IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 October 2024, accepted 24 November 2024, date of publication 2 December 2024, date of current version 12 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3510556

RESEARCH ARTICLE

Effective Diabetic Retinopathy Classification With Siamese Neural Network: A Strategy for Small Dataset Challenges

MARIA TARIQ[1], VASILE PALADE[1], (Senior Member, IEEE), AND YINGLIANG MA[2]
[1]Centre for Computational Science and Mathematical Modelling, Coventry University, CV1 5FB Coventry, U.K.
[2]Norwich Epidemiology Centre, School of Computing Sciences, University of East Anglia, NR4 7TJ Norwich, U.K.

Corresponding author: Maria Tariq (tariqm16@uni.coventry.ac.uk)

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/X023826/1.

ABSTRACT Early detection of diabetic retinopathy, a complication of vision loss in advanced stages of diabetes, is essential to avoid permanent vision impairment. However, the automatic detection of diabetic retinopathy through medical image processing requires a large number of training data to build a model with good performance. This poses a challenge when working with small datasets as these models need large datasets to perform well on unseen data. In this paper, we design a few-shot Siamese Neural Networks combined with pre-trained models, such as VGG16, ResNet50, and DenseNet121, to effectively differentiate between classes using small lesions in the retinal images. The proposed model is trained based on the similarity between the pair of images using a comparatively small dataset and performs well for a five-class classification problem. We use the Fine-Grained Annotated Diabetic Retinopathy (FGADR) and APTOS 2019 Vision Impairment Detection dataset, where a small ratio of training images is used to train the model. To evaluate our model, we conduct the testing on the remaining data and achieve good accuracy when trained on limited images, with fewer epochs and fewer parameters. The proposed model achieves high accuracy rates on five-class classification of 80% on FGADR and 81% on APTOS 2019 datasets, with a consistent quadratic weighted kappa (QWK) score of 0.89 across both datasets. Furthermore, we conduct an in-depth analysis of hyperparameter optimisation, specifically investigating different pair selection techniques, loss functions, and distance layers to thoroughly evaluate their impact on the performance of the model. Our proposed model demonstrates promising results when combined with an attention mechanism to perform multiclass classification of diabetic retinopathy using a limited number of eye fundus images, outperforming existing approaches with only a small number of epochs in training.

INDEX TERMS Transfer learning, Siamese neural network, diabetic retinopathy, few-shot learning, multiclass classification.

I. INTRODUCTION

Diabetic retinopathy (DR) is a retinal vascular disease that develops in the advanced stages of diabetes. In diabetes, patients suffer from high blood sugar levels, which can cause severe damage to different organs, including the retina [1]. High blood sugar damages the tiny vessels in the retina, leading to blood and small fluid leaks into the eye, resulting in diabetic retinopathy [2]. Initially, it starts with small blood leaks, blocked and swelling blood vessels, fatty deposits in the retina and progressing to more severe changes, including the formation of new blood vessels and the appearance of abnormal vessels. These symptoms lead to blurry vision, floating spots in the vision and eventually permanent vision impairment [2]. It is hard to notice the small changes in vision in the early stages, but it is also the right time to get it controlled with proper check-ups and treatment. Timely treatment and early detection can prevent 95% of severe vision loss from DR [3]. Between 2010 and 2050, the estimated number of people ages 40 or above who

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

have diabetic retinopathy will double from 7.7 million to 14.6 million [3]. According to the U.S. Department of Health and Human Services, DR is the leading cause of vision loss among working-age adults 20-74 years [4].

Different screening tools are available for human experts and ophthalmologists to evaluate the stages of DR. Colour fundus images are used mainly to perform automatic diagnosis of the stages of DR and other eye diseases. For automatic diagnosis, artificial intelligence techniques are highly useful in detecting DR, initial symptoms of DR, and the stages of DR. Deep learning has been used previously to satisfy the need to perform multiclass classification [5], but DR remains a challenge due to noise, quality and a limited number of images. Deep learning models perform best on unseen data only if it has sufficient data in the training set; otherwise, they start to overfit. The major challenge in medical analysis is the unavailability of public datasets due to privacy, ethics and security issues. Some techniques like Generative Adversarial Networks (GANs) [6] can be used to build artificial image datasets; moreover, in some stages, it also needs data to generate high-resolution images. The alternative approach is to use a model that only needs a few images to get trained, such as a few-shot Siamese model [7]. The contributions of this paper are given below:

- In terms of our model design, we have proposed a novel architecture that combines a similarity-based Siamese network with pre-trained models to extract important features and attention mechanisms to enhance the selected features. Specifically, we have employed VGG16 [8], ResNet50 [9], and DenseNet121 [10] for feature extraction combined with custom self-attention layers, enabling better performance of the multiclass classification model for the detection of diabetic retinopathy.
- The proposed model addresses a common challenge in medical research: the limited availability of data, particularly in the context of diabetic retinopathy. Our research contribution lies in the development of a computationally efficient model for few-shot learning. This model is trained on a few images from the dataset, specifically 54% in the training set, which results in reduced training time and rapid convergence, often completing training in only a few epochs. Notably, the model is trained on binary classification based on similarity and dissimilarity between the images, while its testing phase involves multiclass classification.
- Detailed experimentation has been conducted to optimise the pre-trained models combined with attention layers, enhancing their capacity to extract valuable features and detect small lesions within the retinal images. Furthermore, we also provided a valuable discussion about the influence of pairs of images and various distance metrics of the Siamese model to analyse how these hyperparameters affect the model's performance.

The rest of the paper is organised as follows. The background study is given in Section II, while Section III and IV discusses the proposed methodology of our work. Experimental results are stated in Section V with an extensive discussion. Finally, Section VI concludes the paper with some future work.

II. BACKGROUND STUDY

Deep learning is widely used in the field of medical image analysis. Medical images are analyzed for classification, segmentation, localization, and reconstruction tasks to improve clinical diagnosis and treatment [11]. However, a large amount of data is needed to perform these tasks, which is a huge challenge in the medical field. Deep learning models perform effectively well when a large amount of data is available in training, but it gets worse when there is little data to train. For deep learning in medical imaging, the biggest challenge is the requirement of sufficient data for experiments, and training [12]. There are two ways to get the data; one is from public repositories, and the other is directly from hospitals. It is hard to get sufficient data from hospitals as there are privacy concerns about patient data. Sometimes, data provided by hospitals misses important information about the data labelling. Furthermore, in some cases, the data from hospitals is not free for academic research. In the former way of getting data through public repositories, the challenges include the bad quality of data collected in different environments and light conditions, imbalanced classes, and a small number of images.

To resolve the imbalanced nature of the data, [13] have used a reinforcement learning model with transfer learning for binary classification. However, our work focuses more on smaller datasets using transfer learning [14]. Previously, transfer learning has been used to eliminate the need for comparatively big datasets and the number of epochs to train the model, as it transfers the weights of one problem to a different problem. In transfer learning, pre-trained models are used, which already got trained on different datasets, and do not need to be trained from scratch. In medical research, VGG16 [8], Inceptionv3, ResNet50 [9] and DenseNet121 [10] have mainly been used for binary and multiclass classification. These pre-trained models are trained on ImageNet, a completely different dataset that includes daily life images. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [15] includes more than one million annotated images, whereas fewer images are needed during transfer learning. VGG16 has been used to perform binary classification on Kaggle and Messidor 2 datasets for early detection of DR. In the United States, authors of [16] have performed multiple binary predictions using 128,175 DR images from the EyePACS dataset [17]. A high sensitivity of 97.5% and specificity of 93.4% for the EyePACS validation dataset and sensitivity of 96.1% and specificity of 93.9% for the Messidor-2 validation dataset is achieved using the pre-trained Inception-v3 model [16]. Different research efforts have been performed towards binary [[14], [18], [19], [20], [21]], 3-class [19], and

4-class classification [19], [21], [22] using the EyePACS dataset, a task relatively straightforward in achieving high accuracy. In comparison, it is challenging to do a five-class classification based on small lesions, which makes the difference between classes notably minimal. Therefore, our work focuses on five-class classification using pre-trained models to better distinguish the classes based on the lesion details.

Moreover, [23] has highlighted the drawbacks of the Kaggle EyePACS dataset through different experiments using ResNet50 and DenseNet121. This work mostly talked about the problems faced due to bad quality and incorrect annotations in publicly available datasets. An ensemble of different pre-trained models, ResNet50, Inceptionv3, Xception, Dense121, and Dense169, is used to achieve better prediction as compared to the performance achieved from the single model [24]. While working with a small dataset, it is likely to get stuck in overfitting the model where the training accuracy is good, but validation or test accuracy is not good enough to generalize the model. In order to reduce overfitting, there are some traditional techniques like data augmentation and L2 regularization in [19], dropout layer, and early stopping. Moreover, a combined model of an autoencoder and VGG19 has been used to improve test data performance and avoid overfitting [25].

In [13], an extensive discussion on traditional data augmentation techniques is done in addition to generating new data artificial data through GANs. Sometimes traditional real-time data augmentation techniques like random horizontal and vertical rotation, horizontal and vertical flips, and horizontal and vertical shifts are useful [18], but excessive oversampling can also lead to overfitting. Whereas GANs have been successful in the medical field, particularly in reconstructing medical images, however, they also face the challenge of insufficient medical data and achieving high-resolution images [13]. Some GANs have been used to generate high-resolution DR images, but they only focus on normal retinal images, lacking information about the stages of DR [26]. To generate high-resolution images that closely resemble the original images while incorporating lesion details, GANs need a substantial amount of training data. DR-GAN [27] and DR-LL-GAN [28] have been introduced to generate high-resolution synthetic samples using the Kaggle EyePACS dataset. These models have been trained by taking additional information about DR lesions and improving the model's performance to produce images for each class. However, it's important to note that these approaches have their limitations, such as the generation of synthetic lesion patterns of lower quality [27] and results that may not generalize well to other datasets [28]. Another way is to feed patches of images instead of images to eliminate small dataset problems. An image is divided into patches and can be fed with the same label for all the patches [29]. It is useful to feed images into the classifier as patches, as it can save a lot of computational resources and focus on small details in the image.

Another approach used nowadays is a few-shot learning model, the Siamese network [7], for tasks like image classification, similarity analysis, face recognition, and object detection. It is a neural network implemented to perform similarity analysis between a pair of inputs. It has been applied to medical image analysis, but its utilization in this field has seen relatively few contributions. It takes a pair of images as input and learns the similarity between the pairs. In [30], they have designed a deep Siamese network based on content-based medical image retrieval using the Kaggle EyePACS dataset. In [31], they have designed their model for multiclass classification following the pipeline of the Siamese model and achieved good accuracy compared to the well-known pre-trained models. In another work [32], they have done experiments on single-eye and both-eye diabetic retinopathy image analysis with a different attention mechanism. In our work, we have combined the Siamese network with pre-trained models to perform classification on a small dataset to get comparatively good accuracy in order to deal with the need for big datasets in medical problems. Our fusion model is trained on a few shots or a small dataset with 54% training, 13% validation and 33% testing. Our model has successfully learned the small lesions within the classes of diabetic retinopathy. In this work, we have also implemented an attention mechanism combined with pre-trained models to emphasize the relevant part of the image. Some experiments have been done on imbalanced classes to analyse the performance of our model. We have applied hyperparameter optimisation using different loss functions and distance layers to differentiate between the feature embedding.

III. METHODOLOGY

A. DATASET AND DISCRIMINATIVE FEATURES

Two datasets have been used in this work; one is FGADR [33], and the second is Kaggle APTOS 2019 Vision Impairment Detection [34]; the description is given in Table 2. The FGADR dataset has fine-grained annotations on 1842 fundus images with both pixel-level and image-level labels. For our study, we have adopted image-level grading of the dataset to access the overall severity and stages of DR rather than the specific location of the disease within the images. The image-level grading is done by three ophthalmologists over a period of 10 months. The dataset is taken from the UAE hospitals and is the property of the Inception Institute of Artificial Intelligence, Abu Dhabi. It comes under complete privacy protection, and the personal information of the patient was anonymised. The data has then been pre-processed to select the best quality images, with the same resolution of 1280×1280.

In Fig. 1, the discriminative features in the dataset are highlighted manually and can be seen. In class 0, there are no lesions; however, in class 1, there are microaneurysms (MAs), highlighted in green. Microaneurysms are small red stains on the retina, and it is the first sign that appears in stage 1 of DR. Retinal haemorrhages are also red spots that become visible in stage 2, highlighted in blue, so it is hard to separate them from microaneurysms. In stage 3, small

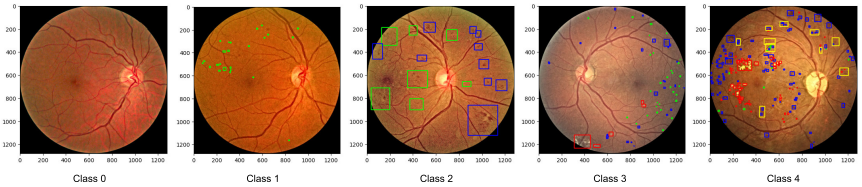

FIGURE 1. Sample Images from each class of FGADR dataset. Class 0 has no lesions, Class 1 has microaneurysms in green, Class 2 has haemorrhages in blue, Class 3 has exudates in red, and Class 4 has neovascularization highlighted in yellow.

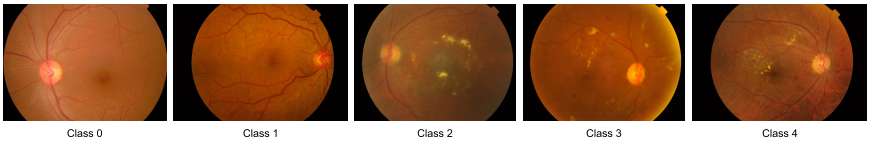

FIGURE 2. Sample Images from each class of APTOS 2019 dataset. Class 0 has no lesions, Class 1 has mild symptoms, Class 2 has moderate DR, Class 3 has severe DR, and Class 4 has proliferative DR. This dataset does not include any information on lesions and lesion location.

TABLE 1. Comparative analysis of previous studies.

Paper	Classification Model	Dataset	Classes	No of Images in Training	Training details	Results
[24]	Ensemble Model (Resnet50, Inceptionv3, Xception, DenseNet121, DenseNet169)	Kaggle EyePACS	5	35,126	Epochs=50 Optimiser= SGD and Adam	Accuracy=83.68%
[31]	Siamese Capsule Network	Messidor	2 (binary)	1200	Epochs= not given Optimiser= Adam	Accuracy=99.1%
[29]	AlexNet, VGG16, GoogLeNet, ResNet, Inception-v3	Kaggle EyePACS, eOphta dataset	2 (binary), 5	1050	Epochs=600	Accuracy =98%
[25]	Autoencoder and VGG Network	Kaggle EyePACS	5	31,631	Epochs= 100 Optimiser= Adam	Accuracy=76.27%
[32]	Siamese Network based CNN	Kaggle EyePACS	5	35,126	Epochs=60 Optimiser=SGD	Accuracy =84.6%, Kappa score (QWK)= 0.86
[30]	Deep Siamese CNN + ResNet50	Kaggle EyePACS	5	35,126	Epochs=not given Optimiser=Adam	MAP=0.6492 MRP=0.7737
[19]	GoogLeNet and AlexNet	Kaggle EyePACS, Messidor-1 dataset	2-ary, 3-ary, and 4-ary	35,126, 1200	Epochs=100	Accuracy=74.5% (2-ary), Accuracy=68.8% (3-ary), Accuracy=57.2% (4-ary)

TABLE 2. Dataset overview.

Dataset	Class 0	Class 1	Class 2	Class 3	Class 4	Total no of Training Images
FGADR	101	212	595	647	287	1842
Aptos2019	1805	370	999	193	295	3662

white and yellowish-white deposits get noticeable, known as hard exudates, and superficial white, pale yellow-white, and greyish-white are called soft exudates. These lesions can be seen in Class 3 in red colour. Then intra-retinal microvascular abnormalities (IRMA) and neovascularization (NV) appear in stage 4, which can be seen in yellow colour.

The grading distribution in the dataset is given below: ([‘grade’: the number of images] ‘0’: 101, ‘1’: 212, ‘2’: 595, ‘3’: 647, ‘4’: 287). Lesions start getting visible in the first and second stages, and it is also hard to differentiate those stages because the lesions are very small in the initial stages. It gets even harder for the later stages 3 and 4 because almost all lesions can be seen in those stages.

On the contrary, the Kaggle APTOS 2019 dataset comprises a total of 18,590 fundus images. However, access is restricted to the training set, which includes 3,662 images. The dataset also contains 1,928 images in the validation set and 13,000 images in the test set, gathered by the organizers of the Kaggle competition in collaboration with the Aravind Eye Hospital in India. Unfortunately, these validation and testing images are not publicly accessible.

This dataset is annotated by highly skilled medical professionals who have provided diagnoses for each individual image. These diagnoses categorize the images into one of five stages of Diabetic Retinopathy, ranging from 0 to 4. The sample images of the dataset are given in Fig. 2. The grading distribution in the dataset is given below: (['grade': the number of images] '0': 1805, '1': 370, '2': 999, '3': 193, '4': 295).

B. PREVIOUS EXPERIMENTS

In [23], we did a predictive analysis of the Kaggle EyePACS dataset [17], the largest publicly available dataset. This dataset has many highlighted challenges like poor quality of images, imbalanced classes and incorrect annotations. Different pre-trained models like ResNet50, EfficientNetB0 and DenseNet121 were used to perform transfer learning on the EyePACS dataset. In conclusion, we analysed conflicts between some of the classes of this dataset due to small unidentifiable lesions and incorrect annotations. The imbalanced nature of the dataset can be handled using different down-sampling and augmentation techniques, but incorrect annotations have no solution. The other available dataset, APTOS 2019, is relatively small in size. As a result, we cannot build a generalised model using that dataset primarily due to a high risk of overfitting. However, an accuracy of 93% was achieved on APTOS 2019 with the same model used for the Kaggle EyePACS dataset.

There are some ways to deal with small dataset problems; one is by generating synthetic images using GANs (Generative Adversarial Networks), and the other is by using a few-shot Siamese model. A Siamese model can be trained with fewer shots to build a generalised model as it learns to differentiate between the classes of the dataset with few examples in the training set.

TABLE 3. Experimental results on five classes using pre-trained models.

Dataset	Model	Classes	Test Accuracy
Kaggle EyePACS	SVM	5	52.57%
	DenseNet121	5	48%
APTOS 2019	DenseNet121	5	93%
FGADR	ResNet50	5	68%
	VGG16	5	60%

In Table 3, we achieved the best test accuracy, out of deep learning models, of 48% and 93% with DenseNet121 designed for five-class classification using Kaggle EyePACS [17] and APTOS 2019 dataset [34], respectively. We got slightly more accuracy of 52.57% on Kaggle EyePACS using support vector machines (SVM) as it gives the upper estimate of the model's performance. These results are taken from [23] to discuss the limitations of these datasets. Kaggle EyePACS dataset has a considerable number of images, but it has other challenges like the huge ratio of imbalanced classes in the dataset, bad quality of images and incorrect annotations. The other dataset, APTOS 2019, was used to train the model but was comparatively small to produce a generalised model. Later, FGADR was used with DenseNet121, though it did not perform well with this model; in contrast, ResNet50 worked well and showed an accuracy of 68% on five classes. In this work, we have used the FGADR dataset and proposed a different methodology to generalise the model well for multiclass classification. This methodology resolved the problem of the unavailability of a large number of images in medical research.

Algorithm 1 Training Pipeline for Siamese Neural Network

Input: $\mathcal{I} = \{x_i \mid i \in [1, n]\}$ (Unannotated Dataset), $\mathcal{M}$ (Pre-trained CNN Models with Custom Attention Layers)

Output: $\mathcal{E}_t$ (Set of predicted labels, if pair of images are similar or not.)

Function `PairOfImages` $(\mathcal{I})$ **:**
  if $x_i \in$ *SameClass and* $x_{i+1} \in$ *SameClass* **then**
    $S \leftarrow (x_i, x_{i+1}, 1)$ `// Images are Similar`
  else
    $D \leftarrow (x_i, x_{i+1}, 0)$ `// Images are Dissimilar`

Function `FeatureExtraction` $(\mathcal{M}, x_i)$ **:**
  $F(\mathcal{M}, x_i) \leftarrow$ `// Output vector of` (64×1)

Function `GetDistance` (x_i, x_{i+1}) **:**
  GetDistance(x_i, x_{i+1})
  $\leftarrow d(x_i, x_{i+1}) = \sqrt{\sum_{j=1}^{n}(x_{i,j} - x_{(i+1),j})^2}$

Function `GetPrediction` $(\mathcal{M}, x_i)$ **:**
  $\mathcal{E}_t \leftarrow$ GetPrediction$(\mathcal{M}, x_i) \leftarrow$ `// Prediction Results as 0 or 1`

Training:

$S, D \leftarrow$ PairOfImages$(\mathcal{I})$
for $t = 1, 2, 3, \ldots,$ *num_epochs* **do**
  for *each mini_batch*(S, D) **do**
    $f_i \leftarrow$ FeatureExtraction$(\mathcal{M}, x_i)$
    $d_i \leftarrow$ GetDistance(f_i, f_{i+1})
    $\mathcal{E}_i \leftarrow$ GetPrediction$(\mathcal{M}, d_i)$
    `// Compute the loss for the current prediction and actual label`
    $\mathcal{L}_i \leftarrow$ LossFunction$(\mathcal{E}_i, y_i)$
    `// Calculate the gradient of the loss with respect to the parameters`
    $\Lambda_i \leftarrow$ ComputeGradient$(\mathcal{L}_i, \theta)$
    `// Update the model parameters using an optimiser`
    $\theta \leftarrow$ Optimise$(\Lambda_i \mathcal{L}_i)$

C. SIAMESE NETWORK FOR FEW SHOT LEARNING

The Siamese neural network analyses the relationship between two instances by calculating the distance or similarity between their feature vectors. The architecture diagram of our model can be seen in Fig. 3 (a). The images from the

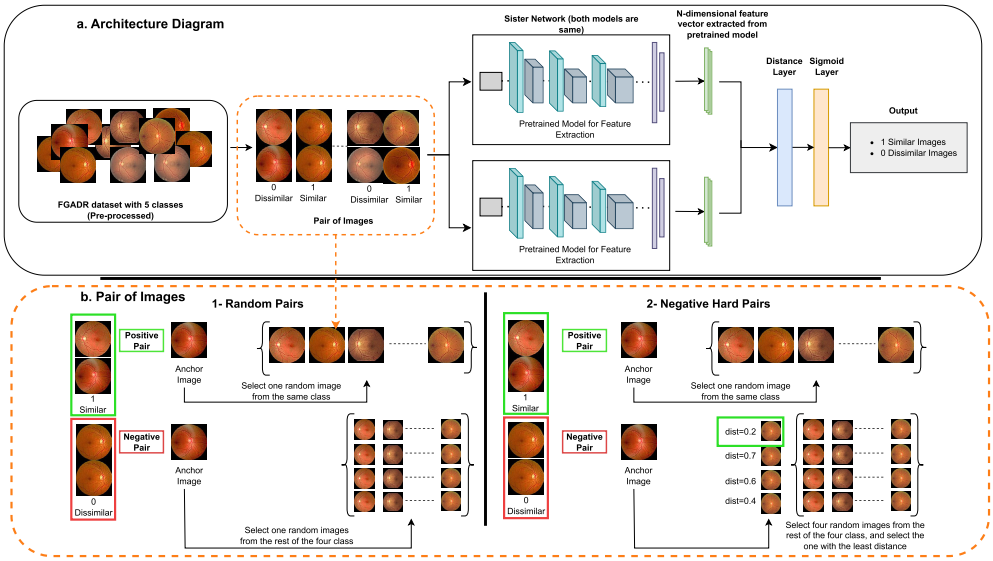

FIGURE 3. (a) Architecture diagram of siamese neural network. Pre-processed images are taken from the FGADR dataset to create the pairs of images, explained in (b). The images are fed into separate pre-trained models (e.g., VGG16, ResNet50, or DenseNet121) to extract N-dimensional feature vectors from their respective final layers. Each model operates independently, and their weights are not shared. A sigmoid layer is used to predict the relation between pairs of images based on similarity. (b) There are two ways to get the pair of images, Random pairs and Negative Hard Pairs. In Random pairs, both positive and negative pairs are selected randomly. However, in the case of a negative hard pair, a specific criterion based on distance is used to select the negative image from the rest of the four classes.

FGADR dataset are arranged into pairs based on similarity and dissimilarity. The same number of images from similar and dissimilar pairs are fed into a pre-trained model to get feature vectors. The model has two subnetworks with the same feature extraction model to build the feature space for the instances. Two resultant N-dimensional feature vectors are extracted, each from the twin networks. The feature vectors are fed into the distance layer to learn the relationship between the feature vectors and calculate a scalar value. The model is followed by a sigmoid layer to train the model for final prediction. The detailed pipeline of the model training can also be seen in Algorithm. 1.

1) FEATURE EXTRACTION

In this section, pre-trained models VGG16, ResNet50, and DenseNet121 have been used to extract feature vectors. VGG16 has 16 layers, which makes it effective for smaller datasets. ResNet50, with a suitable number of layers, does not have much depth like other pre-trained models, making it less complex for challenging datasets. It also has a previous record of good performance on medical images. DenseNet121 has 121 layers and is chosen due to its depth to extract complex features from the images. These pre-trained models have been chosen based on their frequent use and better performance on medical images. We have applied transfer learning to extract informative features from our resized images, which are used to feed to the Siamese network. Pre-trained models perform better on small datasets as they have already been trained on the 'Imagenet' dataset. We have used transfer learning, and only the two last layers of the model are trained to get useful features. A global average pooling layer is added to get a feature space of 64 dimensions.

In the Siamese network, we have used the same pre-trained model as a sister network, indicating the model takes two images as input. The input image is given to the model to get n-dimensional feature space. These features are then fed to the distance layer to evaluate the relation between the two images.

2) PAIR OF IMAGES

The Siamese model is designed to check the similarity between two inputs, which means it takes a pair of images each time for the training. We have converted our data to positive and negative pairs of images for the training. There are different ways to make pairs of images, as can be seen in Fig. 3 (b); in our work, we have done experiments using three different ways of arranging pairs.

Random pairs: There are two types of pairs; one is a positive pair, which has both images from the same class, and the other is a negative pair, which has both images from different classes. The pairs are determined against an anchor image. We randomly selected an image from the same class for the positive pair, and the label was given as 1 for similarity. Whereas for the negative pair, the image was randomly selected from the rest of the four different classes, and the label was given as 0 for dissimilarity.

Hard negative pairs: In this pair selection type, we followed the same random pair technique and randomly chose a positive pair for our anchor image. However, it is different for negative pairs; we realised that the random selection needs to be fixed. The validation accuracy worked fine, but

the models sometimes did not perform well for test images. We have incorporated some complex negative examples in pairs. We randomly picked four images from the rest of the four distinct classes, calculated the distance between each dissimilar image with the anchor image, chose the one with the maximum distance, and labelled this pair as 0. The distance is calculated using two different metrics, distance metric and Structural Similarity Index Metric (SSIM) to get the relationship between two vectors.

On the other hand, the structural similarity index metric (SSIM) is used to calculate the relation between two images based on the luminance, contrast and structure of the images. The formula of SSIM is given below:

$$SSIM = \frac{(2\mu_x\mu_y + C1)(2\sigma_{xy} + C2)}{(\mu_x^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C2)} \quad (1)$$

where x and y are the two feature vectors, μx and μy represent the mean of vectors respectively, σ_x^2 and σ_y^2 give the standard deviation of the vectors respectively, σ_{xy} is the covariance of the two vectors, and C1 and C2 are constants added for numerical stability.

3) DISTANCE METRICS

After the sister network, we have two feature embeddings of identical dimensions. The Siamese model adds a distance layer at the end of the base model to learn the relationship between the pair of images. The relationship between the pairs is calculated using different distance metrics. The metrics used in our work are given below:

Euclidean distance: It is a similarity metric to compute the relationship between the feature vectors and convert them to a scalar value. It is the straight line distance between two points in the Euclidean space. If the distance between the pair of images is small, then the images are similar. In comparison, the images are considered dissimilar if the Euclidean distance between the images is large.

The similarity between the images can be computed using the Euclidean distance formula as follows:

$$similarity = exp(-||A - B||^2) \quad (2)$$

where A and B are the feature vectors.

Cosine similarity: This metric calculates the similarity between two images based on the cosine value of the angles between two feature vectors. It determines whether the feature vectors are the same, depending on the direction of the vectors. If the images are similar, the vectors point in the same direction, and the cosine similarity value is close to one. Whereas if the images are dissimilar, the direction of the vectors is different, and the similarity value is close to zero. Cosine similarity is roughly the dot product of the two vectors; the formula of cosine similarity is given as follows:

$$similarity = \frac{(A.B)}{(||A||||B||)} \quad (3)$$

where A and B are the feature vectors.

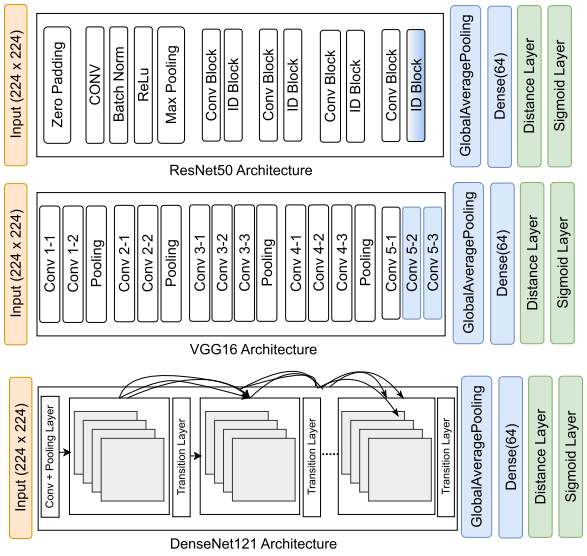

FIGURE 4. Architecture of ResNet50, VGG16, and DenseNet121 for feature extraction. The white-coloured layers of the models show that the weights of those layers are frozen. Custom layers are added to the end of the models to get informative features.

TABLE 4. Data splitting ratio of FGADR and APTOS 2019 dataset.

Dataset	Original Train Img	Pair of Images	Validation Images	Test Images	Reference Images
FGADR	266 (53%)	532	67 (13%)	167 (33%)	266
APTOS 2019	732 (53%)	1464	182 (13%)	458 (33%)	732

D. EXPERIMENTAL CONFIGURATION

For training, all the experiments in this study were conducted using a high-performance NVIDIA Quadro RTX 8000 graphics card. We used 100 images from each class and fed 53% in training and 33% in testing as shown in Table 4.

1) DATA PREPROCESSING

In the FGADR dataset, the images were already preprocessed to enhance the quality of the images. The extra black background was also removed; moreover, all images were resized to the 1280 × 1280 × 3 resolution. We further resized all the images for our work to 224 × 224 x 3 according to the input of our pre-trained models. Furthermore, pixel normalisation was applied to the images before feeding into the model. The same was done for APTOS 2019; the images were resized to 224 × 224 × 3 and pixel normalisation was applied.

2) FEATURE EXTRACTION AND TRANSFER LEARNING

We have primarily applied transfer learning for feature extraction using the VGG16, ResNet50, and DenseNet121 models, the architecture can be seen in Fig. 4. We have used these pre-trained models with the input size of 224 × 224 × 3. Only the last two layers were trained with informative features, and all the initial layers were frozen to prevent overfitting and speed up the training cycle. A global average

pooling layer was added at the end of the base model with an embedding of 64 and 128 dimensions to get a feature vector. It helped our model capture more complex features from our images without significantly increasing the trainable parameters. Some of the experiments have been done with different optimisers and batch sizes to analyse how this feature extraction affects the test accuracy of the model.

3) DEEP SIAMESE TRAINING

The images were divided into pairs, positive pairs and negative pairs. VGG16 and DenseNet121 performed well with random pairs; however, ResNet50 did best with hard negative pairs, using the NumPy function to calculate the distance between two negative images to get the best negative pairs. Features were extracted from these pairs, and then those features were fed to the Siamese network, where the distance between the feature space will be calculated to analyse the similarity between the pairs. Euclidean distance is well suited to calculate the distance between two feature vectors to give a scalar value, followed by a fully connected layer with a sigmoid activation function.

4) HYPERPARAMETER OPTIMISATION

In hyperparameter optimisation, we have used binary cross-entropy and contrastive loss functions as we have two classes to train if two images are similar or not. We applied different learning rates according to the optimisers. For the distance layer, we have done experiments with Euclidean distance and Cosine distance.

The learning rate used in these experiments is the default learning rate set for the Adam optimiser. Different experiments have been done using cropped images and full images, whereas, for some experiments, we have added batch normalization to preserve the useful information within the layers.

5) LOSS FUNCTIONS

There are two types of problems; one is a closed set problem, and the other is an open set problem. A closed set problem is a straightforward classification problem; we have some training data from specific classes and train the model for those classes. In comparison, an open set is different; we train our model based on the similarity and dissimilarity of the points and then evaluate it on a classification problem with entirely different classes. The model is considered good if it performs well on unseen images. The most used loss functions for the Siamese network are contrastive loss and binary cross-entropy loss.

Binary cross-entropy loss is the most commonly used loss function for binary classification problems. It is implemented in our work to calculate the loss between the predicted output and the actual output. In a classification task, the model's output would be a single scalar value of 0 or 1.

The equation for binary cross-entropy is as follows:

$$\text{Loss}_{\text{binary}} = -(y \log(p) + (1 - y) \log(1 - p)) \quad (4)$$

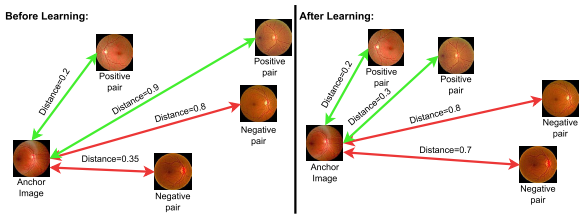

FIGURE 5. Visualisation of contrastive loss.

TABLE 5. Classification results for feature extraction models on FGADR dataset.

Model	Epochs	Optimizer	Training Accuracy	Validation Accuracy	Test Accuracy
VGG16	80	Adam	0.74	0.65	0.60
Siamese + (VGG16)	5	SGD	0.50	0.45	0.79
ResNet50	80	Adam	0.62	0.59	0.68
Siamese + (ResNet50)	5	Adam	0.50	0.41	0.80

where y is the true label (0 or 1), and p is the predicted probability of the positive class. For such metric learning problems, contrastive loss works well as it calculates the error based on the relationship between the two points. It preserves the relationship between the pairs of images after transformation. Logically, the images close to each other are similar, and those far away from each other are dissimilar. It penalises similar points from being too far from each other and dissimilar images from being close to each other. The formula of the contrastive loss function is given below:

$$\text{Loss}_{\text{contrastive}} = (1 - Y)(D^2) + Y_{max}((m - D), 0)^2 \quad (5)$$

where Y is the binary label showing the relationship between the pair of inputs, if similar (Y = 0) or dissimilar (Y = 1). D is the distance calculated between the feature vectors of the two data points, and m is the margin, which determines the distance threshold between the two inputs to preserve the relationship.

The contrastive loss can be visualised in Fig. 5. The concept of contrastive loss minimises the distance between two similar images while maximising the distance between two dissimilar images. The model gets evaluated based on the ability of the model to distinguish between similar pairs or different pairs.

E. MODEL EVALUATION FOR FIVE-CLASS CLASSIFICATION

Testing a Siamese model is different for multiclass classification because it requires the model to learn the representation space and distinguish well between the pairs to get better results. Our model is trained based on the pair's similarity and dissimilarity and is tested on multiple classes.

The model is tested on 33% of the whole dataset, and we have fed all of the training images for reference images. The data is pre-processed and fed to the model for feature embeddings and then for prediction. Table 5 presents a

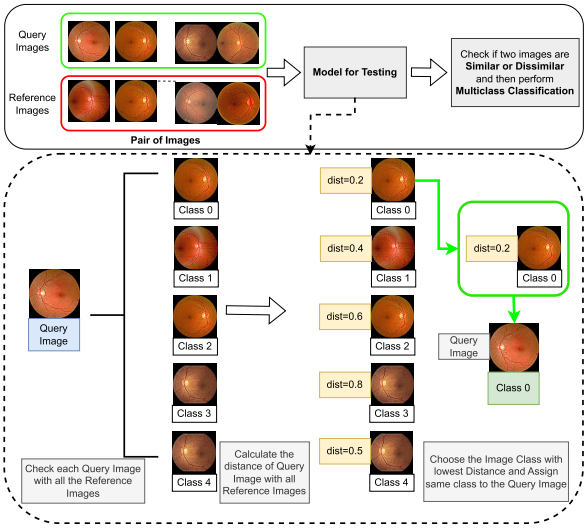

FIGURE 6. Model's evaluation phase for multiclass classification. Test images, considered as query images, undergo a comprehensive evaluation involving comparisons with all reference images. The model calculates distances based on image similarities and dissimilarities. The query image is assigned the label of the closest reference image, as determined by these distances.

comparison of results of simple VGG16 and ResNet50 models with Siamese Network combined with VGG16 and ResNet50 for feature extraction. Notably, the results show a significant improvement when incorporating the Siamese Network with the pre-trained models, and it's noteworthy that this improved performance is achieved with just five training epochs. In these findings, the training accuracy and validation accuracy tell us how well the model is learning similarities between the images. We use early stopping to prevent it from learning too much and avoid overfitting. But the test accuracy is completely different; it gives us the accuracy of five-class classification.

During the testing phase, each test image serves as the query image and undergoes evaluation against all reference images from the training set to obtain distance embeddings. The minimum value of all the distance values is selected, which means that these two images, the reference and query image, have less distance between each other and are more similar to each other. Consequently, the label of the closest reference image is assigned to the query image. This process is repeated for all test images, resulting in a multiclass classification of the model using the Siamese Network, as illustrated in Fig. 6. The model performance is assessed using different performance metrics:

Accuracy: Accuracy gives the percentage of correctly predicted labels out of the total number of samples. To evaluate the accuracy of our model, the predicted labels are compared with the ground truth labels for a given set of images.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{6}$$

Quadratic Weighted Kappa: QuadraticWeighted Kappa (QWK) takes into account the agreement between predicted and actual labels, adjusted for the possibility of chance

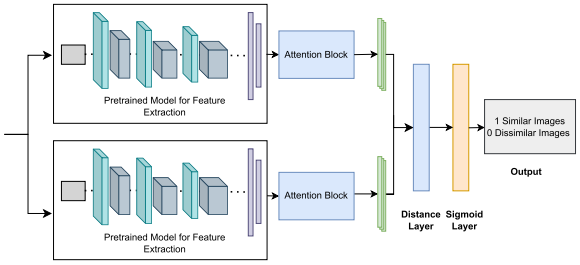

FIGURE 7. In our updated model architecture, we have integrated an attention mechanism with a pre-trained model. The attention block is seamlessly integrated into the pipeline after obtaining features from the pre-trained model. Subsequently, a feature vector is extracted following the application of the attention block.

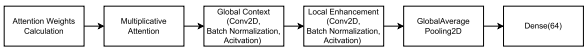

FIGURE 8. Attention layers combined with DenseNet121.

FIGURE 9. Attention layers combined with VGG16.

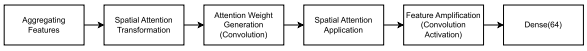

FIGURE 10. Attention layers combined with ResNet50.

agreement. It considers class imbalance by applying class weights.

$$QWK = \frac{p_o - p_e}{1 - p_e} \tag{7}$$

where: p_o is the observed agreement ratio and p_e is the expected agreement ratio.

IV. ATTENTION MECHANISM COMBINED WITH PRE-TRAINED MODELS

There are certain limitations while working with CNN-based pre-trained models. These models process images as a whole and do not pay explicit attention to informative parts of the image. The convolutional filters are applied directly to the entire image, extracting hierarchical features from lower-level edges and textures to higher-level patterns and object representations. In our problem, we have lesions that appear gradually with each stage of diabetic retinopathy, where existing lesion also stays. This is a complex scenario which includes the evolution of lesions across different stages of DR, and it shows overlapping between the features of classes. Due to lesion progression, it gets difficult for neural networks to differentiate between classes of DR. However, pre-trained models need to be more advanced to address the problem of overlapping features. Due to this, the attention mechanism is introduced with the pre-trained models to extract features. The attention mechanism allows the model to focus on important features of the image, which can be ignored easily by traditional CNNs [35] while

de-emphasizing other non-important features. It enhances the feature extraction process and improves the accuracy of the classifier [35]. The updated architecture of our model is shown in Fig. 7. Different attention layers are used for each model to ensure that they are specifically optimized to complement the unique feature representations and architectural characteristics of each pre-trained model.

A. ATTENTION LAYERS COMBINED WITH DENSENET121, VGG16 AND RESNET50

Implementing an attention mechanism with DenseNet121, the initial attention weights calculated determine how much emphasis the network should place on each pixel while processing the successive layers. Element-wise multiplication is used to scale the input image according to the weights. The approach directs attention to the global features of the image and then applies the local enhancement to get local features fused with the global features. It is utilised thoroughly to make full use of the global information of eye fundus images. The features extracted by the attention networks are filtered through the global average pooling layer and augmented by batch normalisation with sigmoid activation. The layers used for the attention mechanism are given in Fig. 8.

Attention block is integrated into the architecture following VGG16, which can be seen in Fig. 9. This block comprises a series of layers dedicated to directing the model's focus towards the most informative regions within the input image. Initially, the relevant features extracted from VGG16 are aggregated for further refinement. The features are reshaped to align with the next layers, where transformation is done to prepare data for the calculation of attention weights. The attention weights are calculated for different parts of the input features to indicate the score of each feature element. The weights are then applied to the original input feature data through element-wise multiplication. This process serves to selectively emphasise useful features while de-emphasizing irrelevant ones.

The attention layers from Fig. 10, when combined with ResNet50, empowered the model to extract meaningful features from the input image. The process begins with the aggregation of the features derived from the following layers of the ResNet50 model, where it undergoes a spatial transformation layer to align the data with the subsequent layer of attention weight calculation. Attention weights are calculated through convolutional operations, allowing the model to prioritise salient features from the input image. The computed attention weights are applied to the original input image to highlight the informative features and diminish the influence of less important ones. Further refinement is done through feature amplification by convolutional layers with activation functions, which fine-tunes the selected features extracted from the input image.

V. RESULTS AND DISCUSSION

Small datasets have significant challenges in training a reliable model for classification due to the limited number of samples from which to learn. Many previous researchers have already worked on small dataset problems to improve the performance of the model. Some used data augmentation, transfer learning and few-shot techniques. Our proposed model follows the few-shot approach combined with transfer learning to perform classification on DR images. It is then further improved by using attention layers in combination with the pre-trained models to improve the feature extraction process. We extensively analysed the dataset, the hyperparameters of the Siamese model combined with pre-trained models ResNet50, VGG16, and DenseNet121.

A comprehensive analysis has been done between our work and the previous studies, which utilise the same dataset, FGADR. The results are given in the Table. 6. In [36], the training is done on 80% of the data with QWK of 0.8389 on 20% test set, and 93.9% of the data is trained to get an accuracy of 71% on 6.1% of the test set [37]. In contrast, our work has achieved a prominent accuracy of 80% and QWK of 0.89 on 33% of the test set when the model is trained on just 54% of the dataset. From the FGADR dataset, we randomly selected 100 images from each class to fetch balanced data from all classes, as class 0 only has 101 images. We aimed to eliminate any potential bias caused due to class imbalance, so we maintained an equal number of images in each class. In total, we got 500 images, then we distributed the data into 54% training, 13% validation and 33% testing sets. We followed a different training pattern and trained our model

TABLE 6. Previous studies utilizing the FGADR dataset for comparative evaluation with our research.

Model	Data split	Data Aug.	Test Accuracy	Kappa Score (QWK)
[[36]]	Train (80%) Test (20%)	Yes	-	0.8389
[[37]]	Train (93.9%) Test (6.1%)	Yes	71%	-
Ours				
Siamese + ResNet50	Train (54%) Val (13%) Test (33%)	No	80%	0.87
Siamese + VGG16	Train (54%) Val (13%) Test (33%)	No	79%	0.87
Siamese + DenseNet121	Train (54%) Val (13%) Test (33%)	No	76%	0.89

TABLE 7. Previous studies utilizing the APTOS 2019 vision impairment detection dataset for comparative evaluation with our research.

Model	Data Split	Data Aug	Test Accuracy	Kappa Score (QWK))
[[38]]	-	Yes	-	0.79
[[39]]	-	-	77%	0.78
Ours				
Siamese + VGG16	Train (54%) Val (13%) Test (33%)	No	81%	0.89
Siamese + ResNet50	Train (54%) Val (13%) Test (33%)	No	79%	0.86

TABLE 8. Comparative analysis with previous studies employing the siamese network architecture.

Model	Loss Function	Pair of Images	Distance Metric	Image Emb.	QWK
[[40]]	Adam	-	Euclidean	-	0.83
[[32]]	MSE	-	-	-	0.86
Ours					
Siamese + ResNet50	Contrastive	Hard Negative	Euclidean	64	0.87
Siamese + VGG16	Contrastive	Random	Cosine	128	0.87
Siamese + DenseNet121	Contrastive	Random	Cosine	64	0.89

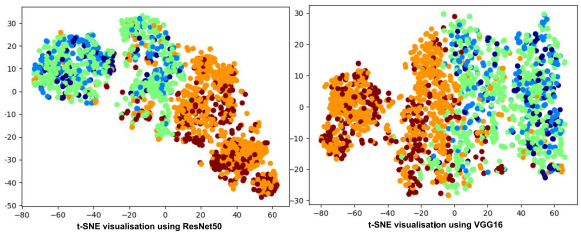

FIGURE 11. t-SNE visualisation of FGADR dataset: Distribution and clustering of classes.

on a few images, just 266 images in training without data augmentation. The model is trained for 11 epochs to achieve an average test accuracy of 80% and an average Kappa score of 0.89 on 3-fold cross-validation when performing five-class classification. To emphasise the key findings, our test results show the accuracy and Kappa score on 33% of the test data.

On the other hand, Table 7 presents a comprehensive view of our study's application to the APTOS 2019 dataset, along with a comparative analysis against previous research studies. Reference [38] conducted a comparative analysis of different pre-trained models on the APTOS dataset, achieving the highest Kappa score of 0.79 using EfficientNet-b4. Furthermore, [39] employed a deep VGG model, yielding an accuracy of 77% and a Kappa score of 0.78. In our prior work, we achieved a commendable accuracy of 93% with DenseNet121 on the APTOS 2019 dataset, as indicated in Table 3. However, this achievement raised concerns about potential overfitting and dataset dependency. In contrast, our present study focuses on training based on similarity and assesses the model's performance in a five-class classification scenario, thereby enhancing its versatility and applicability in a broader context.

In Table 8, [32], [40] are compared with our work as these studies used siamese architecture to perform classification using all the images of the Kaggle EyePACS dataset for training. They used the original Kaggle EyePACS dataset, the largest available dataset and applied data augmentation to improve the model's performance and got a QWK score of 0.83 [40] and 0.86 [32]. In contrast, our research presents a novel approach featuring a few-shot Siamese Network, trained with pre-trained models, using a relatively small dataset. To facilitate learning, we employ a contrastive loss function, enabling the model to learn the comparative distances between similar and dissimilar images. The performance of the model is highly affected by the pair of images, introducing an element of randomization that profoundly influences its performance.

We analysed that selecting pairs of images plays a crucial role in the training process. The selection of pairs of images is central to the Siamese architecture, which aims to learn the similarity or dissimilarity between two input samples. By providing well-suited pairs of similar and dissimilar images, the model can learn to distinguish well between the classes. In this concept, we did not directly apply data augmentation, but making pairs allows effective indirect data augmentation. Positive and negative pairs were created from the available data to improve the model's generalizability.

Having diverse and representative pairs ensures that the model can accurately assess the relationships between unseen samples, enabling the model's ability to learn meaningful representations and effective multiclass classification. Talking about diversity, it is important during the training process if we have examples which are good representatives of the whole data. However, an adverse scenario emerges when the training set lacks the required diversity, which, in turn, impacts the model's efficacy, potentially leading to less favourable outcomes for specific folds of evaluation.

The occurrence of less favourable outcomes within a single fold of cross-validation can be further explained by visualizing the overlapping patterns inherent within the dataset. For this purpose, we employ t-SNE (t-distributed Stochastic Neighbor Embedding) to observe the behaviour or distribution of classes in reduced dimensional space. We have observed some valuable insights within the dataset, as there are overlapping classes because of shared patterns and features within the images, which can be seen in Fig. 11. It is because the lesions we see in Class 1 also appear in Class 2, and the features in Class 2 appear in Class 3. This visualisation also indicates that specific attributes or characteristics are not distinct enough to separate the classes. On the other hand, we can also analyse that there is much intra-class diversity; the classes are not tightly clustered but instead spread out, leading to overlapping regions with other classes. The presence of overlapping classes poses challenges for classification tasks. It suggests that distinguishing between certain classes based solely on the available features might be more difficult.

The presence of overlapping classes in our dataset contributes to variations in performance across different folds during cross-validation. We analysed this variation of performance across different folds during our experiments; specifically, in one fold, we analysed comparatively less accuracy than the other two folds. It is due to the overlapping classes which introduce ambiguity and make it more challenging for the model to distinguish between them accurately. As a result, the performance varied across our 3-folds, leading to a reducing effect on the scores. The average score of two folds with a majority score can be seen in Table 9, which represents the reducing effect of less score in one fold. Some classification models may be

TABLE 9. Comparison of multiclass classification performance using pre-trained models for diabetic retinopathy.

Papers	Backbone Model	Dataset Size	Epochs	Optimiser	Accuracy	Kappa Score (Quadratic Weighted)	Avg. Accuracy Score for Majority Folds
[[18]]	CNN	78000	125	SGD	75% (Validation)	-	
[[41]]	ResNet50 InceptionV3 Ensemble	3562	75	-	-	0.776 0.820 0.824	
[[42]]	InceptionV3	4000	-	-	48.8%	-	
[[43]]	VGG16 InceptionV3	166		SGD		50.03% 63.23%	
Ours							
-	ResNet50	500	11	Adam	80%	0.87	95%
-	VGG16	500	11	Adam	79%	0.87	95%
-	DenseNet121	500	11	Adam	76%	0.89	95%

TABLE 10. Comparing models with and without attention mechanisms for multiclass classification.

	FGADR		APTOS 2019	
Our Siamese Model + Pre-trained Model	Epochs	Avg. QWK	Epochs	Avg. QWK
VGG16	5	0.77	5	0.89
VGG16 (With Attention)	11	0.89	5	0.86
ResNet50	5	0.77	5	0.86
ResNet50 (With Attention)	11	0.88	5	0.87
DenseNet121	11	0.89	-	-
DenseNet121 (With Attention)	11	0.88	-	-

more sensitive to overlapping classes than others. Our choice of model architecture, hyperparameters, and optimisation approach impact how well our model handles overlapping instances and generalises across different folds.

Table 9 gives a comparative analysis of our pre-trained models with the previous work done to perform multiclass classification on Diabetic Retinopathy. Our models showcase adequate accuracy and a QWK score on only 500 images from the dataset to perform training and testing in just 11 epochs using Adam optimiser. The model in [18] and [41] utilised 125 and 75 training epochs, respectively, to get trained on five-class classification. The model achieved an accuracy rate of 75% when evaluated against the validation set [18], and a QWK score of 0.824 [41]. In [42], they have reported an accuracy of 48.8% using InceptionV3, whereas in [43], a significantly smaller dataset of only 166 images is used to attain comparatively less QWK of 50.03% and 63.23%, making our model stand out with a QWK score of 0.89.

Many pre-trained models, especially those trained on large-scale datasets with diverse classes, have demonstrated robust performance despite overlapping instances. We chose class balancing, feature selection, transfer learning and attention mechanism with different hyperparameters to learn representations well and improve our model's ability to handle class overlap. We experimented with different approaches and evaluated the impact on the model's performance to determine the most suitable combination of our hyperparameters. ResNet50 performed well with hard negative mining of image pairs when the contrastive loss was applied to train the model. This indicates that ResNet50 can effectively learn from pairs of images, focusing on differentiating between similar and dissimilar images. On the other hand, VGG16 demonstrates good performance when trained with random pair selection and binary cross-entropy loss function. This suggests that VGG16 may be more adept at handling random pairs and binary classification tasks, where the focus is on separating classes rather than explicitly learning from image pairs. These observations highlight the importance of choosing the appropriate model architecture and training techniques based on the characteristics of the dataset and the specific problem at hand. It's crucial to experiment and evaluate different approaches to identify the most suitable combination for achieving optimal performance and handling class overlap effectively.

Discussing the integration of attention mechanisms with pre-trained models, we observe a modest improvement in the model's performance. While the accuracy of the model shows some fluctuations within the folds, a notable improvement is evident in the Quadratic Weighted Kappa (QWK) metric, as shown in Table 10. Specifically, for VGG16, the QWK rises to 0.89, while for ResNet50, it attains a value of 0.88. In the case of DenseNet121, the QWK maintains the score. To provide context, the Quadratic Weighted Kappa is a metric used to measure the agreement between predicted and true classifications and in our work, this metric is used to measure the performance of the model. To our conclusion, we have succeeded in extracting salient features from the images, and it is important to acknowledge that, given the characteristics of this dataset, we may have approached an upper limit in terms of achievable scores. The current performance levels achieved with attention mechanisms likely represent a notable milestone, leaving limited room for further improvement.

To verify this hypothesis, we have undertaken ensemble modelling to evaluate our model's potential for further improvement. We have integrated pre-trained models, VGG16, ResNet50, and DenseNet121, implemented with attention layers and applied a majority voting technique to extract optimal results. The concept of majority voting

TABLE 11. Ensemble model results: Average accuracy across three folds.

Dataset	Ensemble Models	Epochs	Avg. Acc. (3 folds)
FGADR	With Attention Mechanism (VGG16 + ResNet50 + DenseNet121)	11	78%
	Without Attention Mechanism (VGG16 + ResNet50 + DenseNet121)	11	78%
APTOS 2019	With Attention Mechanism (VGG16 + ResNet50)	5	80%
	Without Attention Mechanism (VGG16 + ResNet50)	5	81%

mirrors the decision-making process done by medical experts in clinical settings. When differences arise during annotations and examinations, clinicians often take a majority decision to arrive at a consensus. In our research, we imitate this practice by implementing the majority voting technique. Our comprehensive analysis has led us to a conclusion where the ensemble model, guided by majority voting, yields results similar to the Quadratic Weighted Kappa (QWK) scores previously attained, as illustrated in Table 11. It reaffirms our model's performance, indicating that we have approached a plateau in terms of achievable metrics.

VI. CONCLUSION AND FUTURE WORK

In this comprehensive study, we designed a similarity-based Siamese network combined with VGG16, ResNet50, and DenseNet121, which is generalised well on the FGADR and APTOS 2019 dataset, to perform an effective multiclass classification of diabetic retinopathy. Our proposed model achieved promising results in learning the small lesions and effectively distinguishing between overlapped classes with an average accuracy of 80% and Kappa score of 0.87 on 3-fold cross-validation. Through extensive experimentation and analysis of hyperparameters, we gained insights into the factors that influence the training and testing of our model. Despite the limited number of images, it allowed us to optimise the model's performance and training time by attaining comparatively good classification results in just five epochs.

Furthermore, our research investigates the influence of pairs of images and different distance metrics of the Siamese model, which provides valuable insights into these hyperparameters in optimising the model's performance. The results of this study have significant implications in dealing with small dataset challenges in medical image classification, particularly in the field of diabetic retinopathy. Achieving good model performance demonstrates the potential of our model in aiding the detection of stages of DR and treatment decisions of patients. While our study represents a significant step forward, certain limitations should be considered. The selection of a pair of images and the overlapped nature of the classes may have influenced our results. Future research should focus on the expansion of methodology for the selection of image pairs to further enhance the robustness of our approach.

Moreover, we have conducted an exploration of our research by integrating the attention mechanism with the pre-trained models to obtain informative features from the images. Additionally, an ensemble model is also designed to achieve the optimal performance of the model. In summary, this research contributes to the advancement of medical image classification by addressing the challenges of achieving a generalised model using a small dataset for training. The insights gained from our study can promise to provide guidance in future research to develop more accurate and robust models for multiclass classification tasks within the limited resources of the medical field.

The code for the Siamese neural network implementation can be found at https://github.com/tariqm16/Effective-Diabeti
c-Retinopathy-Classification-with-Siamese-Neural-Network/tree/main.

REFERENCES

[1] D. Soumya and B. Srilatha, "Late stage complications of diabetes and insulin resistance," *J. Diabetes Metabolism*, vol. 2, no. 9, p. 1000167, 2011.

[2] M. Wang, Q. Li, M. Jin, Z. Wang, X. Zhang, X. Sun, and Y. Luo, "Noncoding RNAs are promising therapeutic targets for diabetic retinopathy: An updated review (2017–2022)," *Biomolecules*, vol. 12, no. 12, p. 1774, Nov. 2022.

[3] R. Jacobs, U. Tran, H. Chen, A. Kassim, B. G. Engelhardt, J. P. Greer, S. G. Goodman, C. Clifton, C. Lucid, L. A. Vaughan, B. N. Savani, and M. Jagasia, "Prevalence and risk factors associated with development of ocular GVHD defined by NIH consensus criteria," *Bone Marrow Transplantation*, vol. 47, no. 11, pp. 1470–1473, Nov. 2012.

[4] *Health, United States, 2009: In Brief., no. 2010*. Atlanta, GA, USA: U.S. Department of Health and Human Services, Centers for Disease Control and, National Center for Health Statistics (U.S.), Centers for Disease Control, Prevention (U.S.), Hyattsville, MD, USA, 2010.

[5] R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, "Automated detection of mild and multi-class diabetic eye diseases using deep learning," *Health Inf. Sci. Syst.*, vol. 8, no. 1, p. 32, Dec. 2020.

[6] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for medical image analysis," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101938.

[7] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. CML Deep Learn. Workshop*, vol. 2, Lille, France, 2015, pp. 1–8.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[9] 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.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 2261–2269.

[11] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[12] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps: Automation of Decision Making*, vol. 26. Cham, Switzerland: Springer, 2018, pp. 323–350.

[13] M. Tariq, V. Palade, Y. Ma, and A. Altahhan, "Diabetic retinopathy detection using transfer and reinforcement learning with effective image preprocessing and data augmentation techniques," in *Fusion of Machine Learning Paradigms: Theory and Applications*. Cham, Switzerland: Springer 2023, pp. 33–61.

[14] M. Tsighe Hagos and S. Kant, "Transfer learning based detection of diabetic retinopathy from small dataset," 2019, *arXiv:1905.07203*.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[16] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016.
[17] *Kaggle: Diabetic Retinopathy Detection*. Accessed: May 20, 2020. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy detection
[18] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Proc. Comput. Sci.*, vol. 90, pp. 200–205, Jan. 2016.
[19] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," *AMIA Summits Transl. Sci. Proc.*, vol. 2018, p. 147, May 2018.
[20] S. Mohammadian, A. Karsaz, and Y. M. Roshan, "Comparative study of fine-tuning of pre-trained convolutional neural networks for diabetic retinopathy screening," in *Proc. 24th Nat. 2nd Int. Iranian Conf. Biomed. Eng. (ICBME)*, Nov. 2017, pp. 1–6.
[21] 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.
[22] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Opthalmology Vis. Sci.*, vol. 57, no. 13, p. 5200, Oct. 2016.
[23] M. Tariq, V. Palade, and Y. Ma, "Transfer learning based classification of diabetic retinopathy on the kaggle eyepacs dataset," in *Proc. Int. Conf. Med. Imag. Comput.-Aided Diagnosis*, Dec. 2022, pp. 89–99.
[24] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. Ahmed Khan, and W. Jadoon, "A deep learning ensemble approach for diabetic retinopathy detection," *IEEE Access*, vol. 7, pp. 150530–150539, 2019.
[25] N. Barhate, S. Bhave, R. Bhise, R. G. Sutar, and D. C. Karia, "Reducing overfitting in diabetic retinopathy detection using transfer learning," in *Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Oct. 2020, pp. 298–301.
[26] M. Kim, Y. N. Kim, M. Jang, J. Hwang, H.-K. Kim, S. C. Yoon, Y. J. Kim, and N. Kim, "Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization," *Sci. Rep.*, vol. 12, no. 1, p. 17307, Oct. 2022.
[27] Y. Zhou, B. Wang, X. He, S. Cui, and L. Shao, "DR-GAN: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 56–66, Jan. 2022.
[28] S. Hameed Abbood, H. N. A. Hamed, M. S. Mohd Rahim, A. H. M. Alaidi, and H. T. S. Alrikabi, "DR-LL gan: Diabetic retinopathy lesions synthesis using generative adversarial network," *Int. J. Online Biomed. Eng. (iJOE)*, vol. 18, no. 3, pp. 151–163, Mar. 2022.
[29] C. Lam, C. Yu, L. Huang, and D. Rubin, "Retinal lesion detection with deep learning using image patches," *Investigative Opthalmol. Vis. Sci.*, vol. 59, no. 1, p. 590, Jan. 2018.
[30] Y.-A. Chung and W.-H. Weng, "Learning deep representations of medical images using Siamese CNNs with application to content-based image retrieval," 2017, *arXiv:1711.08490*.
[31] G. U. Nneji, J. Cai, J. Deng, H. N. Monday, S. Nahar, G. T. Mgbejime, E. C. James, and S. K. Woldeyes, "A dual weighted shared capsule network for diabetic retinopathy fundus classification," in *Proc. Int. Conf. High Perform. Big Data Intell. Syst. (HPBD&IS)*, Dec. 2021, pp. 297–302.
[32] R. Nirthika, S. Manivannan, and A. Ramanan, "Siamese network based fine grained classification for diabetic retinopathy grading," *Biomed. Signal Process. Control*, vol. 78, Sep. 2022, Art. no. 103874.
[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.
[34] Kaggle, "Aptos 2019 diabetic retinopathy dataset," 2019. [Online]. Available: https://www.kaggle.com/competitions/ aptos2019-blindness-detection/data
[35] J. Zou, X. Zhang, and X. Lin, "A diabetic retinopathy classification method based on novel attention mechanism," in *Proc. 17th Int. Conf. Intell. Comput. (ICIC)*, Shenzhen, China. Cham, Switzerland: Springer, 2021, pp. 129–142.
[36] L. Tian, L. Ma, Z. Wen, S. Xie, and Y. Xu, "Learning discriminative representations for fine-grained diabetic retinopathy grading," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.
[37] B. Hou, "High-fidelity diabetic retina fundus image synthesis from freestyle lesion maps," *Biomed. Opt. Exp.*, vol. 14, no. 2, p. 533, 2023.
[38] A. Pak, A. Ziyaden, K. Tukeshev, A. Jaxylykova, and D. Abdullina, "Comparative analysis of deep learning methods of detection of diabetic retinopathy," *Cogent Eng.*, vol. 7, no. 1, Jan. 2020, Art. no. 1805144.
[39] O. Dekhil, A. Naglah, M. Shaban, M. Ghazal, F. Taher, and A. Elbaz, "Deep learning based method for computer aided diagnosis of diabetic retinopathy," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Dec. 2019, pp. 1–4.
[40] 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.
[41] M. Al-Smadi, M. Hammad, Q. B. Baker, and S. A. Al-Zboon, "A transfer learning with deep neural network approach for diabetic retinopathy classification," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 11, no. 4, p. 3492, Aug. 2021.
[42] S. Masood, T. Luthra, H. Sundriyal, and M. Ahmed, "Identification of diabetic retinopathy in eye images using transfer learning," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, May 2017, pp. 1183–1187.
[43] 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.

MARIA TARIQ is currently pursuing the Ph.D. degree in deep learning for the automatic detection of diabetic retinopathy using eye fundus images with the Centre for Computational Sciences and Mathematical Modeling, Coventry University, U.K. She is conducting in-depth research and analysis on eye retinal images for the detection of diabetic retinopathy using deep and transfer learning techniques. She joined Coventry University, in 2021, after being associated with the Biomedical Research Laboratory, NUCES-FAST, Pakistan, for one year, where she was working in the field of image pre-processing and deep learning for the detection of breast cancer. Her research interests include deep learning, machine learning, image pre-processing, transfer learning, and deep reinforcement learning.

VASILE PALADE (Senior Member, IEEE) received the Ph.D. degree from the University of Galati, Romania, in 1999. He is a Professor of artificial intelligence and data science with the Centre for Computational Sciences and Mathematical Modelling, Coventry University, U.K. He joined Coventry University, in 2013, after working with the Department of Computer Science, University of Oxford, U.K., from 2001 to 2013. His research interests include machine learning and applications, including deep learning and neural networks, various nature inspired optimization algorithms, computer vision, and natural language processing.

YINGLIANG MA received the Ph.D. degree in computer science from The University of Manchester. He leads the Vision and Graphics Research Group, University of East Anglia. His research primarily focuses on developing innovative computer vision and computational geometric algorithms, with applications in medical image analysis, image-guided surgery, and procedure risk assessment. As a Principal Investigator (PI) or a Co-PI, he has conducted research in computer vision, machine learning, AI, and parallel computing, funded by EPSRC and NIHR. He has been invited to speak at numerous national and international institutions.