IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 15, 2021, accepted March 28, 2021, date of publication April 2, 2021, date of current version April 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3070685

Multi-Scale Attention Network for Diabetic Retinopathy Classification

MOHAMMAD T. AL-ANTARY AND YASMINE ARAFA
School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, U.K.
Corresponding author: Mohammad T. Al-Antary (m.alantary@gre.ac.uk)

ABSTRACT Diabetic Retinopathy (DR) is a highly prevalent complication of diabetes mellitus, which causes lesions on the retina that affect vision which may lead to blindness if it is not detected and diagnosed early. Convolutional neural networks (CNN) are becoming the state-of-the-art approach for automatic detection of DR by using fundus images. The high-level features extracted by CNN are mostly utilised for the detection and classification of lesions on the retina. This high-level representation is capable of classifying different DR classes; however, more effective features for detecting the damages are needed. This paper proposes the multi-scale attention network (MSA-Net) for DR classification. The proposed approach applies the encoder network to embed the retina image in a high-level representational space, where the combination of mid and high-level features is used to enrich the representation. Then a multi-scale feature pyramid is included to describe the retinal structure in a different locality. Furthermore, to enhance the discriminative power of the feature representation a multi-scale attention mechanism is used on top of the high-level representation. The model is trained in a standard way using the cross-entropy loss to classify the DR severity level. In parallel as an auxiliary task, the model is trained using the weakly annotated data to detect healthy and non-healthy retina images. This surrogate task helps the model to enrich its discriminative power for distinguishing the non-healthy retina images. The proposed method when implemented has achieved outstanding results on two public datasets: EyePACS and APTOS.

INDEX TERMS Diabetic retinopathy, image classification, artificial intelligence, multi-scale, convolutional neural network (CNN), attention mechanism.

I. INTRODUCTION

In the healthcare field, early diagnosis of many diseases results in more effective treatment. Recently, diabetes and its impediments are becoming a widespread disease all over the world. In the body of a diabetic person, impairment of insulin secretion and resistance to the action of insulin causes an increasing amount of glucose in the blood [1]. This disease affects different parts of the human body such as the heart, nerves, kidneys, and retinas [1]. The retina is the innermost layer of the eye which lines the posterior part of the eye excluding the area of optic nerve. The function of the retina is to process visual information by transferring the light through neural signals and coordinating with the brain [2]. The retina, like all parts of the human body, receives blood nourishment through the micro blood vessels. It is necessary to retain the level of blood sugar with the uninterrupted blood flow [3]. The tiny blood vessels can be damaged by the high blood sugar level, even in the prediabetes stage. A complication of diabetes may cause the blood vessels of the retina to swell and leak fluids and blood, which is called Diabetic Retinopathy (DR) [4], [5]. DR is a diabetic eye disease that is the most prevalent microvascular complication among patients with diabetes mellitus. DR can be classified into two groups of non-proliferative (NPDR) and proliferative (PDR). NPDR is characterised by lesions such as microaneurysms (MAs) and exudates, while neovascularization of weak blood vessels is the hallmark of PDR. DR may lead to loss of vision and it is known as one of the most common aetiologies of permanent blindness of humans [6], [7]. As per statistics cited in 2019 [8], there were about 463 million diabetic patients in the world and almost 30% of them are suffering from DR. The blindness rate of DR is increasing each year, which makes it one of the leading causes for blindness in near future. With early diagnosis and treatment, it is possible to prevent further progression and save many people from permanent blindness.

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

DR can be detected by examining the appearance of some types of lesions of the retina, i.e., microaneurysms (MA), haemorrhages (HM), soft and hard exudates (EX).

DR is divided into five different grades: no DR (Class 0), mild DR (Class 1), moderate DR (Class 2), severe DR (Class 3), and proliferative DR (Class 4). Figure 1 shows the sample retina images for each class.

The early stage of DR is mild DR in which the formation of MA can be detected. In the moderate stage, swelling of blood vessel can be observed which results in blurred vision. During the severe stage, the abnormal growth of blood vessels can be found and a large number of blood vessels are blocked. The advanced stage of DR is proliferative. In this stage, retinal detachment along with a large retinal break can be detected. This stage results in complete vision loss. It is essential to screen the retina of a diabetic patient regularly since there are no early warning symptoms for DR. Nevertheless, the manual grading of DR needs a laborious and is a time-consuming task, i.e., fundus examination after mydriasis of the patient, and it is quite expensive.

By employing some automatic DR identification, many people can benefit from early diagnosis of this disease. In the past years, Computer-aided diagnosis (CAD) systems have been widely studied in healthcare applications. To reduce the cost of regular screening, the technology of capturing colour fundus images is exploited. Many machine learning-based methods have been proposing to address issues in DR Classification. These approaches help the expert to distinguish patients who require a further referral from those who are classified as low-risk by an automatic screening of retina images. Early work in this field of research was based on handcrafted features. In those approaches, retina features such as vessel enhancement, optic disk detection, and lesions segmentation have been extracted from the input images and a classifier (e.g., Support Vector Machine (SVM)) was utilised to categorise the images [9], [10]. Recently, deep learning-based neural networks, particularly convolutional neural networks (CNN) have achieved great success in all areas of medical image processing [11]–[13]. These networks are able to detect complex patterns by extracting powerful features. These features have been extracted by utilising many filters which exploit the natural structure of the data. Among all deep networks, CNN based ones have been most successful for DR grade classification [14], [15], [17], [19]. Noushin *et al.* [14] proposed a two-step CNN for segmenting MAs in the input retinal scans. Li *et al.* [15] proposed a method based on Deep CNN (DCNN) for the identification of DR, in which fractional max-pooling was employed to derive more discriminative features. The features were then classified by SVM. Tan *et al.* [16] employed CNN to detect exudates, MAs, and HEMs. For detecting MAs and filters false positives, Hatanaka *et al.* [17] utilised a two-step DCNN. Gargeya and Leng [18] exploited CNN for extracting features from images, which were then fed to a tree-based model that classifies binary DR. Gulshan *et al.* [20] used inception-v3 to detect DR grades. A main challenge on the fundus images is that most parts of the retina images are irrelevant to the DR while some parts of the input image have more influences on the final label of an image. Most of the CNN approaches employed for DR classification, process the input data without considering this fact. To address this problem, this paper proposes using a Multiscale Spatial Attention network (MSA-Net) for DR classification in which a multi-scale attention mechanism is inserted on top of the high-level representation. The multi-scale helps the network to learn where to look for retina damages and scale the representation space. To produce multi-scale representation, the idea of Atrous convolution has been employed. Then the attention maps produced by the attention mechanism have been utilised to focus on more informative parts of the multi-scale feature representation. The model has been trained using both supervised data and weakly annotated data to boost the classification performance. The proposed method has been evaluated on the EyePACS and APTOS public datasets. The experimental results demonstrate the effectiveness of the proposed method compared to other methods. The proposed method has the following contributions:

1- Multi-level feature encoding structure is considered to include both local and semantic features.
2- A deep neural network architecture with multi scale attention mechanism on top of the high-level feature representation is proposed to improve and scale the discriminative power of the representation space.
3- Boosting performance with multi-task learning.
4- State-of-the-art performance on two public datasets.

The remainder of this paper is structured as follows. An overview of the related work is discussed in Section II. The proposed method is presented in Section III. Section IV shows the experimental results and evaluates the proposed method with respect to different metrics. Finally, the conclusions are presented in Section V.

II. RELATED WORK

Manual detection of DR images had many problems. Lack of expertise (professional ophthalmologists) and expensive experiments cause many problems for patients in undeveloped countries. Therefore, automated processing techniques have been developed to simplify access to accurate and rapid diagnosis and treat the patients at early stages to prevent blindness. In recent years, machine learning models that focused on analysing eye fundus images, have managed to attain accurate automatic DR classification. Many efforts have been made to establish faster and cheaper automatic approaches [18]. Therefore, these approaches have become more efficient than manual ones for all human beings. The research background of DR classification can be divided into two groups of old handcrafted approaches and modern deep learning-based ones. These methods are discussed in more detail as follows.

A. HANDCRAFTED APPROACHES

Previous automatic DR frameworks strongly depend on the variables manually measured, i.e., handcrafted features. Akram *et al.* [21] proposed a hybrid structure of Gaussian Mixture Model (GMM) and Support Vector Machines (SVM) for DR classification. The authors then improved that method by augmenting the feature set with shape, intensity, and statistics of the affected region [22]. Adal *et al.* [23] utilised several intensity and shape features to capture the changes in red lesions and then classify the DR grade by employing SVM. K-nearest Neighbour (KNN) algorithm classifies test data based on their distance in the feature space to the K training samples. Tang *et al.* [24] utilised KNN for classifying the haemorrhages candidate from DR. To do that, the author first segmented the retinal images into splat partitions, and by applying some filters, they then selected some optimal features. These features were used as the input to the classifier. KNN has worked well in many cases, however, this algorithm has some computational efficiency and generalisability problems. Random forest is another classifier that has been used in many DR classification approaches. It includes an ensemble of trees that are trained based on random subsamples of the training set. In each node of a tree, an optimum feature is selected based on its entropy to classify the remaining training samples. Zhang *et al.* [25] exploited random forest as a binary classification to separate lesions from non-lesions in retina images. The random forest has been successful in many classification approaches, nevertheless, when there is not a clear class distinction in data (e.g. mild and moderate) the algorithm may fail. A genetic algorithm-based feature extraction method [26] and AdaBoost [27] have been also employed in this field of research. The main problem with all the handcrafted features that they need a heuristic feature extraction stage. The manual feature extraction may not work well in more complex tasks. Therefore, with the emergence of deep learning-based approaches, researchers moved to utilise deep feature-based models for DR classification.

B. DEEP LEARNING APPROACHES

Deep neural networks are a new branch of machine learning tools. These networks consist of a set of consecutive convolutional layers to learn both features and classifiers together. In medical image analysis, the most popular and effective deep learning methods are convolutional neural networks (CNN). The performance of these networks significantly depends on the size of the training data. CNN includes three kinds of layers, i.e., convolutional, pooling, and fully connected layers. Each convolutional layer contains several filters that are convolved with the input data to extract features from the input data. Pooling layers are used to reduce the input data size. The fully connected layers are also employed to produce compact feature sets. Quellec *et al.* [28] utilised a heat map optimization scheme to introduce a system for identifying DR by employing a deep convolutional neural network (CNN) and automatically detect lesions in retinal images. Zeng *et al.* [29] trained inception-V3 with metric learning techniques to classify the colour retinal fundus photographs into two grades. In [18] a residual CNN was used for the assessment of glaucoma. For classifying DR, after a pre-processing step, the fundus images are fed to the network to learn discriminative features, similar to [30]. Gulshan *et al.* [20] introduced a fundus image dataset with about 128 thousand images and utilised inception-V3 to detect DR and diabetic macular edema. Bodapati *et al.* [31] proposed a multi-modal fusion module by combining multiple pre-trained CNNs a multi-modal fusion module to extract discriminative features. The authors employed the VGG16 model to learn the lesions, and Xception to learn the global representation of the images. Kassani *et al.* [32] introduced a modified version of the Xception model for classifying DR. They insert a deep layer aggregation that receives multilevel features from different convolutional layers in the Xception model. The extracted features were then classified by a multi-layer perceptron (MLP). Jain *et al.* [33] evaluated the performance of three networks: VGG16, VGG19, and InceptionV3 for both binary and 5-class DR classification. The authors employed different data augmentation strategies to balance the input data. Their results demonstrate that the best performance was achieved by the network with a larger number of convolutional and pooling layers, i.e., VGG19. Hagos *et al.* [34] exploited the idea of transfer learning with the InceptionV3 model for small dataset. The authors used the cosine loss function with an SGD optimiser. A Siamese-like structure was employed for DR classification by receiving binocular (two fundus images corresponding to the left eye and right eye) fundus images as inputs [29].

Pratt *et al.* [35] utilised different data augmentation approaches to enrich the size of input data to about 80,000 samples for training a CNN model. A microaneurysm detection system was introduced by Haloi *et al.* [36] for DR detection. The authors used a deep neural network (DNN) to identify microaneurysm from the original input without any pre-processing steps. The OCTD-Net [37] was proposed for DR detection in its early stages. OCTD-Net consists of two networks. The first network extracts features from the original optical tomography (OCT) images and the second one is used for extracting retinal layer information. Poplin *et al.* [38] proposed a DNN model for predicting cardiovascular risk factors from the retinal fundus photograph. Mateen *et al.* [39] first extracted features from different layers of a pre-trained VGG19 model, and then in order to avoid overfitting, the authors applied Principal Component analysis (PCA) and singular value decomposition (SVD) to reduce the feature dimension. Vo *et al.* [40] introduced a DR detection model by combing kernels with multiple losses network (CKML Net) and VGG Net with Extra Kernel (VNXK). Dai *et al.* [41] integrated a multi sieving CNN framework with an image-to-text mapping scheme to detect microaneurysms from fundus images. The authors used the image-to-text mechanism as a clinical report. A challenge in retina images is that these images contain more irrelevant

information rather than relevant ones (microaneurysms which are critical for ophthalmologists). Wang *et al.* [42] proposed Zoom-in-Net to classify DR. The Zoom-in-Net generates suspicious areas by employing an attention mechanism. Another bilinear learning strategy [43] with an attention mechanism was also utilised for DR classification. The authors employed an attention approach to boost the meaningful features while suppress the weak ones.

III. PROPOSED METHOD

In this section, the proposed method is presented in more details. The general diagram of the proposed architecture is depicted in figure 2. In the proposed architecture, first, the pre-processing approach applied to the input image to normalise the retina image and reduce the illumination and intra-class variation effect. The pre-processed images were then fed to the MSA-Net to estimate the DR severity level. The MSA-Net consists of four blocks: the encoder block which encodes the input retina image into a high-level representation space, multi-level and multi-scale feature representation block, a Multi-scale attention mechanism, and finally the decoder block to generate the DR score. In the following subsections, each block is discussed in more details.

A. PRE-PROCESSING

The retina fundus images are usually collected from different clinics, captured by different devices. Therefore, they have considerable intensity variation. Like [32], in order to optimise the training process, a pre-processing step was performed in this approach. The input images were first resized to 512×512 based on their aspect ratio by using bicubic interpolation. These images were then cropped from the centre to 320×320 pixels such that each retinal circle is located at the centre of the image. The approach introduced by Graham [44] was employed to improve the clarity of blood vessels and lesion areas. To do that, the black pixels of the input images were first removed. Next, a min-pooling filtering approach was employed to normalise the images as [44] (1):

$$Ic = \alpha I + \beta G(\rho) * I + \gamma \tag{1}$$

where $*$ is the convolution operation, I denotes the input image, and G (ρ) denotes the Gaussian filter with a standard deviation of ρ. Pre-defined parameters are also used: α; β; γ. to remove feature bias and achieve uniform distribution across the dataset, the intensity values of cross channels of all images of the dataset have been normalised to $[-1, 1]$. Figure 3 shows the sample of pre-processing results, which results in the normalised retina images.

B. RESNET AS ENCODER

The first part of the network is the encoder part. As it can be seen in Figure 2 the ResNet architecture was used as the encoder in the proposed network. Hypothetically, by increasing the number of layers in a deep network, the performance of the network should be improved. However, there is a problem in the deep networks, called vanishing gradients. The

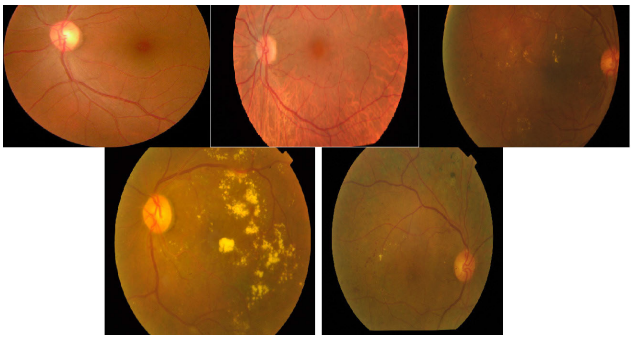

FIGURE 1. Sample of retinopathy images for different grades. In order from the top left to the bottom right, grade 0-4.

ResNet has been proposed to solve this problem by using short connections, i.e., direct connections between the output of each layer with the input of the adjacent layer. The network learns the residuals. Since this network is relatively easy to optimise, the accuracy can be increased by adding more layers. The first layer of the network is a 7*7 convolutional layer. The network then consists of 4 stages. The first to the fourth stages include 3, 8, 12, and 3 residual blocks, respectively. At the end, the network has an average pooling layer. It is worth mentioning that the ResNet structure was used without any fully connected layers as the encoder of MSA-Net. Mathematically the encoder model G with parameters θ_1 in the proposed network takes the retina image I and generates the representation tensor F_{enc} (equation 2):

$$F_{enc} = G(\theta_1; I) \tag{2}$$

C. MULTI-LEVEL & MULTI-SCALE REPRESENTATION

The encoder part extracts features through a consecutive residual blocks. The extracted features close to the input images have higher resolution and therefore they have more information about the local features while the features close to the last layer include more semantic information. To include both kinds of information in the next steps, multi-level features were combined, i.e., mid-level and high-level features. Since these features have different spatial resolution, a scaling mechanism was employed to make them evenly sized (semantic learning in figure 2). A concatenation of these feature sets was then fed to an Atrous convolution to extract information with different scales. The Atrous convolution applies convolutional filters with different field of view sizes. Using a small field of view for extracting helped the network to encode more local information. On the other hand, more global information was taken into account with larger image context [11]. The produced multi-level and multi-scale representation encodes the input image in a compact representation space, which is capable to learn diabetic signs with different scales, locality and severities.

D. MULTI-SCALE ATTENTION MECHANISM

Depending on the DR severity level, the structure of the retina can be deformed. This deformation can cause special damage

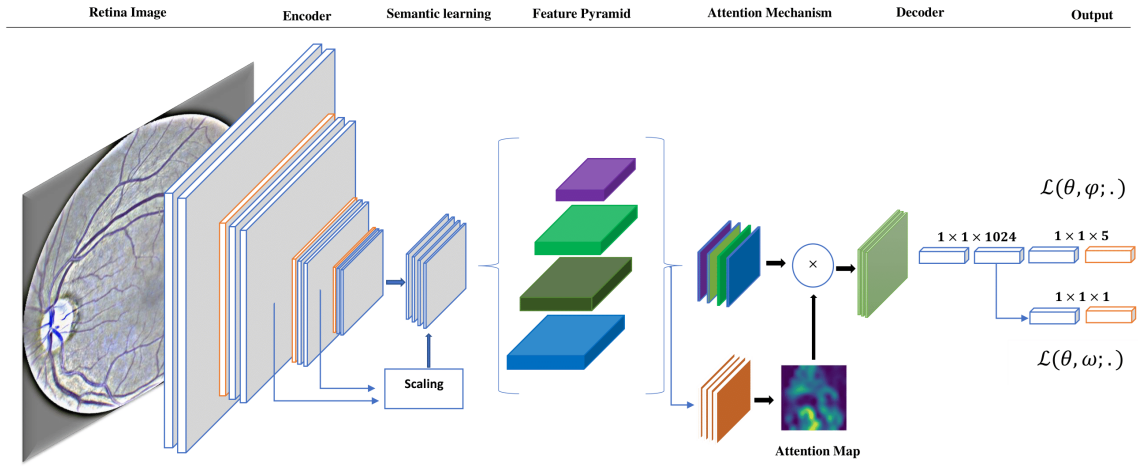

FIGURE 2. Structure of the proposed MSA-Net for DR classification. The MSA-Net utilised the multi-level and multi-scale representation to encode high-level representation. Further, it applied a multi-scale attention mechanism to scale the discriminative power of the representation space.

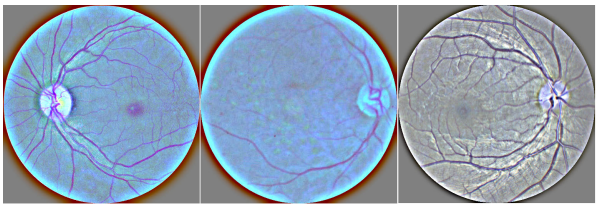

FIGURE 3. The pre-processed retina images resulted from the pre-processing algorithm.

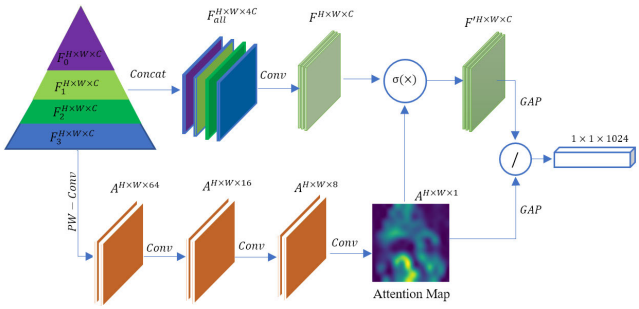

FIGURE 4. Multi-scale attention mechanism proposed in MSA-Net.

to the retina fundus. In order to classify such damage, the literature work used the high-level representation achieved by the deep CNN model. Although, the high-level representation is capable of distinguishing different classes, its effectiveness for precisely detecting the damage is limited due to the scarcity of diabetic patterns. To increase the discriminative power of such representation, the proposed method included the attention mechanism on top of the multi-scale representation. The objective of the attention mechanism is to learn where to look for retina damages and scale the representation space. In other words, the attention mechanism in the MSA-Net model highlights the diseased parts in the retinal image with less emphasis on the normal regions. Figure 4 illustrates the proposed attention mechanism.

In the proposed MSA-Net structure, the attention mechanism is a series of convolution layers applied to the multi-scale representation. More precisely, at first, the point-wise convolution applied between the pyramid representations to form a compact representation:

$$A^{H\times W\times 64} = \left(F_{all}^{H\times W\times 4C} * K_{pw}\right) \tag{3}$$

where A shows the attention tensor, F represents the representation tensor generated by the multi-scale block and K_{pw} shows the point-wise convolution kernel. The generated compact representation was then fed to the series of convolution to generate the attention map $A^{H\times W\times 1}$. The obtained attention map was then multiplied with the high-level representation $F^{H\times W\times C}$ to scale the representation. The sigmoid σ activation function was included to normalise the output in range [0, 1]. Finally, the global representation was obtained using the global average pooling (GAP) of the final representation. The final representation was normalised using the GAP information of the attention map. Thus, the model generated the final representation vector F as:

$$F^{1\times 1024} = \frac{GAP(\sigma(F^{H\times W\times C} \odot A^{H\times W\times 1}))}{GAP(A^{H\times W\times 1})} \tag{4}$$

where $\odot$ shows the point wise multiplication. It is worthwhile to mention that the learnable parameters for the multi-scale and attention blocks were determined as θ_2. The attention mechanism implemented in the MSA-Net architecture improves the model learning and contributes towards boosting the accuracy of retinal images classification based on DR severity level, since it considers the outputs of the previous layers with varying importance. This distinguishes the proposed model from standard neural network models that do not consider variation.

E. DECODING AND TRAINING

The decoder block in the proposed method consists of fully connection layers for mapping the features vectors to the

desired outputs. In this paper, two objectives are defined as follows:

DR classification: The main objective of the proposed network is to classify the retinopathy images; thus, the loss function is defined $\mathcal{L}(\theta, \varphi.)$ as a classification loss for model with encoder + attention parameters $\theta = \theta_1 \cup \theta_2$ and the classification branch parameters φ. The cross-entropy loss is utilised between the predicted class and the true class. In addition, in the loss function $\mathcal{L}(\theta, \varphi; .)$, the non-trainable weight is included to scale the importance of each class loss on the final loss values. The objective of this weighted loss is to control the effect of unbalanced samples on the training process.

Healthy and non-healthy retina: as an auxiliary task, the model was trained to classify the retina fundus images as either healthy or non-healthy. Since the ultimate objective of automatic retinopathy detection is to assist the ophthalmologist/hospitals and reduce the monitoring burden, this objective is designed to help the doctor in recognising retinopathy. Besides that, annotating the healthy and non-healthy without precise DR score is much easier than precise scoring for the doctors. Thus, this weak annotation can be provided in easier fashion. Given the fact that such a dataset can be available, it is aimed to include this dataset in the training process to increase the main objective's performance. The auxiliary task was trained with parameters $\mathcal{L}(\theta, \omega; .)$ using the cross-entropy loss function.

The network was trained for 100 epochs with Adam optimisation with batch size 4 and learning rate 10^{-4}. Also, the geometrical data augmentation techniques like flipping, rotation and scaling were included in the training process to avoid overfitting.

IV. EXPERIMENTAL RESULTS

The proposed network was evaluated on two public datasets, namely EyePACS, APTOS. This section discusses further the details of the datasets, performance measuring metrics, experiment setups, and visual results.

A. DATASET

Two public datasets EyePACS and APTOS were utilised in the experimental results. In the following subsection more details about these datasets and the used setting are provided.

1) EYEPACS KAGGLE

EyePACS published Kaggle dataset [45] for a competition and to facilitate the researchers without any cost. This dataset is a well-known dataset that has been widely used for the detection of DR. The data is a challenging one since the images vary by camera, eye polarity (left/right), inversion or view, and noise-like artifacts and exposure issues. It includes a total number of 88,702 retinal fundus photographs. The train set contains 33,566 fundus photographs with 5 classes, where 0 represents no disease and 4 represents the highest severity level of the disease. Kaggle EyePACS test set has more than 50,000 retinal images for which there are no ground truth labels. In experiment, similar to [29], the annotated dataset was divided into a training and test set, where the test set contains 10% of each class. Figure 5 shows sample of retina images from this dataset which shares a high variation inside the dataset.

FIGURE 5. A sample of retinopathy images from EyePACS data set [36].

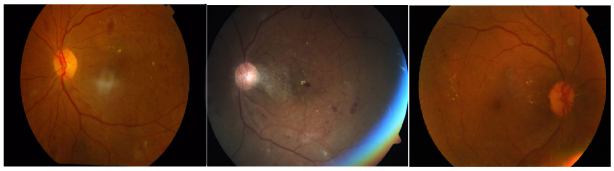

FIGURE 6. A sample of retinopathy images from APTOS data set [37].

2) APTOS KAGGLE

The APTOS dataset [46] is a large dataset of retinal images which have been taken using fundus photography under different imaging conditions. The dataset consists a total of 3662 retina images collected from multiple clinics from Aravind Eye Hospital in India. The fundus images included in this dataset are categorized into five classes: No DR (Class 0), Mild DR (Class 1), Moderate DR (Class 2), Severe DR (Class 3), Proliferative DR (Class 4). Figure 6 shows sample of images from the APTOS dataset. In this dataset, the class distribution is highly imbalanced, i.e., 49%, 10%, 27%, 5%, and 8% of images belong to normal, mild, moderate, severe, proliferative DR, respectively. The same setting was followed as in [32], which used 10% of the labelled samples as a test set and the rest for the training purpose.

B. METRICS

The performance of the proposed method is compared to other approaches using the following metrics: accuracy, sensitivity, specificity, area under ROC curve, F1 score, and Kappa score. These metrics are mathematically calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{8}$$

$$Kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \quad (9)$$

where true positive (TP) shows the number of samples that are correctly classified as the positive class, true negative (TN) indicates the number of samples that are correctly classified as the negative class, false positive (FP) is the number of samples which has the original negative label while they are classified as positive class, and false negative (FN) shows the number of instances belonging to the positive class, but are predicted as negative class. In the Kappa equation O, $EandW$ are the $N \times N$ matrices where the O is the confusion matrix, E shows the expected rating and W is the weight matrix calculated based on the difference between the ground truth and the predicted value. The ROC curve is plotted with true positive rate (as x-axis) against the false positive rate (as y-axis). AUC is calculated as the area under this curve.

C. EVALUATION

The proposed deep model was evaluated on both EyePACS and APTOS datasets for DR classification. The model was trained to classify the DR score in form of multi-class classification. Accuracy, sensitivity, specificity, Kappa and F1 metrics were utilised to compare the proposed method with the literature work. The performance of the proposed method on healthy and non-healthy retina classification is also reported.

1) APTOS KAGGLE RESULTS

Following the literature work on DR classification on APTOS dataset, the labelled samples were divided into a training and test set using the same setting mentioned in [32]. In order to provide a comprehensive comparison results, the well-known deep structures such as Inception v3, MobileNet, VGG and Resnet models were also trained with transfer learning. To this end, in each model we modified the last fully connection layer to the five nodes and the models were trained for 100 epochs on the same setting. The baseline models' results are similar to the results mentioned in [32]. Table 1 shows the comparison results for the APTOS dataset.

According to table 1, the proposed method outperformed literature work under the same setting. For the best baseline model (VGG) the accuracy metric was improved by 4.59% and comparing to the recent method modified Xception method [32] 1.51% accuracy improvement was achieved. In addition, the proposed method outperformed both blended [31] and hybrid [47] methods where these methods used a combination of different deep models while the proposed model used a single model. The proposed method was also evaluated using Kappa score where it achieved Kappa score 0.896. This fact reveals the robustness of the proposed method in precise DR classification. In figure 7, the normalised confusion matrix is provided to further analyse the model accuracy on classifying each particular DR level. The confusion matrix was obtained by applying the proposed method on the test set.

TABLE 1. Performance comparison between the proposed method and Previous WORK for DR classification on Aptos.

Method	Accuracy	Sensitivity	Specificity
Inception v3 [48]	0.787	0.636	0.853
MobileNet [49]	0.790	0.764	0.846
VGG [50]	0.800	0.853	0.866
ResNet50 [51]	0.746	0.565	0.857
Blended [31]	0.809	-	-
Hybrid [47]	82.18	-	-
Modified Xception [32]	0.830	0.882	0.870
Proposed Method	**0.846**	**0.910**	**0.905**

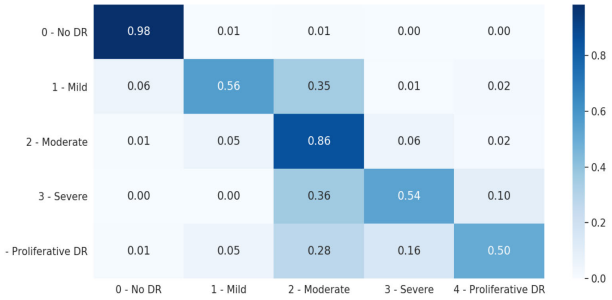

FIGURE 7. Normalised confusion matrix of the proposed method on APTOS dataset.

According to the confusion matrix, the proposed method is able to recognise healthy retina (no diabetic retinopathy) with 98% accuracy, which is highly desirable in real-world application. Furthermore, the confusion matrix reveals that the MSA-Net misclassified the samples between moderate class and other diabetic classes. In fact, this happened because of the high correlation between diabetic classes. In table 2, the experimental results in term of healthy and non-healthy retina classification are provided. To do so, the proposed method and the baseline models were applied on the APTOS test set to classify the retina either as healthy or non-healthy.

It is clear from table 2 that the proposed method outperformed the baseline methods in all F1, accuracy, sensitivity and specificity scores. To further analyse the true positive rate against the false positive rate, the ROC curve is provided in figure 8. The ROC curve in figure 8 demonstrates the trade-off between the sensitivity and the specificity of the model. In other words, with increasing the number of FPR the model is able to recognise non-healthy retina with high probability. This matter is highly useful when our objective is to drop patients with healthy retinas while keeping all the patients with non-healthy retinas for further analyses by the ophthalmologist.

TABLE 2. Performance of the proposed method for healthy and non-healthy retina classification on Aptos dataset.

Method	F1	Accuracy	Sensitivity	Specificity
Inception v3 [48]	0.968	0.965	0.972	0.969
MobileNet [49]	0.969	0.963	0.974	0.969
ResNet50 [51]	0.965	0.958	0.968	0.962
VGG [50]	0.971	0.969	0.974	0.971
Proposed Method	**0.982**	**0.981**	**0.983**	**0.982**

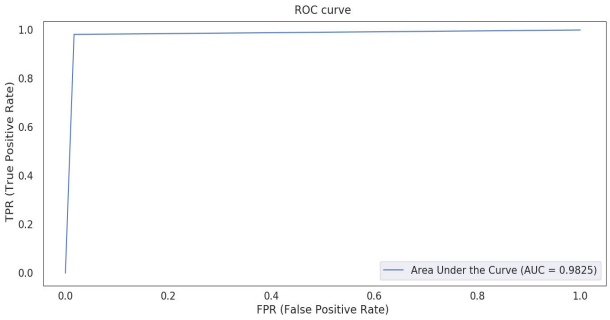

FIGURE 8. ROC curve of the proposed method on APTOS dataset.

TABLE 3. Performance comparison between the proposed method and Previous work for DR classification on eyepacs.

Method	Kappa score
Inception v3 [48]	0.792
MobileNet [49]	0.770
ResNet50 [51]	0.786
Monocular Model [29]	0.808
Binocular Model [29]	0.829
Proposed method	**0.878**

2) EYEPACS KAGGLE RESULTS

To further analyse the effectiveness of the proposed MSA-Net, the model was also evaluated on EyePACS dataset. The evaluation setting followed the [29] and divided the labelled samples into a training and test sets where the test set contains 10% of the labelled samples. Table 3 provides the comparison results between the proposed method and the previous work described in Section II. The Kappa score was used for the comparison.

As it is clear from table 3, the proposed MSA-Net outperformed both baseline models, Monocular and Binocular models. The performance improvement (5.5) clarifies the effectiveness of the multi-scale attention mechanism on retinopathy recognition. To further analyse the discriminative power of the MSA-Net for distinguishing different diabetic

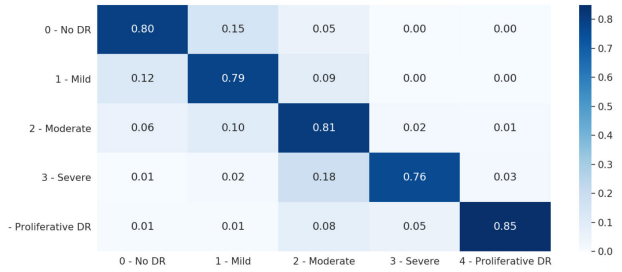

FIGURE 9. Normalised confusion matrix of the proposed MSA-Net on EyePACS dataset.

TABLE 4. Performance of the proposed method for healthy and non-healthy retina classification on Eyepacs dataset.

Method	F1	Accuracy	Sensitivity	Specificity
Inception v3 [48]	0.746	0.840	0.853	0.767
MobileNet [49]	0.738	0.820	0.752	0.863
ResNet50 [51]	0.741	0.836	0.893	0.727
Proposed Method	**0.767**	**0.875**	**0.906**	**0.787**

classes, the normalised confusion matrix is provided in figure 9. According to the confusion matrix the model has almost similar classification error for each class. This fact resulted from the heavy training loss weight that was allocated for the DR classes than healthy class. The weighted loss may have reduced the performance on healthy retina classification, but it improved the overall classification performance.

The proposed method was also evaluated for healthy and non-healthy retina classification on EyePACS dataset. To do so, N samples from each class was randomly selected to form a training set. The test set also contains balanced samples for each class (M each class). In which, 1000 and 200 for N and M were chosen, respectively. Repetition was used if necessary. The main purpose behind this setting was to analyse the effect of class balance on healthy and non-healthy classification. Table 4 shows the comparison results.

Among all methods, the proposed MSA-Net provided more robust classification results for healthy and non-healthy retina classification. Furthermore, the ROC curve is provided in the figure 10 to demonstrate the FPR ratio to the TPR.

D. DISCUSSION

In this section, the effect of multi-level multi-scale (MM) representation on the MSA-Net performance is discussed. The generated attention maps were also visualised to discuss its effectiveness in retinopathy recognition. In order to analyse the effect of multi-level multi-scale representation, the proposed model was trained with and without multi-level multi-scale representation. Table 5 demonstrates the experimental results. In the experimental results, it is concluded that using MM structure in both datasets improved the performance.

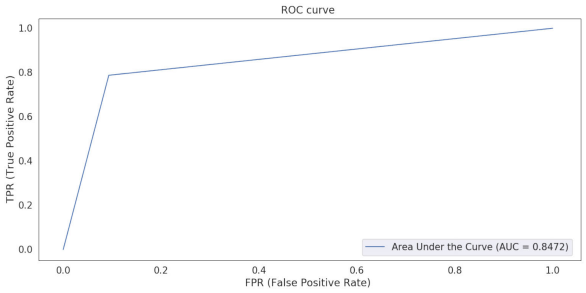

FIGURE 10. ROC curve of the proposed method for EyePACS dataset.

TABLE 5. Performance measurement based on different model setting.

Dataset	Setting	Accuracy	Kappa score
APTOS	MSA-Net without MM	0.821	0.857
	MSA-Net with MM	0.846	0.896
EyePACS	MSA-Net without MM	0.775	0.842
	MSA-Net with MM	0.799	0.878

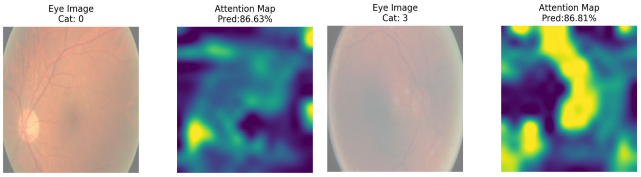

FIGURE 11. Sample of attention maps extracted by MSA-Net for healthy and moderate diabetic type retina images.

This fact showed that using multi-scale receptive field can generate feature representation with different locality. Combining all these local and global representations helped the model to learn the DR structure with different scales. Moreover, combining mid and high-level representations improved feature reusability and consequently resulted in better performance.

The attention mechanism in the proposed method aims to focus more on the informative area and scale the representation space. Thus, it is highly effective in recognising the DR severity level. Figure 11 demonstrates two attention maps extracted for healthy and Moderate DR. The extracted attention maps demonstrate the importance of the input area for recognising the DR. As in our experiment, removing the multi-scale attention mechanism from the proposed MSA-Net reduced the kappa score of the proposed method by 2.5% on APTOS dataset.

E. COMPUTATIONAL TIME

Ophthalmologists take up to five minutes to closely analyse and grade a retinal image based on the severity level of diabetic retinopathy [52]. Moreover, in special cases such as the presence of macular degeneration, the task of manual grading and analysis of a retinal image may take longer time [52]. To automate this process and accelerate the diagnosis time, machine learning-based approaches are proposed. In this section, the complexity and computational time of the proposed method are reported. To this end, the number of operations (multiplication and sum) required to execute in the Graphics Processing Unit (GPU) device is calculated. Furthermore, the number of model parameters is reported in table 6 to demonstrate the complexity of the model. Table 6 shows the training time per one epoch on APTOS dataset, number of operations as well as the number of model parameters.

TABLE 6. Complexity of the proposed method.

Training time one epoch (APTOS)	Million Operation	Million Parameters
4 minutes	31.5	15.7

As demonstrated in table 6, the proposed method has 31.5 million operations which can be executed in a single GPU device (GTX 1080). In our experiments, the classification results for a batch of 32 images took only 5 seconds, which is comparably more efficient than manual grading.

V. CONCLUSION

In this paper, a novel deep learning model (MSA-Net) is proposed for the classification of the damage caused by DR on retina images. To improve the representation power of the network, the multi-scale attention mechanism on top of the high-level feature representation has been introduced. The multi-scale mechanism consists of the Atrous convolution which processed the input feature with different scales. The attention maps were produced with a series of convolutional layers. The attention maps were employed to focus on more informative parts of the multi-scale representation and suppress the weak ones. Furthermore, the multi-level and multi-scale representation layers were included in the network to boost the performance. Training model in form of multi-task learning achieved better performance than previous work described in the literature. The experimental results demonstrate the effectiveness and efficiency of the proposed model in diagnosing and classifying the DR disease. Accordingly, the proposed method has great clinical application potential in the future.

REFERENCES

[1] D. Mellitus, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 28, no. 37, pp. S5–S10, 2005.
[2] H. Wassle and B. B. Boycott, "Functional architecture of the mammalian retina," *Physiolog. Rev.*, vol. 71, no. 2, pp. 447–480, Apr. 1991.
[3] R. W. Rodieck, "The vertebrate retina: Principles of structure and function," Freeman, San Francisco, CA, USA, Tech. Rep., 1973.
[4] H.-P. Hammes, J. Lin, O. Renner, M. Shani, A. Lundqvist, C. Betsholtz, M. Brownlee, and U. Deutsch, "Pericytes and the pathogenesis of diabetic retinopathy," *Diabetes*, vol. 51, no. 10, pp. 3107–3112, Oct. 2002.
[5] S. Roy, J. Ha, K. Trudeau, and E. Beglova, "Vascular basement membrane thickening in diabetic retinopathy," *Current Eye Res.*, vol. 35, no. 12, pp. 1045–1056, 2010.

[6] K. Mazhar, R. Varma, F. Choudhury, R. McKean-Cowdin, C. J. Shtir, and S. P. Azen, "Severity of diabetic retinopathy and health-related quality of life: The Los Angeles Latino eye study," *Ophthalmology*, vol. 118, no. 4, pp. 649–655, 2011.

[7] J. R. Willis, Q. V. Doan, M. Gleeson, Z. Haskova, P. Ramulu, L. Morse, and R. A. Cantrell, "Vision-related functional burden of diabetic retinopathy across severity levels in the united states," *JAMA ophthalmol.*, vol. 135, no. 9, pp. 926–932, 2017.

[8] M. Sharif and J. H. Shah, "Automatic screening of retinal lesions for grading diabetic retinopathy," *Int. Arab J. Inf. Technol.*, vol. 16, no. 4, pp. 766–774, 2019.

[9] A. Pinz, S. Bernogger, P. Datlinger, and A. Kruger, "Mapping the human retina," *IEEE Trans. Med. Imag.*, vol. 17, no. 4, pp. 606–619, Aug. 1998.

[10] M. Al-Antary, M. Hassouna, Y. Arafa, and R. Khalifah, "Automated identification of diabetic retinopathy using pixel-based segmentation approach," in *Proc. 2nd Int. Conf. Watermarking Image Process.*, Sep. 2019, pp. 16–20.

[11] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with densley connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 406–415.

[12] R. Azad, M. Asadi Aghbolaghi, M. Fathy, and S. Escalera, "Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 251–266.

[13] Z. Zeng, Y. Xulei, Y. Qiyun, Y. Meng, and Z. Le, "SeSe-Net: Self-supervised deep learning for segmentation," *Pattern Recognit. Lett.*, vol. 128, pp. 23–29, Dec. 2019.

[14] N. Eftekhari, H.-R. Pourreza, M. Masoudi, K. Ghiasi-Shirazi, and E. Saeedi, "Microaneurysm detection in fundus images using a two-step convolutional neural network," *Biomed. Eng. OnLine*, vol. 18, no. 1, p. 67, Dec. 2019.

[15] Y.-H. Li, N.-N. Yeh, S.-J. Chen, and Y.-C. Chung, "Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network," *Mobile Inf. Syst.*, vol. 2019, pp. 1–14, Jan. 2019.

[16] J. H. Tan, H. Fujita, S. Sivaprasad, S. V. Bhandary, A. K. Rao, K. C. Chua, and U. R. Acharya, "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," *Inf. Sci.*, vol. 420, pp. 66–76, Dec. 2017.

[17] Y. Hatanaka, K. Ogohara, W. Sunayama, M. Miyashita, C. Muramatsu, and H. Fujita, "Automatic microaneurysms detection on retinal images using deep convolution neural network," in *Proc. Int. Workshop Adv. Image Technol. (IWAIT)*, Jan. 2018, pp. 1–2.

[18] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, Jul. 2017.

[19] G. S. Scotland, P. McNamee, A. D. Fleming, K. A. Goatman, S. Philip, G. J. Prescott, P. F. Sharp, G. J. Williams, W. Wykes, G. P. Leese, and J. A. Olson, "Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy," *Brit. J. Ophthalmol.*, vol. 94, no. 6, pp. 712–719, Jun. 2010.

[20] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, and R. Kim, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[21] M. U. Akram, S. Khalid, and S. A. Khan, "Identification and classification of microaneurysms for early detection of diabetic retinopathy," *Pattern Recognit.*, vol. 46, no. 1, pp. 107–116, 2013.

[22] M. U. Akram, S. Khalid, A. Tariq, S. A. Khan, and F. Azam, "Detection and classification of retinal lesions for grading of diabetic retinopathy," *Comput. Biol. Med.*, vol. 45, pp. 161–171, Feb. 2014.

[23] K. M. Adal, P. G. van Etten, J. P. Martinez, K. W. Rouwen, K. A. Vermeer, and L. J. van Vliet, "An automated system for the detection and classification of retinal changes due to red lesions in longitudinal fundus images," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1382–1390, Jun. 2018.

[24] L. Tang, M. Niemeijer, J. M. Reinhardt, M. K. Garvin, and M. D. Abramoff, "Splat feature classification with application to retinal hemorrhage detection in fundus images," *IEEE Trans. Med. Imag.*, vol. 32, no. 2, pp. 364–375, Feb. 2012.

[25] X. Zhang, G. Thibault, E. Decencière, B. Marcotegui, B. Lay, R. Danno, G. Cazuguel, G. Quellec, M. Lamard, P. Massin, A. Chabouis, Z. Victor, and A. Erginay, "Exudate detection in color retinal images for mass screening of diabetic retinopathy," *Med. Image Anal.*, vol. 18, no. 7, pp. 1026–1043, Oct. 2014.

[26] R. A. Welikala, M. M. Fraz, J. Dehmeshki, A. Hoppe, V. Tah, S. Mann, T. H. Williamson, and S. A. Barman, "Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy," *Computerized Med. Imag. Graph.*, vol. 43, pp. 64–77, Jul. 2015.

[27] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic retinopathy analysis using machine learning," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 5, pp. 1717–1728, Sep. 2014.

[28] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Med. Image Anal.*, vol. 39, pp. 178–193, Jul. 2017.

[29] X. Zeng, H. Chen, Y. Luo, and W. Ye, "Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network," *IEEE Access*, vol. 7, pp. 30744–30753, 2019.

[30] P. Zang, L. Gao, T. T. Hormel, J. Wang, Q. You, T. S. Hwang, and Y. Jia, "DcardNet: Diabetic retinopathy classification at multiple levels based on structural and angiographic optical coherence tomography," *IEEE Trans. Biomed. Eng.*, early access, Aug. 28, 2020, doi: 10.1109/TBME.2020.3027231.

[31] J. D. Bodapati, V. Naralasetti, S. N. Shareef, S. Hakak, M. Bilal, P. K. R. Maddikunta, and O. Jo, "Blended multi-modal deep ConvNet features for diabetic retinopathy severity prediction," *Electronics*, vol. 9, no. 6, p. 914, May 2020.

[32] S. H. Kassani, P. H. Kassani, R. Khazaeinezhad, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Diabetic retinopathy classification using a modified xception architecture," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2019, pp. 1–6.

[33] A. Jain, A. Jalui, J. Jasani, Y. Lahoti, and R. Karani, "Deep learning for detection and severity classification of diabetic retinopathy," in *Proc. 1st Int. Conf. Innov. Inf. Commun. Technol. (ICIICT)*, Apr. 2019, pp. 1–6.

[34] M. T. Hagos and S. Kant, "Transfer learning based detection of diabetic retinopathy from small dataset," 2019, *arXiv:1905.07203*. [Online]. Available: http://arxiv.org/abs/1905.07203

[35] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.

[36] M. Haloi, "Improved microaneurysm detection using deep neural networks," 2015, *arXiv:1505.04424*. [Online]. Available: http://arxiv.org/abs/1505.04424

[37] X. Li, L. Shen, M. Shen, F. Tan, and C. S. Qiu, "Deep learning based early stage diabetic retinopathy detection using optical coherence tomography," *Neurocomputing*, vol. 369, pp. 134–144, Dec. 2019.

[38] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomed. Eng.*, vol. 2, no. 3, p. 158, 2018.

[39] M. Mateen, J. Wen, Nasrullah, S. Song, and Z. Huang, "Fundus image classification using VGG-19 architecture with PCA and SVD," *Symmetry*, vol. 11, no. 1, p. 1, Dec. 2018. [Online]. Available: https://www.mdpi.com/2073-8994/11/1/1

[40] H. H. Vo and A. Verma, "New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2016, pp. 209–215.

[41] L. Dai, R. Fang, H. Li, X. Hou, B. Sheng, Q. Wu, and W. Jia, "Clinical report guided retinal microaneurysm detection with multi-sieving deep learning," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1149–1161, May 2018.

[42] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-Net: Deep mining lesions for diabetic retinopathy detection," in *Proc. Int. Conf. Med. Imag Comput. Comput.-Assist. Intervent.* Quebec City, QC, Canada: Springer, 2017, pp. 267–275.

[43] Z. Zhao, K. Zhang, X. Hao, J. Tian, M. C. Heng Chua, L. Chen, and X. Xu, "BiRA-Net: Bilinear attention net for diabetic retinopathy grading," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1385–1389.

[44] B. Graham. (Aug. 6, 2015). *Kaggle Diabetic Retinopathy Detection Competition Report*. Accessed: Nov. 20, 2020. [Online]. Available: https://kaggle-forum-message-attachments.storage.googleapis.com/88655/2795/competitionreport.pdf

[45] Kaggle. *Diabetic Retinopathy Detection*. Accessed: Nov. 16, 2020. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection
[46] Kaggle. *Aptos 2019 Blindness Detection*. Accessed: Nov. 15, 2020. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection
[47] A. K. Gangwar and V. Ravi, "Diabetic retinopathy detection using transfer learning and deep learning," in *Evolution in Computational Intelligence*, vol. 1176. Pune, India: Springer, 2021, pp. 679–689.
[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[49] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[52] N. Aujla. (Jan. 28, 2015). *Retinal Imaging: How it Works & Why It's Important*. Accessed: Mar. 8, 2020. [Online]. Available: https://visionaryeyecentre.com/retinal-imaging-how-it-works-why-its-important

MOHAMMAD T. AL-ANTARY received the B.Sc. degree in computer information systems from The University of Jordan, Amman, Jordan, in 2015, and the M.Sc. degree in enterprise systems and database administration/computing and information systems from the University of Greenwich, U.K., in 2017, where he is currently pursuing the Ph.D. degree in computing and information systems with the School of Computing and Mathematical Sciences. He is currently working as a Teaching Assistant with the School of Computing and Mathematical Sciences, University of Greenwich. His research and professional interests include data science, medical image processing, machine learning, business intelligence, and big data.

YASMINE ARAFA received the Ph.D. degree from Imperial College London. She is currently a Senior Lecturer in disruptive technologies with the School of Computing and Mathematical Sciences, University of Greenwich. She has over 15 years' experience in software engineering and IT project management across vertical segments commercially and in academia. She has led research and development projects that involved the multi-disciplinary areas of affective computing, social interaction design, multi-agent systems, and their application in real-time smart systems. Her research interests include affective computing, data analytics, disruptive technologies, and machine learning.