DIARETDB0: Evaluation database and methodology for diabetic retinopathy algorithms," Mach. Vis. Pattern Recognit. Res. Group, Lappeenranta Univ. Technol. Finland, Tech. Rep. 73, 2006, pp. 1–17.
[45] L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, S. Garg, K. W. Tobin, and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," *Med. Image Anal.*, vol. 16, no. 1, pp. 216–226, Jan. 2012.
[46] M. Niemeijer et al., "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Jan. 2010.
[47] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, B. Charton, and J.-C. Klein, "Feedback on a publically distributed image database: The Messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, Aug. 2014.
[48] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis, "TeleOphta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, Apr. 2013.
[49] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci.*, vol. 501, pp. 511–522, Oct. 2019.
[50] R. Liu et al., "DeepDRiD: Diabetic retinopathy–grading and image quality estimation challenge," *Patterns*, vol. 3, no. 6, Jun. 2022, Art. no. 100512.
[51] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian Diabetic Retinopathy Image Dataset (IDRiD): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, Jul. 2018.
[52] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, L. Giancardo, G. Quellec, and F. Mériaudeau, "Retinal fundus multi-disease image dataset (RFIDM): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, Feb. 2021.
[53] Z. Han, B. Yang, S. Deng, Z. Li, and Z. Tong, "Category weighted network and relation weighted label for diabetic retinopathy screening," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106408.
[54] S. Atasever, N. Azginoglu, D. S. Terzi, and R. Terzi, "A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning," *Clin. Imag.*, vol. 94, pp. 18–41, Feb. 2023.
[55] A. Solano, K. N. Dietrich, M. Martínez-Sober, R. Barranquero-Cardeñosa, J. Vila-Tomás, and P. Hernández-Cámara, "Deep learning architectures for diagnosis of diabetic retinopathy," *Appl. Sci.*, vol. 13, no. 7, p. 4445, Mar. 2023.
[56] M. M. Islam, H.-C. Yang, T. N. Poly, W.-S. Jian, and Y.-C. J. Li, "Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis," *Comput. Methods Programs Biomed.*, vol. 191, Jul. 2020, Art. no. 105320.
[57] N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, F. Scarpa, A. Scarpa, D. I. Fotiadis, and K. Marias, "Deep learning for diabetic retinopathy detection and classification based on fundus images: A review," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104599.
[58] D. Das, S. K. Biswas, and S. Bandyopadhyay, "A critical review on the diagnosis of diabetic retinopathy using machine learning and deep learning," *Multimedia Tools Appl.*, vol. 81, no. 18, pp. 25613–25655, Jul. 2022.
[59] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12116–12128.
[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[61] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 30, 2023, doi: 10.1109/TNNLS.2022.3227717.
[62] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802.
[63] R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, and Y. Zhang, "Lesion-aware transformers for diabetic retinopathy grading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10933–10942.
[64] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, and S. A. Baker, "VTGAN: Semi-supervised retinal image synthesis and disease prediction using vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3228–3238.
[65] S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, and Y. Zheng, "MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Strasbourg, France, Sep. 2021, pp. 45–54.

[66] A. Papadopoulos, F. Topouzis, and A. Delopoulos, "An interpretable multiple-instance approach for the detection of referable diabetic retinopathy in fundus images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, Jul. 2021.
[67] N. S. Kumar and B. R. Karthikeyan, "Diabetic retinopathy detection using CNN, transformer and MLP based architectures," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2021, pp. 1–2.
[68] Z. Zhang, G. Sun, K. Zheng, J.-K. Yang, X.-R. Zhu, and Y. Li, "TC-Net: A joint learning framework based on CNN and vision transformer for multi-lesion medical images segmentation," *Comput. Biol. Med.*, vol. 161, Jul. 2023, Art. no. 106967.
[69] C. Adak, T. Karkera, S. Chattopadhyay, and M. Saqib, "Detecting severity of diabetic retinopathy from fundus images using ensembled transformers," 2023, *arXiv:2301.00973*.
[70] Z. Gu, Y. Li, Z. Wang, J. Kan, J. Shu, and Q. Wang, "Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention," *Comput. Intell. Neurosci.*, vol. 2023, Jan. 2023, Art. no. 1305583.
[71] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—A contemplative retrospection," *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 106126.
[72] E. Z. Ye, J. Ye, and E. H. Ye, "Applications of vision transformers in retinal imaging: A systematic review," Blyth Acad., Toronto, ON, Canada, Tech. Rep., 2023, doi: 10.22541/au.167528318.80645903/v1.
[73] G. M. Weiss, H. He, and Y. Ma, "Foundations of imbalanced learning," in *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2013.
[74] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.
[75] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[76] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.
[77] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*.
[78] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[79] H. Gandhi, K. Agrawal, U. Oza, and P. Kumar, "Diabetic retinopathy classification using pixel-level lesion segmentation," in *Proc. 4th Int. Conf. Futuristic Trends Netw. Comput. Technol. (FTNCT)*, 2022, pp. 405–417.
[80] S. Garg, T. Vu, and A. Moschitti, "TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 7780–7788.
[81] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, and J. Zhang, "Transformers in medical image analysis," *Intell. Med.*, vol. 3, no. 1, pp. 59–78, 2023.
[82] M. Mediouni, R. Madiouni, M. Gardner, and N. Vaughan, "Translational medicine: Challenges and new orthopaedic vision (Mediouni-model)," *Current Orthopaedic Pract.*, vol. 31, no. 2, pp. 196–200, Mar. 2020.
[83] M. Mediouni, D. R. Schlatterer, H. Madry, M. Cucchiarini, and B. Rai, "A review of translational medicine. The future paradigm: How can we connect the orthopedic dots better?" *Current Med. Res. Opinion*, vol. 34, no. 7, pp. 1217–1229, Jul. 2018.
[84] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease," *Sci. Rep.*, vol. 11, no. 1, p. 2660, Jan. 2021.

WALEED NAZIH received the B.Sc. and M.Sc. degrees from the Faculty of Computers and Information, Cairo University, Egypt, and the Ph.D. degree from the Faculty of Computers and Information Sciences, Ain Shams University, Egypt. He is currently a Lecturer of computer science with the Department of Computer Science, Prince Sattam Bin Abdulaziz University, Saudi Arabia. His research interests include machine learning, healthcare, natural language processing, Arabic language resources, and voice over IP security.



AHMAD O. ASEERI received the bachelor's degree in computing from King Saud University, Saudi Arabia, the M.Sc. degree in computer science from the University of Wisconsin-Milwaukee, USA, and the Ph.D. degree in computer science from Texas Tech University, USA. He is currently an Assistant Professor with the Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Saudi Arabia. His research interests include artificial intelligence, having the primary focus in machine learning-based optimizations and vulnerability analysis for resource-constraint IoTs and physical unclonable functions. He has also researched works in optimization and modeling methods for healthcare applications, time series forecasting, and data mining with direct application to clustering techniques, including K-means and bisecting memory-aware K-means for big data.

OSAMA YOUSSEF ATALLAH received the B.Eng. and Ph.D. degrees from the Faculty of Computing Sciences and Engineering, De Montfort University, Leicester, U.K. He is currently a Senior Lecturer/a Researcher with the Department of Biomedical Engineering, Medical Research Institute, Alexandria University, Alexandria, Egypt. His research interests include applied machine learning/deep learning, natural language processing, and applied machine learning/deep learning for medical image diagnosis.

SHAKER EL-SAPPAGH received the bachelor's and master's degrees in computer science from the Information Systems Department, Faculty of Computers and Information, Cairo University, Egypt, in 1997 and 2007, respectively, and the Ph.D. degree in computer science from the Information Systems Department, Faculty of Computers and Information, Mansura University, Mansura, Egypt, in 2015. In 2003, he joined the Department of Information Systems, Faculty of Computers and Information, Minia University, Egypt, as a Teaching Assistant. Since June 2016, he has been with the Department of Information Systems, Faculty of Computers and Information, Benha University, as an Assistant Professor. He was a Research Professor with the Department of Information and Communication Engineering, UWB Wireless Communications Research Center, Inha University, South Korea, for three years (2018–2020); and Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain, for one year (2021). He has been an Associate Professor with Galala University, Egypt, since 2021. He has also been a Senior Researcher with the College of Computing and Informatics, Sungkyunkwan University, South Korea, since 2021. He has publications in clinical decision support systems and semantic intelligence. His current research interests include machine learning, medical informatics, (fuzzy) ontology engineering, distributed and hybrid clinical decision support systems, semantic data modeling, fuzzy expert systems, and cloud computing. He is a reviewer in many journals, and he is very interested in the diseases' diagnosis and treatment research.