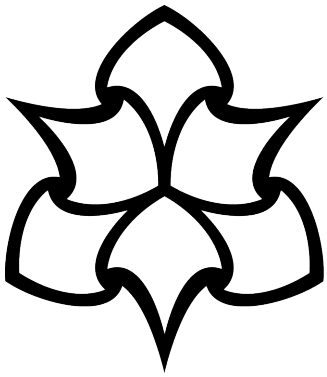

Computers in Biology and Medicine 146 (2022) 105602

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images

Md Robiul Islam [a], Lway Faisal Abdulrazak [b], Md Nahiduzzaman [a], Md Omaer Faruq Goni [a], Md Shamim Anower [c], Mominul Ahsan [d], Julfikar Haider [e], Marcin Kowalski [f,*]

[a] Department of Electrical & Computer Engineering, Rajshahi University of Engineering & Technology, Rajshahi, 6204, Bangladesh
[b] Department of Computer Science, Cihan University Sulaimaniya, Sulaimaniya, 46001, Kurdistan Region, Iraq
[c] Department of Electrical & Electronic Engineering, Rajshahi University of Engineering & Technology, Rajshahi, 6204, Bangladesh
[d] Department of Computer Science, University of York, Deramore Lane, Heslington, York, YO10 5GH, UK
[e] Department of Engineering, Manchester Metropolitan University, Chester St, Manchester M1 5GD, UK
[f] Military University of Technology, Warsaw, Poland

ARTICLE INFO

Keywords:
Diabetic retinopathy
Image analysis
CNN
Supervised contrastive learning
CLAHE
t-SNE

ABSTRACT

Diabetic Retinopathy (DR) is a major complication in human eyes among the diabetic patients. Early detection of the DR can save many patients from permanent blindness. Various artificial intelligent based systems have been proposed and they outperform human analysis in accurate detection of the DR. In most of the traditional deep learning models, the cross-entropy is used as a common loss function in a single stage end-to-end training method. However, it has been recently identified that this loss function has some limitations such as poor margin leading to false results, sensitive to noisy data and hyperparameter variations. To overcome these issues, supervised contrastive learning (SCL) has been introduced. In this study, SCL method, a two-stage training method with supervised contrastive loss function was proposed for the first time to the best of authors' knowledge to identify the DR and its severity stages from fundus images (FIs) using "APTOS 2019 Blindness Detection" dataset. "Messidor-2" dataset was also used to conduct experiments for further validating the model's performance. Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied for enhancing the image quality and the pre-trained Xception CNN model was deployed as the encoder with transfer learning. To interpret the SCL of the model, t-SNE method was used to visualize the embedding space (unit hyper sphere) composed of 128 D space into a 2 D space. The proposed model achieved a test accuracy of 98.36%, and AUC score of 98.50% to identify the DR (Binary classification) and a test accuracy of 84.364%, and AUC score of 93.819% for five stages grading with the APTOS 2019 dataset. Other evaluation metrics (precision, recall, F1-score) were also determined with APTOS 2019 as well as with Messidor-2 for analyzing the performance of the proposed model. It was also concluded that the proposed method achieved better performance in detecting the DR compared to the conventional CNN without SCL and other state-of-the-art methods.

1. Introduction

In 2019, the International Diabetes Federation (IDF) announced that over 460 million people aged between 20 and 79 suffered from diabetes around the world [1]. According to their statistical measure, the number of affected people is expected to reach 700 million by 2045. Diabetic Retinopathy (DR), vision impairment, heart attacks, renal failure, and stroke are the serious health issues associated with the diabetes. DR, a frequent diabetes consequence, happens when the blood vessels in the retina are damaged by high blood sugar levels, causing swelling and leakage [2]. In the fundus retina image, lesions are appeared as leaking blood and fluids. Red and bright lesions are the types that can be commonly identified during diagnosis of the DR. Microaneurysms (MA) and hemorrhage (HM) are involved in the red lesions, whereas soft and hard exudates (EX) are involved in the bright lesions. MA refers to the small dark red dots, whereas HM refers to the larger spots. Soft EX shows as yellowish-white and fluffy dots caused by nerve fiber injury, whereas hard EX appears as distinctive yellow spots [3]. Fig. 1 shows a Fundus Image (FI) with various lesions for DR identification [4]. A person's eyesight may be entirely lost at the severe stage of the DR. The global

* Corresponding author.
E-mail address: marcin.kowalski@wat.edu.pl (M. Kowalski).

https://doi.org/10.1016/j.compbiomed.2022.105602
Received 25 January 2022; Received in revised form 26 April 2022; Accepted 6 May 2022
Available online 10 May 2022

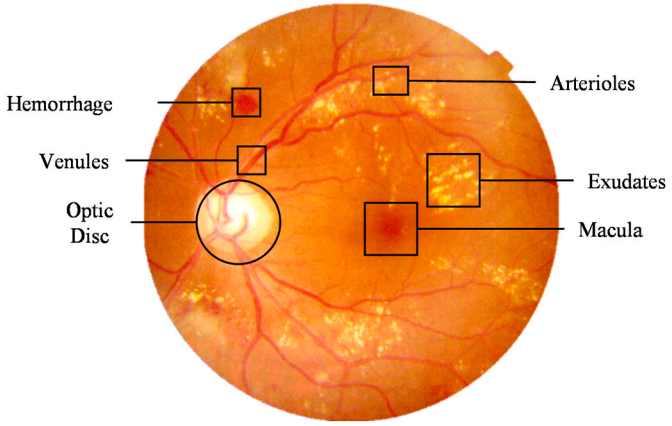

Fig. 1. Fundus image with various lesions for DR classification.

prevalence of the DR-related blindness is estimated to be approximately 2.6% [5]. The risk of blindness can be minimized if the DR is detected at an early stage.

Various imaging modalities have been developed so far for diagnosing retinal related diseases such as DR, diabetic macular edema (DME), glaucoma, and choroidal neovascularization (CNV). Optical coherence tomography (OCT) is a non-invasive retinal imaging technology that is used to obtain high-resolution micro-meter level cross-sectional images of the retina useful for assessing structural changes associated with the DR [6]. OCT images provide the retinal thickness and tortuosity through the morphology and reflectivity of retinal layers and DME is the thickening of macula, the central part of the retina [7]. OCT is more sensitive to small changes in the retinal thickness and therefore mainly used for the DME diagnosis. Optical coherence tomography angiography (OCTA) imaging technology can detect and illustrate movement in ocular structures and choroidal vasculature in the posterior segmentation of the eye [6]. OCTA generates the volumetric view of blood vessels [8] by utilizing sequential OCT scans. Fundus photography (FP) is another retinal imaging modality used for clinical studies to grade and monitor the severity progression of DR over time [9]. Though the OCT images capture the thickness changes, the OCTA captures volumetric blood vessels and can be used for classifying images as normal or DR. However, for grading its different severity stages (five stages) various lesions like microaneurysm, hemorrhages, exudates etc. Available in FP are used. Therefore, the OCT and OCTA imaging modalities are basically used for DME whereas grading severity levels is assessed by the FP imaging modality. In this study, the severity of the DR was classified using the images obtained from the FP imaging modality.

However, manual examination of Optical Coherence Tomography (OCT) or color FIs of the retina is conducted in traditional methods for detecting the presence of DR and this requires experienced and professional ophthalmologists due to the use of sophisticated grading systems during the DR diagnosis. Furthermore, there is a high probability of misdiagnosis during the manual examination, and it is time-consuming and expensive.

Since it is necessary to detect the DR correctly at an early stage for reducing the chances of blindness, in the last decade many researchers have proposed several computer-aided intelligent diagnosis systems. Many researchers have developed several machine learning and deep learning algorithms for the automatic detection of DR. However, still there is a huge scope of improvement in the case of detecting the DR accurately using image analysis by machine learning (ML). So far various works have been carried out on the DR detection using end-to-end one stage training. Here, a two-stage training approach was applied with SCL method. To the author's best knowledge, so far, the classic cross-entropy loss function was used in most of the cases for detecting the DR severity. However, more recently some limitations of the cross-entropy loss function have been identified. For instance, it is sensitive to noisy data and hyperparameter variations, and it also provides poor margin which produces erroneous results if the inputs vary slightly from the training data [10,11]. To overcome these issues, a novel loss function called supervised contrastive loss function (SupCon) was used in SCL and this shows significant improvement in obtaining model accuracy and robustness. Now-a-days, SupCon found widespread applications in computer vision research. As the images are collected from different sources, most of the cases they are noisy and five stages labelling by human effort is also erroneous.

Therefore, SCL, a very simple method, was applied for the DR grading from the FIs. Xception deep learning model was employed as the encoder. CLAHE was applied for preprocessing the FIs for improving the quality of the images. A binary classification was conducted to identify whether a FI is DR or not along with a multiclass classification for further detection of five stages of the DR.

The novel contributions of this work are as follows.

1) A two-stage training method (SCL) was proposed for the first time to detect the DR and its severity levels.
2) Xception CNN model was used as the encoder for representation learning.
3) CLAHE was applied for enhancement of the image quality.
4) t-SNE method was followed to visualize the embedding space learned by the SCL.
5) Earlier stages were detected with a greater accuracy before it progresses to the severe and PDR stages.

The rest of the paper is organized as follows: Section 2 describes the previous work related to the DR while Section 3 presents the datasets that have been considered in this study. The proposed framework for conducting this research is described in Section 4 and the results are presented and analyzed in Section 5. Finally, Section 6 presents the key conclusions and recommendations for future work.

2. Literature review

In the past two decades, several research works have been proposed for the automatic detection of the DR through image analysis by ML. Approximately 425 articles have been published in prestigious journals in the past 19 years presenting different strategies for detecting the DR [12]. In this section, a few of the recent works have been reviewed briefly.

Traditional ML approaches such as, support vector machine (SVM), Random forests (RF), neural network (NN), naive Bayes (NB), multilayer perceptron (MLP) and extreme learning machine (ELM) were used in some of the research works for identification of the DR. Features were retrieved using image processing techniques like HOG features, texture features, Wavelet features etc., and employed to the traditional machine learning or neural network-based classifiers. Ramasamy et al. [13] extracted features based on textural gray-level features like co-occurrence, run-length matrix, as well as the coefficients of the Ridgelet Transform and fused them. Finally Sequential Minimal Optimization (SMO) classifier was used to classify DR images. They achieved 98.87% sensitivity, 95.24% specificity, 97.05% accuracy on DIARETDB1 dataset, and 90.9% sensitivity, 91.0% specificity, 91.0% accuracy on KAGGLE dataset. Asha et al. [14] utilized three ML models NB, MLP, and ELM for classifying the DR from the FIs achieved an accuracy of 90% using ELM that outperformed the other models. Ali et al. [15] introduced a novel clustering-based region growing framework and utilized different types of ML algorithms for detecting the DR. It was calculated that 245 pieces of hybrid feature were obtained from each FI than 13 features were selected using four feature selection methods. A custom dataset 2500 FIs were used in the investigation and obtained an accuracy of 99.73% using the ML algorithm. Gayathri et al. [16]

proposed a lightweight CNN model for detecting the DR from the FIs. First, the features from the images were extracted using a six-layer CNN model. Five ML algorithms (SVM, RF, AdaBoost, Naïve Bayes, J48) were used for classifying the images. In several cases, their proposed strategy produced near-perfect categorization findings. Sikder et al. [17] proposed a decision tree-based ensemble learning algorithm for detecting five stages of the DR. They used gray-level intensity on FIs and extracted features using a genetic algorithm (GA) and achieved accuracy and F-measure of 94.20% and 93.51% respectively. Chetoui et al. [18] introduced various texture features for DR and used SVM as classifier. They obtained an accuracy of 0.904, and AUC score of 0.931 with SVM-RBF kernel. Huda et al. [19] applied classification algorithms on several features like Optic disk, microaneurysms, exudates, hemorrhages of DIARET-DB dataset having region based lesion information and final decision was made using SVM, Decision Tree (DT), and Logistic Regression algorithms which achieved 88% accuracy.

Though ML algorithms achieved a favorable result but the features extraction with image processing techniques needs extra effort. Recently, deep CNN models have showed greater success in computer vision, bio-informatics. Therefore, various works have been reported based on deep CNN models to detect the DR from the FIs. Transfer learning approach was used in some works to adapt with the relatively small size DR Datasets. Liu et al. [20] utilized transfer learning approach trained with 'imagenet' models EfficientNetB4, EfficientNetB5, NASNetLarge, Xception, and InceptionResNetV2 for predicting the DR from the EyePACS dataset. An enhanced cross-entropy loss function and three hybrid model structures for the classification of the DR were developed and achieved accuracy and sensitivity of 86.34% and 98.77% respectively. Sheikh et al. [21] applied four transfer learning algorithms which were VGG16, ResNet50, InceptionV3, and DenseNet121 for identifying the DR from the FIs. Better prediction performance was obtained using the DenseNet121 model. A deep convolutional neural network was proposed by Xu et al. [22] obtained an accuracy of 94.5% for automatic DR classification. They used various augmentation to reduce the overfitting problem of small dataset. Gangwar and Rav [23] proposed a hybrid model where a custom block of convolutional neural network (CNN) was accumulated on top of pre-trained Inception-ResNet-v2. For training these hybrid models, two Kaggle datasets were employed: Messidor-1 and the APTOS 2019. They achieved 72.33% and 82.18% test accuracy for the Messidor-1 and APTOS 2019 datasets, respectively. Hemanth et al. [24] presented a DR detection and classification method based on a CNN. They used both HE and CLAHE for image contrast enhancement and obtained a classification accuracy of 97% and F-measure of 94% using CNN model. Das et al. [25] proposed a novel CNN for classifying normal and abnormal patients using the FIs. The blood vessels were extracted from the images using maximal principal curvature method. To enhance and eliminate falsely segmented regions adaptive histogram equalization (AHE) and morphological opening were applied. DIARETDB1 dataset was considered and attained an accuracy and precision of 98.7% and 97.2% respectively. Pires et al. [26] gradually build a bigger CNN model, performed different types of augmentation, and multi-resolution training using APTOS 2019 dataset and the tested model using the Messidor-2 dataset achieved an area under the receiver operating characteristic (ROC) curve of 98.2%. Liu et al. [27] designed a new model where multiple weighted paths CNN, named WP-CNN, for detecting the DR. Three models were designed including WP-CNN-32, WP-CNN-52, and WP-CNN-105 that consist of 32, 52, and 105 convolutional layers respectively. A high prediction accuracy of 94.23%, sensitivity of 90.94%, and specificity of 95.74% were obtained. Math et al. [28] introduced a segment-based learning method for predicting the DR. They adapted a segment-level DR estimation using a pre-trained CNN and merged all the segment levels for classification, which obtained an area under the ROC curve of 0.963. Zeng et al. [29] proposed a Siamese-like binocular CNN model and obtained an AUC score of 0.951 which is 0.011 higher than existing monocular model for detecting the DR automatically. During the training phase, 3062 DR images were utilized and conducted external validation afterward to obtain sensitivity and specificity values above 97%.

Also, various ensemble learning methods was used for DR classification with multiple classifiers rather than a single classifier. Zhang et al. [30] proposed an automated DR identification and grading system named DeepDR for determining the prevalence and severity of the DR using the FIs. Ensemble learning based method was utilized with Inception V3, Xception and InceptionResNetV2 CNN models and achieved an area under the curve of 97.7%, a sensitivity of 97.5%, and a specificity of 97.7%. Kaushik et al. [31] presented a stacked model with three CNN models and achieved an accuracy of 97.92% for binary classification and 87.45% for multi-class classification on EyePACS dataset.

Besides image-level grading for DR classification, various segmentation tasks were conducted for segmenting and localizing various lesions information like blood vessels, microaneurysms, hemorrhages, exudates etc. Maqsood et al. [32] proposed a new 3D CNN model for localizing the early sign of DR called hemorrhages and a pre-trained vgg19 model was used for extracting features from the segmented hemorrhages. 1509 images from HRF, DRIVE, STARE, MESSIDOR, DIARETDB0, and DIARETDB1 databases were used for the experiments and achieved an average accuracy of 97.71%. Xu et al. [33] presented an enhanced U-Net named FFU-net for segmenting lesions of DR. IDRiD dataset was used in this work and achieved 11.97% sensitivity, 10.68% IoU, 5.79% Dice score. Hasan et al. [34] proposed an end-to-end encoder-decoder network named DRNet for the segmentation and localization of optical disk (OD) and fovea centers. For OD segmentation, they achieved mIoU score of 0.845, 0.901, 0.933, and 0.920 for IDRiD, RIMONE, DRISHTI-GS, and DRIVE, respectively. Nazir et al. [35] proposed a Faster-RCNN based model to segment lesions like hard exudates, soft exudates, microaneurysms, and hemorrhages using the diaretdb1 dataset and achieved accuracy of 0.95 and Intersection over union (IOU) of 0.94. Sambyal et al. [36] proposed a modified U-Net with residual network for the segmentation microaneurysm and hard exudate lesions. For this, they trained the model on e-optha dataset and validate on IDRiD dataset achieved 99.88% accuracy, 99.85% sensitivity, 99.95% specificity and dice score of 0.9998 for both microaneurysm and exudate segmentation. Zago et al. [37] developed a patch-based deep CNN model for red lesion segmentation and localization. For this Diaretdb1 dataset was used and obtained an AUC score of 0.912.

Various attention-based works, fuzzy-classifier and hybrid works were also done for DR classification. Li et al. [38] developed a cross disease attention network (CANet) for predicting the DR. Two types of attention modules were generated including disease specific module and disease dependent module. Messidor and IDRiD challenge datasets were considered for training the ML models and attained a prediction accuracy of 85.10%. Mahmoud et al. [39] proposed a hybrid inductive machine learning algorithm (HIMLA) for the automatic detection of the DR. They encoded and decoded FIs for improving image quality. Finally, they extracted features and classified using multiple instance learning (MIL), achieved an accuracy of 96.62%. Afrin & Shill [40] extracted features like blood vessels, microaneurysms, exudates using image processing technique. Then measures blood vessels area, microaneurysms count, exudates area from the processed images and fed these features into a knowledge-based fuzzy classifier for classification, achieved an accuracy of 95.63%. Lal et al. [41] developed a framework for preserving the classification with correct labels free from adversarial attack, training and feature fusion. Their defensive model achieved 99% accuracy. Most of the previous research works were focused on binary-classification and saw great success in identifying the DR. Again, complex features were extracted using image processing techniques in traditional machine learning approaches which could lead to complex processing and often poor results. However, there is still huge scope for developing ML models that can improve the DR prediction accuracy without complex operations particularly for the multiclass classification.

3. Dataset description

Various datasets are publicly available for the grading of the DR such as Messidor-1, Messidor-2, EyePack, and APTOS 2019. In this research work, APTOS 2019 dataset was considered and collected from the Kaggle competition [42]. The dataset of FIs was provided by Aravind Eye Hospital in India and contained five stages to detect the severity levels named No DR, mild stage, moderate stage, severe stage, and proliferative diabetic retinopathy (PDR) stage to detect and prevent this disease among people living in rural areas where medical screening is difficult to conduct. Aravind technicians travelled to these rural areas to capture images using high resolution specialized fundus cameras consisting of an intricate microscope attached to a flash enabled camera and then relied on highly trained doctors to review the images and provided diagnosis [42]. The total number of training and test samples in the dataset were 3662 and 1928 respectively. The images were made available with a variety of sizes for instance 2416 × 1736, 819 × 614, and 3216 × 2136. Though 1928 test samples were available but their label annotations were not publicly available and kept hidden for final assessment of the submitted works from the participants since it was a competition dataset. Therefore, the test samples were not considered for this experimental work. Only the training samples of 3662 FIs along with their label annotations were publicly available and they were further split into training (85%) and testing datasets (15%) for the DR detection and severity grading. The dataset was highly imbalanced, as can be seen after conducting the class distribution (Table 1). 3662 samples were used to classify the DR in a supervised manner. Multi-class classification was performed for five stages severity grading of the DR as well as binary classification for the detection of DR. The presence of the DR is often classified as binary mode: normal or DR [43]. Thus, for convenience, the dataset class was relabeled as either normal or DR for the binary classification.

In addition, messidor-2 [44,45] dataset was experimented. Part of the dataset (Messidor-Original) was kindly provided by the Messidor program partners. The remainder (Messidor-Extension) contained never-before-published examinations from Brest University Hospital, France. The data distribution in Messidor-2 is also presented in Table 1.

4. Proposed framework

SCL has been proposed in this study for detecting the severity of the DR. FIs have been preprocessed using Contrastive Limited Adaptive Histogram Equalization (CLAHE) method to enhance the image quality before starting the training phase. The overall proposed framework has been depicted in Fig. 2. All the sub-modules have been described in detail in the following subsections.

4.1. CLAHE based preprocessing

Processing of FIs before analysis is critical for obtaining a better prediction outcome by using ML. Numerous techniques have been developed to enhance the medical imagery suitable for disease detection by the application of ML techniques. CLAHE was employed to enhance the quality of the FIs [46]. It was primarily created to enhance low-contrast medical images; however, it might also be used in more comprehensive applications [47]. It is an alternative implementation of Adaptive Histogram Equalization (AHE). In case of Histogram Equalization (HE) approach, the image is considered as a whole for equalizing. But in CLAHE approach, a complete image was divided into smaller named clips to convert the image into AHE clips [48]. Then, the AHE was applied to each clip individually which limited the amplification in CLAHE by clipping the histogram at a user-defined value termed as clip limit [49]. The clipping level specified how much noise in the histogram should be reduced, increasing the contrast—using a CLAHE color version. The clip limit 2.0 was used with a tile grid size of 8 × 8 in this case. Fig. 3 presents examples of original and processed images by applying CLAHE. After applying CLAHE, all the images were resized to 224 × 224 × 3 to unify them.

From Fig. 3, it can be seen that after applying the CLAHE based preprocessing, the quality of the raw images is significantly improved, for instance the lesions are sharpened in the processed images.

4.2. Self-supervised contrastive representation learning

Before applying the SCL, an understand of its primary origin called self-supervised contrastive learning is essential. The critical factor of a successful machine learning model depends on how well it can learn the representation or features or latent variables from the dataset during the training period. Representation learning learns the hidden mapping from the raw input data to feature vector that can improve the future downstream task like classification. More often, the latent feature vectors are located in the manifold of the lower-dimensional spaces. While many dimensionality reduction methods only convert the input data from a higher dimension to a lower dimension, the representation learning methods learn the internal mapping to generalize the new data points. With the increasing success of deep learning in various fields for instance medical imaging, computer vision, and Natural Language Processing (NLP) can learn the internal mapping and extract features with convolutional layers. Contrastive learning has created a surge recently as a suitable representation learning method, and many works associated with the contrastive learning have been published. While in the discriminative model, mapping is learned by human-generated labels. The generative model reconstructs the input given to it, and contrastive learning learns the representation by comparing the similarities and dissimilarities among samples within the dataset. Contrastive learning is one of the leading approaches used in self-supervised representation learning. The contrastive learning pulls “similar” samples together, and “dissimilar” samples are pushed apart in the embedding space.

4.3. Supervised contrastive representation learning

Since there is no label available in self-supervised contrastive representation learning as a pretext task, a positive pair is often formed by data augmentation from the main “anchor” sample. The negative sample is chosen randomly from the mini batch. In the random sampling process, there is a great chance of generating negatives from the same class label of anchor, producing low representation quality. SCL is built on top

Table 1
Dataset distribution for binary and multiclass classifications (APTOS 2019 and Messidor-2 datasets).

Classification	DR Stage	Number of representative images		Number of training images (85%)		Number of test images (15%)	
		APTOS 2019	Messidor-2	APTOS 2019	Messidor-2	APTOS 2019	Messidor-2
Multiclass	No DR (0)	1805	1017	1534	864	271	153
	Mild (1)	999	270	943	229	56	41
	Moderate (2)	370	347	220	295	150	52
	Severe (3)	295	75	266	64	29	11
	PDR (4)	193	35	149	30	44	5
Binary	No DR/Normal (0)	1805	1017	1533	864	272	153
	DR (1)	1757	727	1479	618	278	109

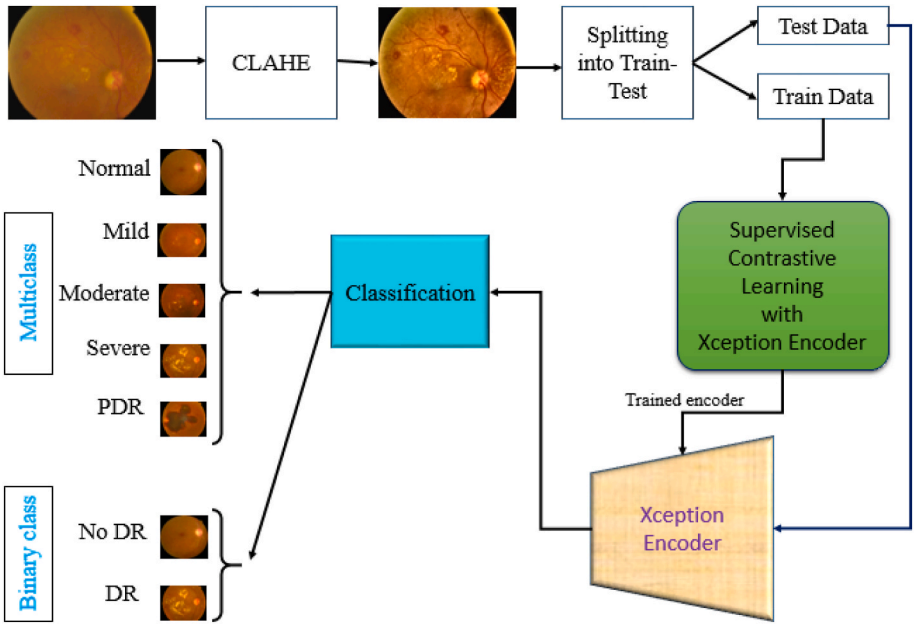

Fig. 2. Proposed Machine Learning framework for Diabetic Retinopathy (DR) identification.

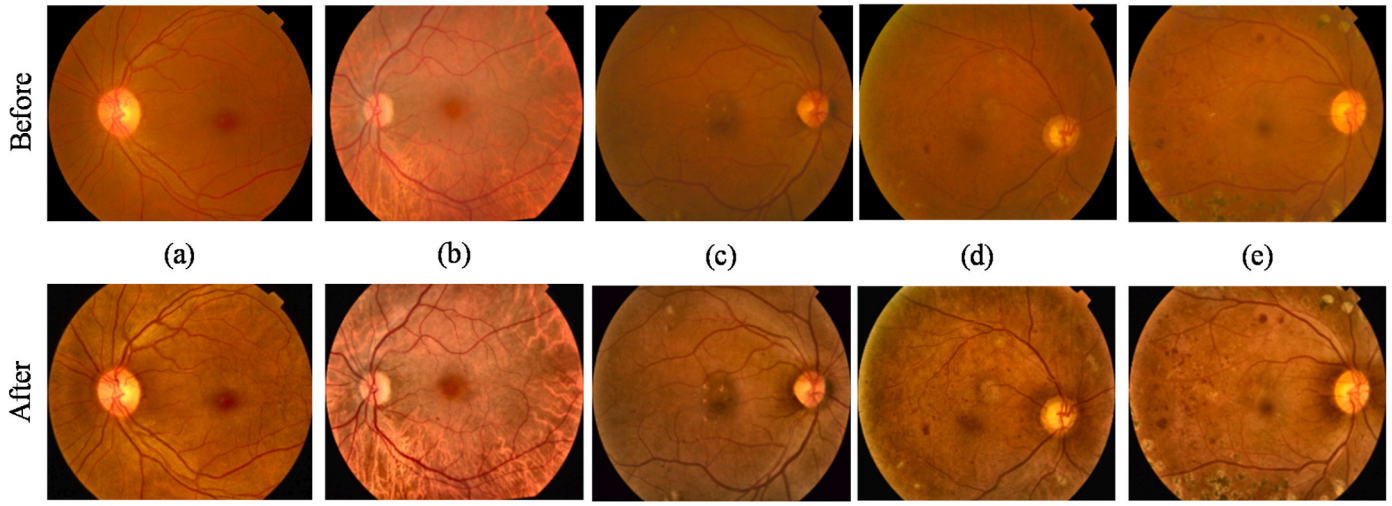

Fig. 3. Original images and corresponding CLAHE processed images (a) No DR, (b) Mild (c) Moderate (d) Severe (e) PDR.

of contrastive self-supervised learning by leveraging the label information of the samples [50]. In this case, the representation feature space is a normalized embedding space where the samples from the same class are pulled closer together, and samples from different classes are pushed apart. Instead of using one single positive sample as in self-supervised contrastive learning, SCL uses many positive samples as well as many negatives per anchor. The positive samples are chosen from the same class instead of data augmentation.

Fig. 4 presents a schematic diagram of self-supervised and SCL. In the self-supervised contrastive learning, one sample highlighted with dotted border is of the same class as the anchor. But in random sampling process, it has fallen into a category as a negative sample in the mini batch. It will degrade the representation learning process. On the other hand, in the SCL, all the samples from the same class as the anchor are considered to be positives. The sample highlighted with dotted border is a positive one for the anchor as it comes from the same class as the anchor.

4.4. Loss functions

Cross-entropy loss is a widely used loss function for training in supervised learning for classification task. Recently some limitations of cross-entropy loss have been identified for example, lack of robustness to noisy labels [11] and possibility of poor margins [10]. Contrastive loss function used in contrastive learning is free from this shortcoming. Firstly, the contrastive loss function used in self- SCL and then this contrastive loss function is adapted in supervised domain. For a batch with N samples given sample/label pairs, $\{x_k, y_k\}$ $k = 1 \ldots .N$, the samples are augmented randomly. Thus, the corresponding batch consists of 2 N pairs which are used for training. Let, $i \in I = \{1 \ldots 2\ N\}$ be the index of an arbitrary augmented sample and let j(i) be the index of the other augmented sample, then the self-supervised contrastive loss function can be defined by Equation (1).

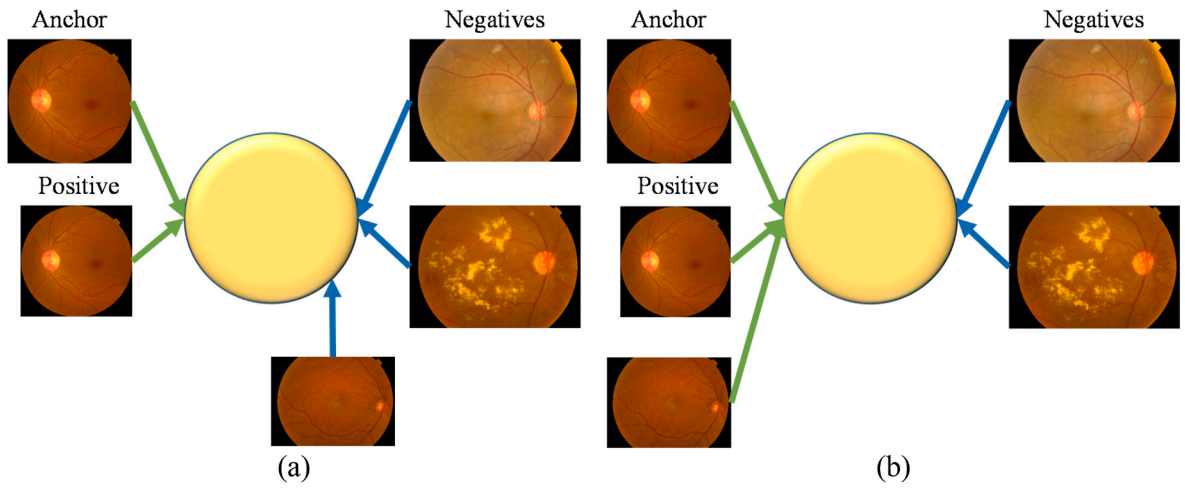

Fig. 4. Representations of (a) self-supervised contrastive learning and (b) supervised contrastive learning.

$$L^{self} = \sum_{i \in I} L_i^{self} = -\sum_{i \in I} \log \frac{\exp\left(\frac{zi \cdot zj(i)}{\tau}\right)}{\sum_{a \in A(i)} \exp\left(\frac{zi \cdot za}{\tau}\right)} \tag{1}$$

where $z_i = H(E(xi)) \in RDp$, represents dot product, $\tau \epsilon R^+$ is temperature parameter, the anchor index is i, positive index is j(i) and the rest 2(N-1) are negative samples.

With the label information, the supervised contrastive loss can be defined by Equation (2).

$$L^{sup} = \sum_{i \in I} L_i^{sup} = \sum_{i \in I} -\log \left\{ \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp\left(\frac{z_i \cdot z_p}{\tau}\right)}{\sum_{a \in A(i)} \exp\left(\frac{z_i \cdot z_a}{\tau}\right)} \right\} \tag{2}$$

4.5. Training strategy

Two stage training was performed where the first phase was representation learning and the second phase was end-to-end classification. Fig. 5 presents the CNN model training strategy. A traditional convolutional neural network (CNN) is presented in Fig. 5(b) where input is given to a convolutional block, then on top of this a classifier is applied to classify five stages of DR for the multiclass classification or identify DR for the binary classification. Fig. 5(b) describes the full training process of the SCL, which has two phases of training as described below in details.

4.5.1. Stage-1: representation learning

Data Augmentation: In the augmentation module Aug (.), each sample image was augmented into two where two views of each sample were observed. Primarily, the batches were created for model training. The batch samples were augmented before deploying into the model. If the primary batch contains N samples having P positive samples and N′ negative samples, then the batch contains 2 N samples containing 2 P positive samples and 2 N' negative samples after conducting augmentation. This augmented batch was then given as an input to the base encoder E (.). For augmentation, a simple strategy [rotation (90°, 270°), vertical flip, and horizontal flip] was used from which one augmentation method was selected randomly for augmentation purpose.

Encoder: A deep CNN model was used as the base encoder E (.) that extracts the features from the input images. Xception CNN pretrained model was used as the base encoder where transfer learning technique has been employed to adapt the 'ImageNet' leaned weights into the DR domain due to having limited datasets [51]. Xception model uses depth-wise separable convolution for reducing computational complexity and memory requirements. Convolutional block of Xception base encoder maps 2048-D features from the given input. In this case, 2048 is the output of the average pool layer of the Xception CNN model.

Projection Head: On top of this convolutional encoder, a projection head H (.) was added, mapping the 2048-D dimensional representation space to 128-D space. To add this projection head, multilayer perceptron (MLP) was used with only one linear layer having 128 nodes and to attach non-linearity, a ReLU(.) function was added. The supervised contrastive loss function was used, and a model was trained to minimize the loss. In this case, 'Adam' the most popular optimizer for deep learning model training was used to update the weights of the model. In a unit hypersphere embedding space, it placed all the positive samples from the same class altogether while pushing negative samples apart. As a similarity measure, a cosine similarity was used in the contrastive loss function.

4.5.2. Stage-2: end-to-end classification

After completing the training phase of representation learning, the encoder learns the representation details for the training samples. The projection head was discarded and kept only the trained Xception encoder.

On top of this encoder, multilayer perceptron (MLP) containing one dense layer was added and trained this end-to-end classifier with a cross-entropy loss function (Equation (3)). During training, the weights of the encoder are frozen and fine-tuned with MLP for the classification of DR.

$$cross-entropy\ loss,\ L = \begin{cases} -(y log(p) + (1-y)\log(1-p);\ Binary \\ -\sum_{c=1}^{M} y_{o,c} log p_{o,c}; Multiclass \end{cases} \tag{3}$$

Algorithm 1. describes the whole working procedure of DR image processing and classification.

Algorithm 1: Proposed framework for DR classification

Requirements	Dataset of fundus images and labels (X_{data}, y_{data}); where y {0(Normal), 1(Mild), 2(Moderate), 3(Severe), 4(PDR)}
Inputs	Training images X_{train} X_{data} and $y_{train}y_{data}$; Test images X_{test} X_{data} and $y_{test}y_{data}$
Output	Trained model with probability score to
Step-1:	*Preprocessing* • Removal of black regions • Apply CLAHE for image enhancement • Resize images into 224×224×3 (RGV images)

Step-2:	*Supervised Contrastive Learning*
	Learning rate=0.001, optimizer = "Adam"
	for epoch 1 to 100 do
	for batch (X_{batch}, y_{batch}) (X_{train} , y_{train}) do
	• Augment images using one random selection from {Rotation [90, 270], vertical flip, horizontal flip}
	• Encode images using Xception encoder and then project images using dense layer to 128-D dimensional embedding space
	• Calculate supervised contrastive loss
	• Update parameters using Adam optimizer

Step-3:	*Classifier*
	Learning rate = 0.001
	Keep only the trained encoder from step-2
	Add a dense layer with 2 neurons for binary class or 5 neurons for multiclass classification
	Freeze the encoder weights
	for epoch 1 to 30 do
	for batch (X_{batch}, y_{batch}) (X_{train}, y_{train}) do
	Predict the probability score
	Calculate cross-entropy loss
	Update parameters using Adam optimizer

Step-4:	*Test* for X_{test} X_{data} do trained model predicts probability scores

5. Results and analysis

The normalized embedding space learned by the SCL method was demonstrated with t-SNE visualization into a lower 2-D space and further evaluation of the proposed method was carried out with various evaluation metrics both for the binary classification and five stages grading. Pytorch Python framework was used, and the experiments were run on Kaggle online platform with GPU support. "Adam" was used as the optimizer, the learning rate was set to 0.001, and the batch size was set to 8. For conducting experiments on the DR, two datasets named APTOS 2019 and Messidor-2 datasets were used.

5.1. Evaluation metrics

When constructing a predictive model, it is critical to use a metric such as confusion matrix to evaluate its success. Accuracy is defined as the percentage of cases that are accurately detected out of all the detected cases (Equation (4)). The method allows seeing how well the algorithm works in classifying [53].

$$Accuracy = \frac{Tp + TN}{Tp + TN + Fp + FN} \tag{4}$$

where T_P = True Positive, T_N = True Negative, F_P = False Positive, F_N = False Negative. Precision as defined by Equation (5) is the most basic to measure the percentage of total positive specimens to total positives [53].

$$Precision = \frac{Tp}{Tp + Fp} \tag{5}$$

An accurate model can identify the majority of True Positives, which is known as recall defined by Equation (6) [53]. Each DR-affected patient must be identified in this investigation.

$$Recall = \frac{Tp}{Tp + FN} \tag{6}$$

F1 score equals the harmonic mean of precision and recall and can be stated by Equation (7) [53].

$$F1 - Score = \frac{2 \times T_p}{2 \times T_p + F_p + F_N} \tag{7}$$

5.2. Experiments on APTOS2019 dataset

5.2.1. Pretrained models- APTOS2019

In order to get the best encoder, various pretrained models named Xception, DenseNet121, ResNet50, and VGG19 were trained to choose the best one. The highest test accuracies obtained with Xception, DenseNet121, ResNet50, and VGG19 were 82.00%, 81.31%, 80.90%, and 49.25% respectively. The variations in training and test accuracies with number of epochs are presented in Fig. 6. The figure clearly

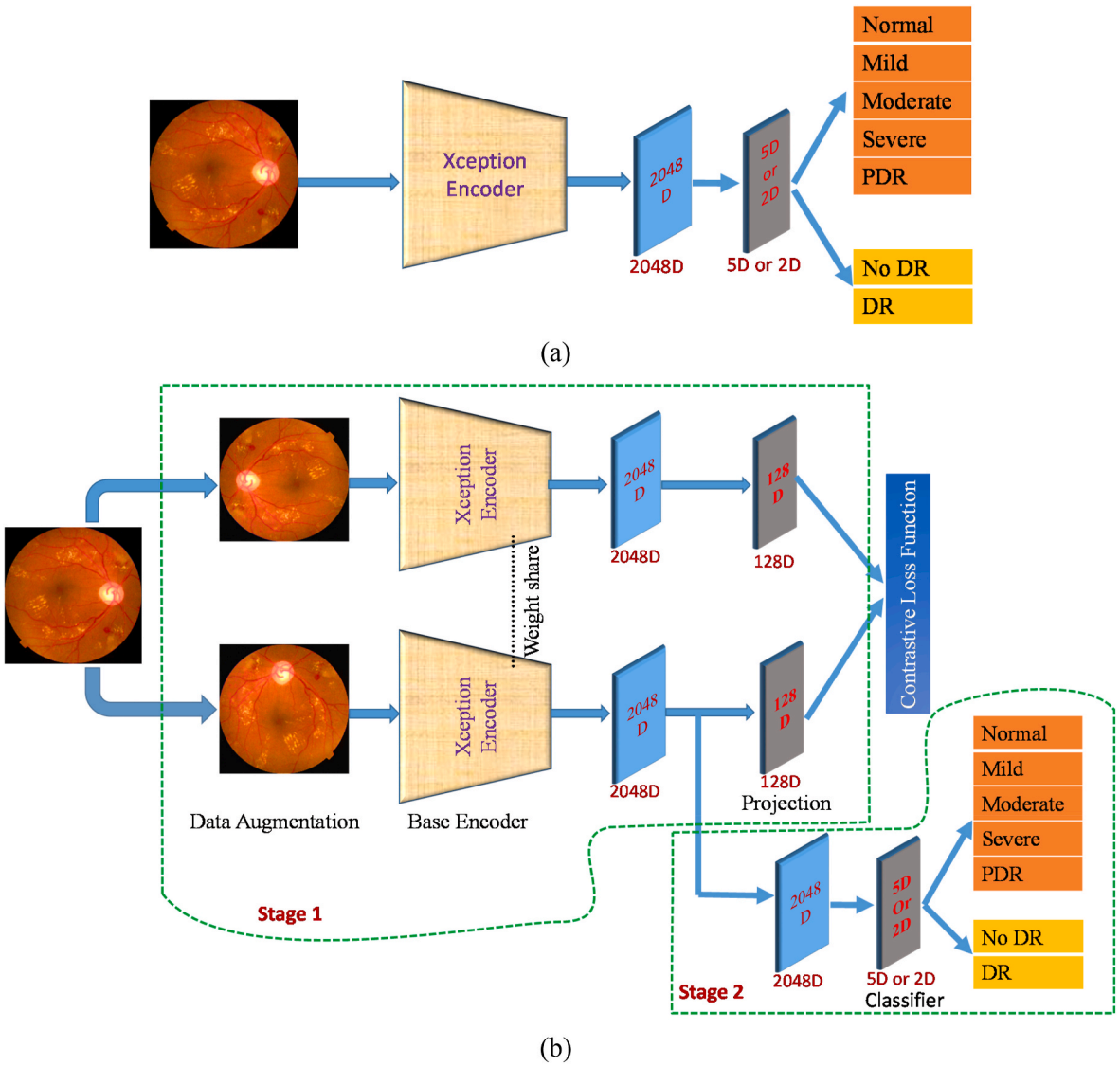

Fig. 5. Model training strategy (a) conventional CNN uses cross-entropy loss (b) supervised contrastive learning uses supervised contrastive loss for representation learning and a second stage classifier on top of the representation learned.

demonstrated that Xception model showed the highest accuracy, and this result justified the selection of the model as an encoder of the proposed two-stage training method called SCL.

5.2.2. t-SNE visualization

t-distributed stochastic neighbor embedding (t-SNE) is a statistical method that is used not only for nonlinear dimensionality reduction but also for visualization of higher-dimensional data [52]. The data visualization is performed by transforming these higher dimensional data into lower dimensions of two or three. The operation is performed in a way such that the adjacent points are used to model similar objects and remote points are used to model distinct objects with high likelihood.

The t-SNE algorithm performed the operation in two steps. Firstly, a probability distribution is created across two high-dimensional objects in this manner that a higher probability is attributed to similar objects, while a lower probability is allocated to dissimilar points. Secondly, in the lower dimensional space, it creates an equivalent probability distribution over the points, and between the two distributions, the Kullback–Leibler divergence (KL divergence) is reduced with respect to the locations of the map's points.

By visualizing what is leaned by the trained model, the model's performance can be interpreted. To visualize the representation learning in the embedded space of a trained model, the popular t-SNE method was used. It showed that the training and test dataset samples' embedding space was in a reduced 2D space both for the binary and multiclass classifications. As shown in Fig. 7, the sample points can easily be separated in the embedded space between the normal or DR after the training phase using the trained model as well as the points in multiclass classification.

5.2.3. Margin hyperparameter tuning- APTOS2019

Margin is a hyperparameter used in contrastive loss function as a threshold distance to separate positive and negative samples. Different margin values provide different representation learning. It was suggested in Ref. [50] that using a lower value (greater than 0) of margin, provides a better representation learning. Therefore, this hyperparameter margin value was tuned for binary and multiclass classifications starting from 0.1 and stopping at 0.9, without going further ahead as the higher-margin values provided higher supervised contrastive loss value during the representation learning. From Fig. 8, the supervised contrastive loss for various margin values can be observed. The trained encoder was chosen with a margin value of 0.1 as it showed lower loss with better representation learning for classification.

After the representation learning with SCL, the projection head H (.)

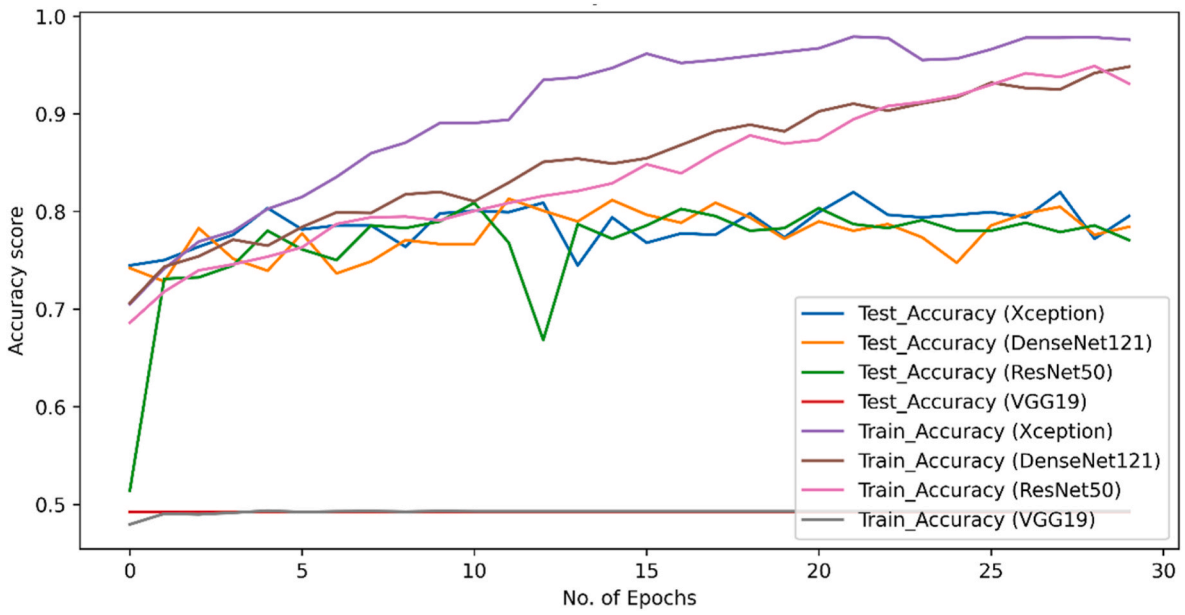

Fig. 6. Training and test accuracy of pretrained models with APTOS2019 dataset for the detection of DR severity.

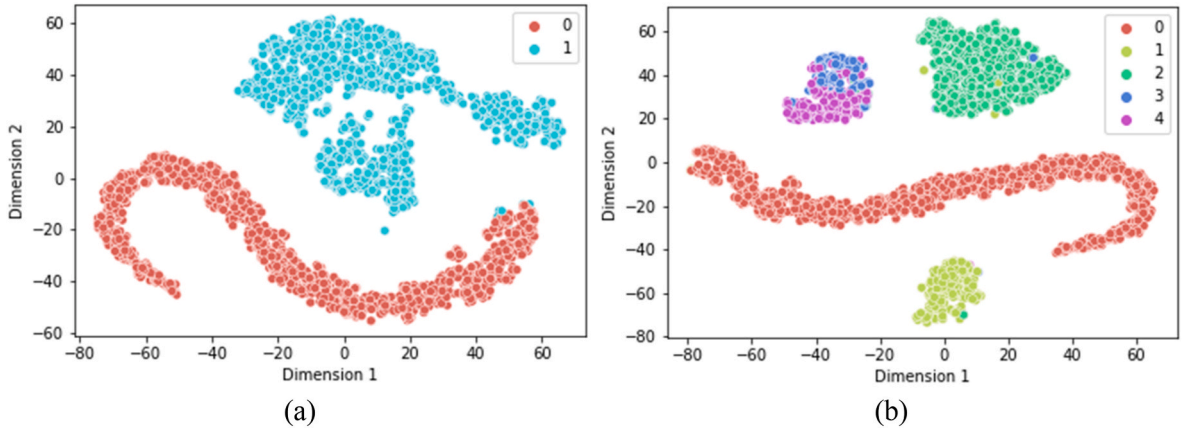

Fig. 7. Embedding space visualization using t-SNE in a 2-D space of training samples after the model training for (a) binary classification (0: Normal and 1: DR) (b) multiclass classification (0: Normal, 1: Mild, 2: Moderate, 3: Severe and 4: PDR).

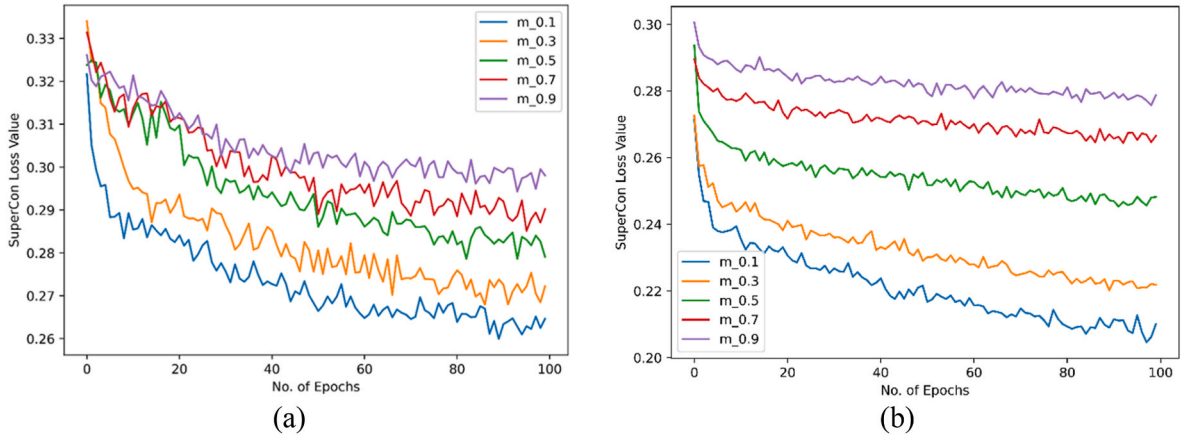

Fig. 8. Supervised contrastive loss during training for (a) binary and (b) multiclass classifications with various margin values with APTOS2019 dataset.

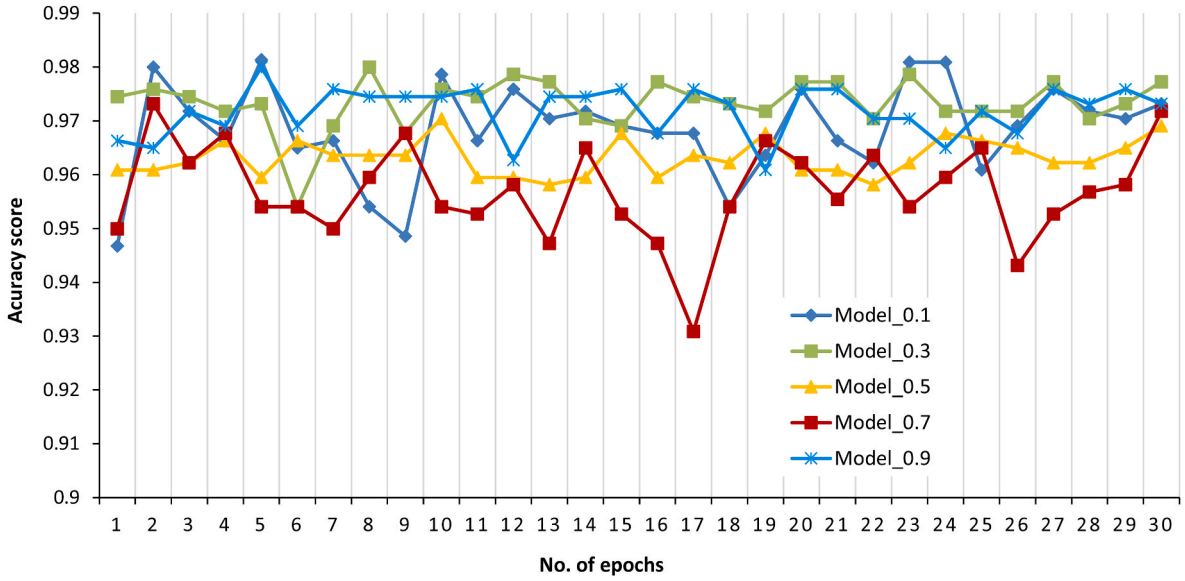

Fig. 9. Test accuracy of trained models for binary classification with various margin values with APTOS2019 dataset.

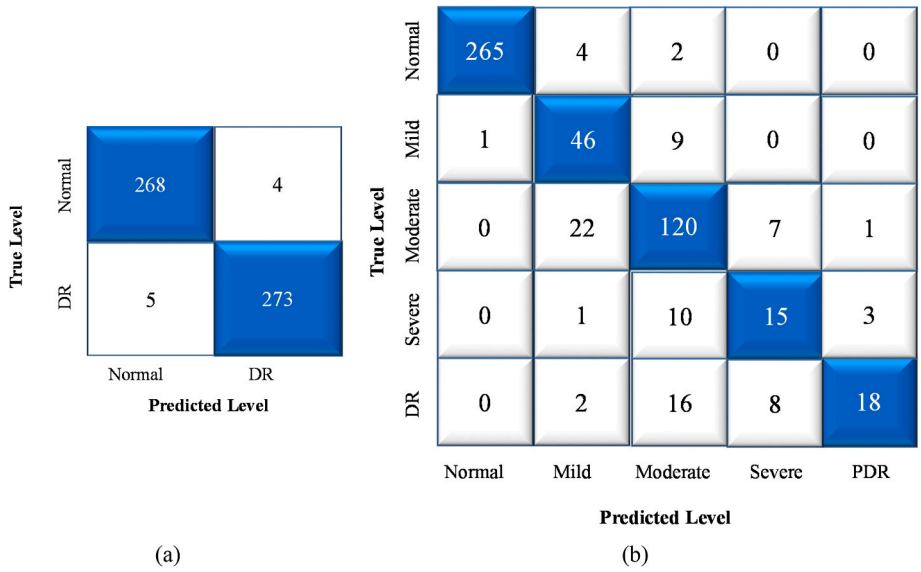

Fig. 10. Confusion matrixes for (a) binary and (b) multiclass classifications with APTOS2019 dataset.

was dropped and only kept the trained encoder E (.). Then one dense layer was added and fine-tuned with the trained encoder E (.) while keeping the weights of the trained encoder frozen. This end-to-end classifier was trained for 30 epochs and the test accuracy results obtained are presented in Fig. 9 for binary classification. The highest test accuracy result obtained was 98.36% from the model with a margin of 0.1 for the APTOS 2019. For the clinical application, only test accuracy does not ensure the performance of a model. For this, other evaluation metrics such as confusion matrix, precision, recall, F1-score are used very often. These evaluation metrics values class-wise and overall scores for both the binary and multiclass classifications were presented only for the best trained encoder chosen with a margin of 0.1. Fig. 10 presents the confusion matrixes (CMs) for both the binary and multiclass classifications.

Table 2 presents the evaluation metrics for the binary and multiclass classification. For the binary classification, the overall precision, recall, and F1-score obtained were 98.37%, 98.36%, and 98.37%, respectively.

Table 2
Class wise performance evaluation for binary and multiclass classifications with APTOS2019.

Evaluation Metrics	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
Binary classification					
Normal	–	98.17	98.53	98.35	–
DR	–	98.56	98.20	98.38	–
Overall	98.36	98.37	98.36	98.37	98.50
Multiclass classification					
No DR	–	99.62	97.79	98.69	–
Mild	–	61.33	82.14	70.23	–
Moderate	–	76.43	80.00	78.18	–
Severe	–	50.00	51.72	50.85	–
PDR	–	81.82	40.91	54.55	–
Overall	84.36	73.84	70.51	70.49	93.82

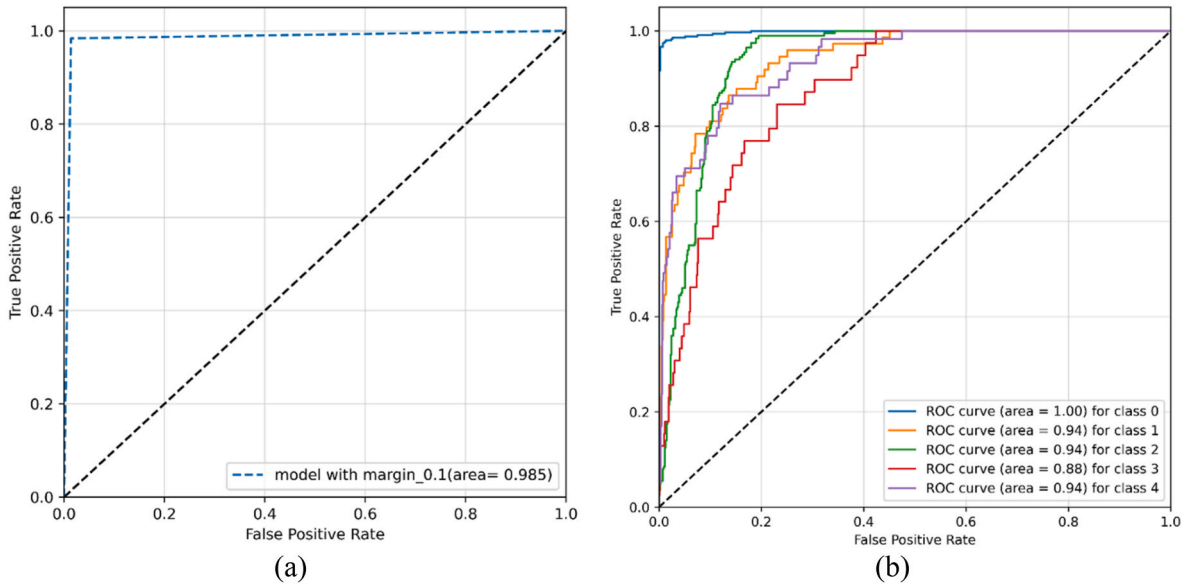

Fig. 11. ROC curves for (a) binary and (b) class wise multiclass classifications (0: Normal, 1: Mild, 2: Moderate, 3: Severe and 4: PDR) with APTOS2019 dataset.

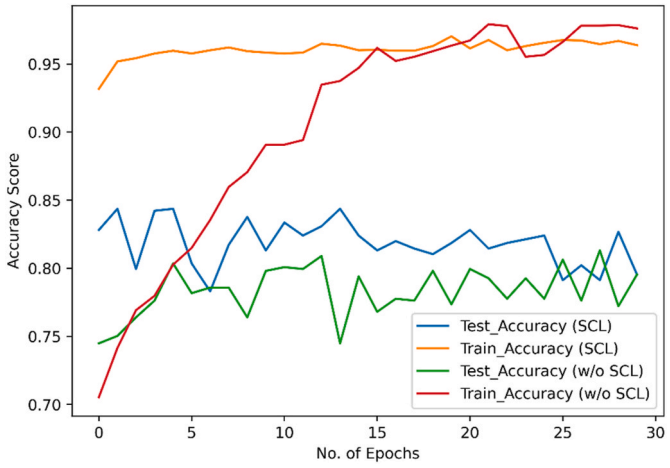

Fig. 12. Training and testing accuracy of SCL method and without SCL end-to-end method for multiclass classification with APTOS2019 dataset.

For the multiclass classification, the overall precision, recall, and F1-score obtained were 73.84%, 70.51%, and 70.49%, respectively.

Also, the effectiveness of the model was demonstrated by the receiver operating characteristics curve (ROC) for the binary and multiclass classifications. Area under the curve (AUC) score provides the capability of a classifier to distinguish among the classes and is a summary of the ROC curve. From Fig. 11(a), it can be seen that the AUC score is 98.50% for the binary classification. Whereas the AUC score was 93.82% for the multiclass classification (Fig. 11(b)). The AUC scores for the normal, mild, moderate, severe, and PDR stages were 100%, 94%, 94%, 88%, and 94%, respectively.

5.2.4. Ablation study- APTOS2019

To show the improvement of using the proposed SCL approach rather than using end-to-end CNN model for the detection of the DR and its severity levels, experimentations were conducted with both the SCL with Xception encoder and end-to-end CNN (Same Xception) model separately. The main difference between the proposed SCL method and conventional CNN method was identified as the two-stage training where the second stage training was the same as like the conventional end-to-end classification. In the first stage of the proposed SCL method, the representation/features of the training dataset were learned in a higher dimensional embedding space with supervised contrastive loss function. In the second stage, the trained encoder was taken and a classifier was added on top of this and fine-tuned like the conventional end-to-end CNN classifier with cross-entropy loss function. From Fig. 12, it was noticed that the testing accuracy of the SCL method was always higher than the end-to-end method with the same Xception model. The results also reflected the effectiveness of using two loss functions (SperCon loss and cross-entropy loss) in two stages of the training in SCL with Xception compared to only the cross-entropy loss function used in end-to-end CNN classifier. For multiclass classification, the proposed SCL method achieved the highest test accuracy of 84.36% which approximately 2% higher than that of the traditional CNN model with the same Xception architecture (82.00%).

From Fig. 13, it was demonstrated that the SCL method showed improved AUC score than the conventional end-to-end model without SCL for every class of the DR.

Again, for the binary cases, similar experiments were conducted to demonstrate superiority of the SCL method. The SCL showed improved the performance for the DR identification using binary classification than the conventional end-to-end deep learning model (Fig. 14) with test accuracy increasing from 98.10% to 98.90%.

Furthermore, other evaluation metrics were obtained from the experiments to demonstrate the superiority of the SCL method over the end-to-end method as shown in Table 3.

5.3. Experiments on Messidor-2 dataset

To demonstrate the superiority of the SCL method over the end-to-end CNN, another dataset named Messidor-2 was also experimented.

5.3.1. Pretrained models- Messidor-2

Similar to the APTOS 2019 dataset, various pretrained models to choose the best encoder for the Messidor-2 dataset. The highest test accuracies obtained with Xception, DenseNet121, ResNet50, and VGG19 were 72.78%, 67.93%, 61.83%, and 58.39% respectively. The variations in training and test accuracies with number of epochs are presented in Fig. 15 and the Xception model producing the highest accuracy justified the selection of the SCL method.

5.3.2. Margin hyperparameter tuning-Messidor-2

The margin hyperparameter was tuned with values starting from 0.1

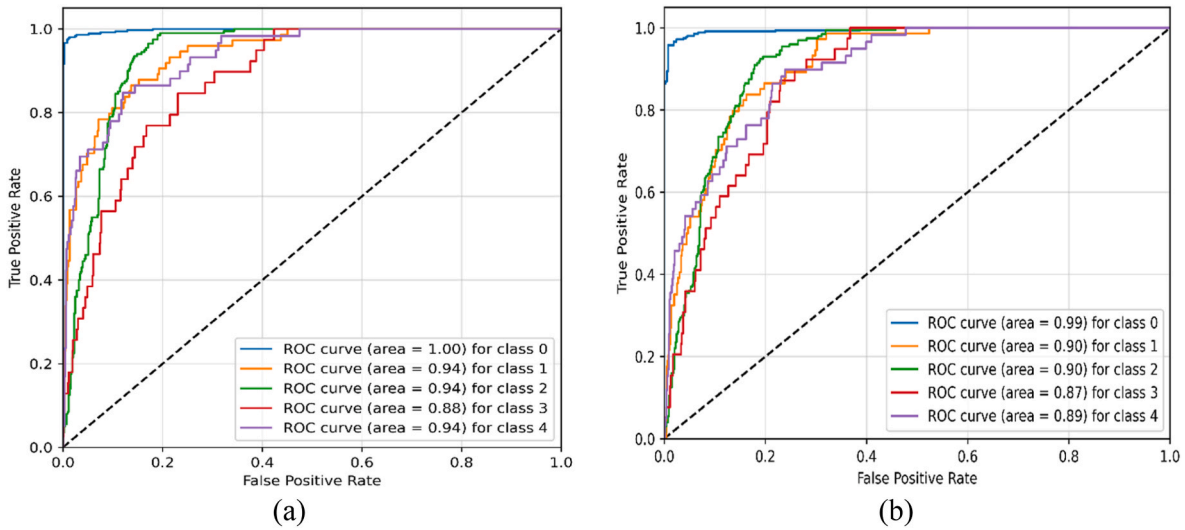

Fig. 13. ROC curves per class of (a) SCL method (b) without SCL end-to-end method for multiclass classification with APTOS2019 dataset.

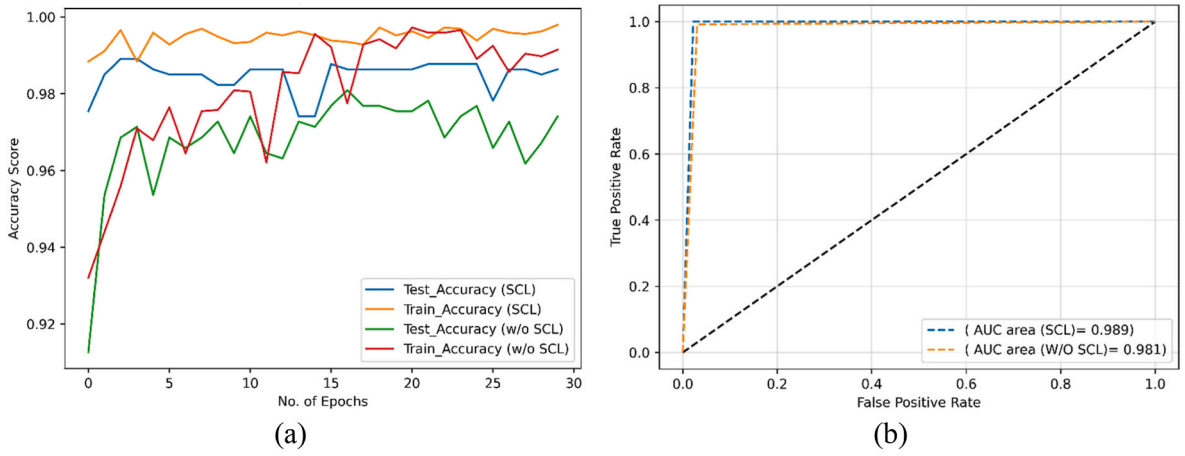

Fig. 14. (a) Training and testing accuracy and (b) ROC curves of SCL method and without SCL end-to-end method for binary classification with APTOS2019 dataset.

Table 3
Overall performance evaluation of SCL and end-to-end methods for multi-class classification with APTOS2019.

Evaluation Metrics	Multiclass		Binary class	
	SCL method	End-to-end method	SCL method	End-to-end method
Accuracy (%)	84.36	82.00	98.90	98.10
Precision (%)	73.84	66.05	100.00	99.15
Recall (%)	70.51	60.79	97.78	96.95
F1-score (%)	70.49	62.40	98.88	98.04
AUC (%)	93.82	90.24	98.89	98.10

and stopping at 0.9, without going further ahead as the higher-margin values provided higher supervised contrastive loss value during the representation learning as demonstrated in Fig. 16. From there, the trained encoder was chosen with a margin value of 0.3 as it showed lower loss with better representation learning for the second phase end-to-end classification.

Other evaluation metrics such as confusion matrix (CM) were demonstrated both for binary and multi-class classification in Fig. 17.

Table 4 presents the evaluation metrics for the binary and multiclass classifications. For the binary classification, the overall precision, recall, F1-score and AUC score obtained were 77.64%, 93.63%, 84.89%, and 84.60% respectively using the proposed SCL method. For the multiclass classification, the overall precision, recall, F1-score, and AUC score obtained were 52.05%, 63.08%, 55.18%, and 87.26% respectively using the proposed SCL method.

The receiver operating characteristics curve (ROC) for the binary and multiclass classifications further proved the model's effectiveness. The overall AUC scores for the binary and multiclass classifications were 84.60% and 87.26% respectively as shown in Fig. 18. The class wise AUC scores obtained were 87%, 75%, 88%, 97%, and 89% for the normal, mild, moderate, severe, and PDR phases, respectively.

5.3.3. Ablation study-messidor-2

Same as the previous ablation results with the APTOS2019 dataset, the testing accuracy of SCL method was always higher than the end-to-end method with the same Xception model (Fig. 19). The proposed SCL method achieved the highest test accuracy of 74.21%, which was a 2% improvement over the traditional CNN model (72.80%) during multi-class classification. It should be noted that in case of the traditional CNN model without SCL, though the training accuracy showed substantial improvement over the SCL, the test accuracy did not show the similar improvement possibly due an overfitting situation. However, in the SCL method, no such overfitting was noticed.

The SCL method also showed improved AUC score than the conventional end-to-end model for almost every class of the DR (Fig. 20).

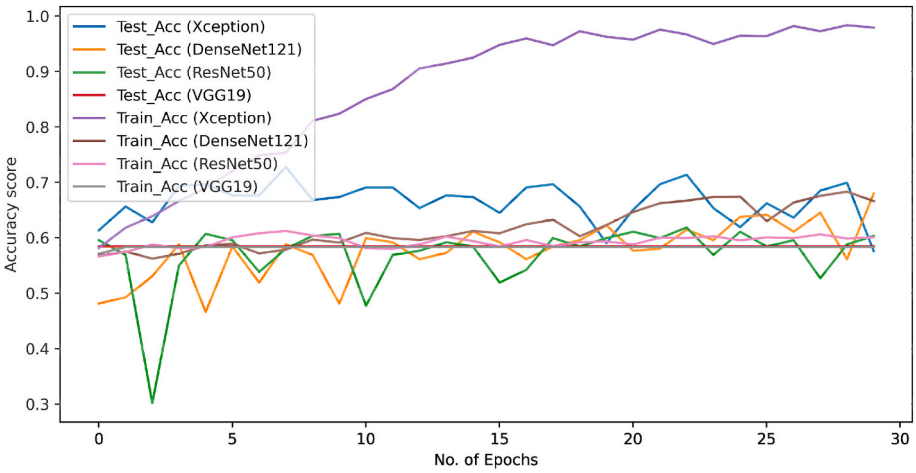

Fig. 15. Training and test accuracy of pretrained models with Messidor-2 dataset for the detection of DR severity.

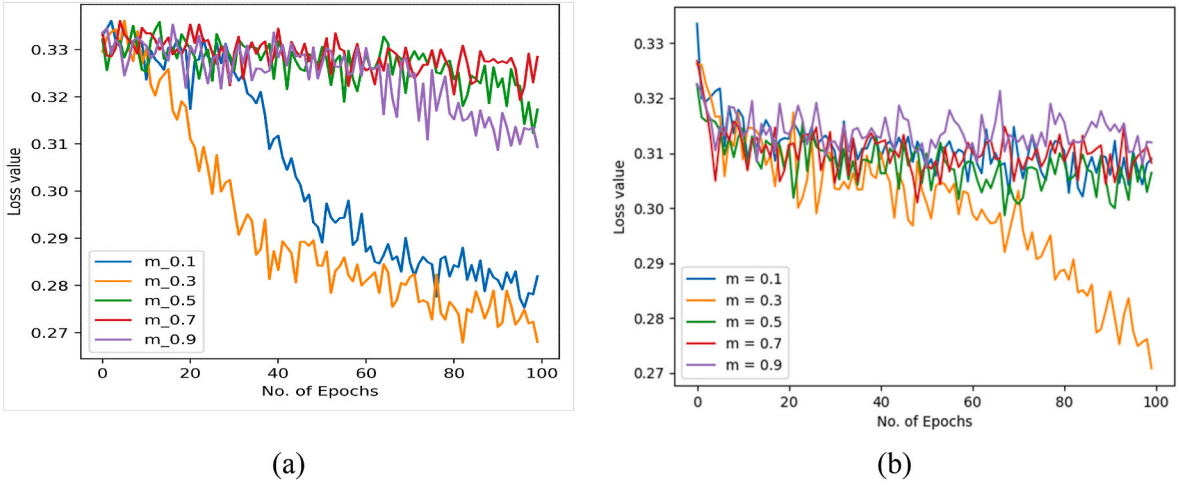

Fig. 16. Training SuperCon loss for (a) binary and (b) multiclass classifications with various margin values with Messidor-2 dataset.

Fig. 21 also demonstrated similar results for the binary classification.

Other evaluation metrics (Recall and f1-score) also showed similar trend (Table 5).

5.4. Discussion

In an end-to-end CNN model training, the model learns the hidden features for classification with cross-entropy loss function using one-stage training approach. However, in the SCL method, firstly the hidden representations/features are learned with an encoder along with a higher dimensional projection head on top of it using contrastive loss function in the first stage of training in order to measure the contrast among the classes. After that the trained encoder is fine-tuned with a classifier layer in the second stage of training.

The SCL method was experimented on both APTOS19 and Messidor-2 datasets. It outperformed for binary classification and moderately well for multiclass classification, owing to the complexity of the DR datasets, which even an expert could not decipher all the time. The hyper-parameter margin value was tuned and achieved the best trained model with lower SuperCon loss of margin 0.1 and 0.3 on APTOS2019 and Messidor-2 dataset respectively. Using both the datasets, test accuracy, precision, recall and F1-score were shown for both the binary and multiclass classifications and the proposed model showed greater AUC scores. Also, to show the superiority of the SCL method over the end-to-end CNN model, an ablation study was demonstrated. It was shown that the SCL method outperformed the standard end-to-end deep learning model in terms of DR grading performance. In the case of multiclass classification with APTOS 2019 dataset, the proposed model achieved higher precision, recall, and F1-score for the earlier stages, but lower values were attained for the severe and PDR stages possibly because of the corresponding small numbers of samples in the datasets. As the early diagnosis can help patients to get recovery than the later stages, therefore high precision, recall, and F1-score in the earlier stages would demonstrate the robustness of the model. Another perspective was that the proposed model achieved a high AUC score indicating that it could separate the classes more accurately.

From all the experimental results, it was clear that the SCL method developed with two-stage of training and APTOS 2019 dataset showed incremental performance improvement than the conventional CNN model with one-stage of training. However, the results with Messidor-2 dataset showed relatively poorer performance than that with the APTOS 2019 dataset. Firstly, in the Messidor-2 dataset, a third party provided the grading annotation, which might be erroneous or a different grading annotation process could be responsible for the inferior performance. Deep learning model generally contains a huge number of parameters and data hungry. Therefore, to train a deep learning model from scratch with only a small sized dataset is challenging. Number of parameters of the proposed model (22.8 M) was too high to be trained by only 3113 images (training set) in the case of APTOS 2019 dataset and 1486 images (training set) in the case of Messidor-2 dataset). The total number of

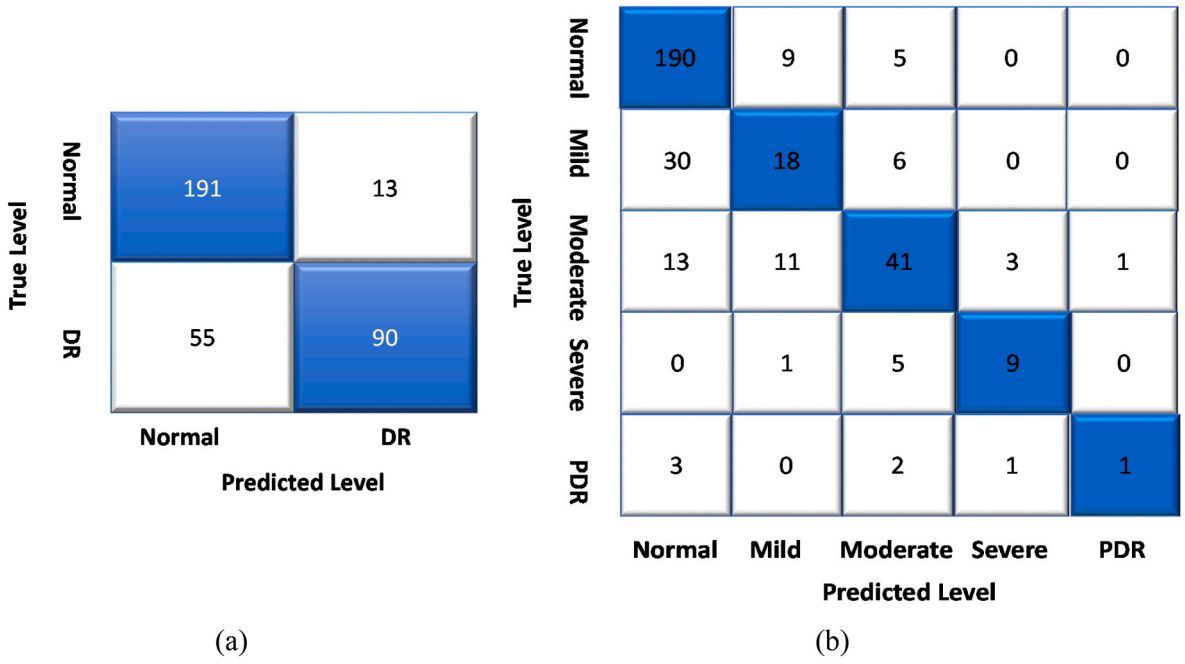

Fig. 17. Confusion matrixes for (a) binary and (b) multiclass classification with Messidor-2 dataset.

Table 4
Class wise performance evaluation for binary and multiclass classifications with Messidor-2.

Evaluation Metrics	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
Binary classification					
Normal	–	77.64	93.63	85.01	–
DR	–	87.38	62.07	73.12	–
Overall	80.52	77.64	93.63	84.89	84.60
Multiclass classification					
No DR	–	93.13	80.51	86.03	–
Mild	–	33.33	46.15	39.21	–
Moderate	–	59.42	69.49	64.41	–
Severe	–	60.00	69.23	64.13	–
PDR	–	14.29	50.00	22.10	–
Overall	74.21	52.05	63.08	55.18	87.26

images of Messidor-2 dataset was significantly smaller compared to the APTOS 2019 dataset. From the distribution of the data in both the datasets, it was observed that most of the images were healthy images. The proportion of the images associated with later stages of the DR in the Messidor-2 dataset were much lower more specifically nearly half of that in APTOS 2019 dataset (Fig. 22). Therefore, for the later stages of the DR (Severe and PDR), the model showed poorer performance particularly for the Messidor-2 dataset.

Some limitations of our works include imbalanced datasets, error in grading and resource restrictions. Both the datasets (APTOS 2019 and Messidor-2) contained highly imbalanced FIs for different categories with significant bias towards the healthy images. From the official documentation of APTOS 2019, it was stated that the grading annotations contained errors and Messidor-2 dataset grading annotation was provided by a third party, hence, no guarantee of correctness could be ensured. Again, because of limitation of the computational resources, the batch size was fixed to 8. Other batch sizes could be experimented to determine their effects.

5.5. Performance comparison with existing work

In traditional machine learning discriminating features are extracted using the image processing techniques for the DR classification purpose.

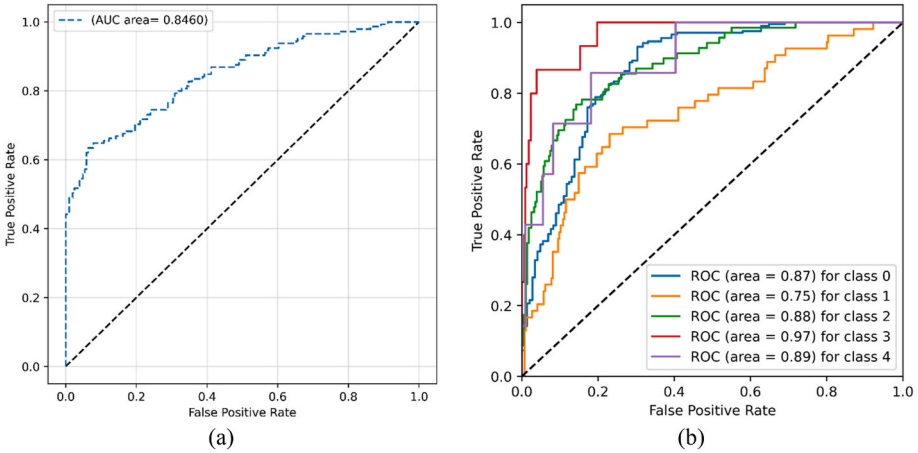

Fig. 18. ROC curve for (a) binary and (b) multi-class classification with Messidor-2 dataset.

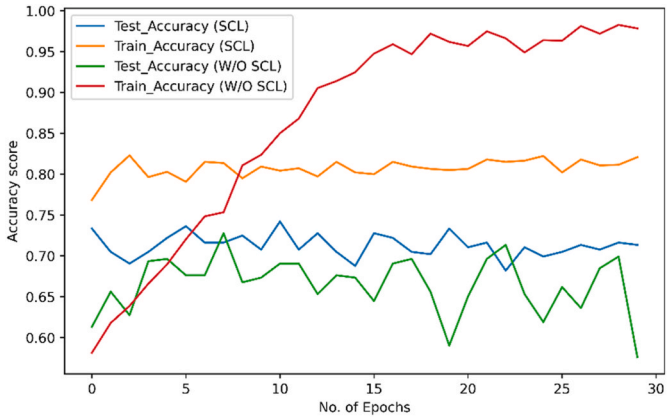

Fig. 19. Training and testing accuracy of SCL method and without SCL end-to-end method for multiclass classification with Messidor-2 dataset.

As the features are very complex and closer from one stage to another, it provides poor results most often. Again, most of the previous works fail to detect the earlier stages accurately which is mandatory to give the patients a chance to recover before it reaches to the later stages [54].

Furthermore, all the previous works done so far used the conventional cross-entropy loss function, which has some limitations. Therefore, a SCL method was proposed for the DR identification and its five stages grading using the publicly available APTOS 2019 dataset and supervised contrastive loss function. The performance of the proposed model for the DR classification from the processed FIs has been evaluated and compared with that of several existing models in this section of the paper. The APTOS-2019 blindness detection dataset has been used to conduct the comparisons.

With the Kaggle APTOS-2019 dataset, Bodapati et al. used a gated-attention method with a deep neural network to detect DR [55]. They

Table 5
Overall performance evaluation for multi-class and binary classifications with Messidor-2.

Evaluation Metrics	Multi-class		Binary	
	SCL	End-to-end	SCL	End-to-end
Accuracy (%)	74.21	72.80	80.52	79.90
Precision (%)	52.05	**61.88**	77.64	**81.60**
Recall (%)	63.08	51.86	93.63	84.80
F1-score (%)	55.18	55.20	84.89	83.17
AUC (%)	87.26	86.00	84.60	83.21

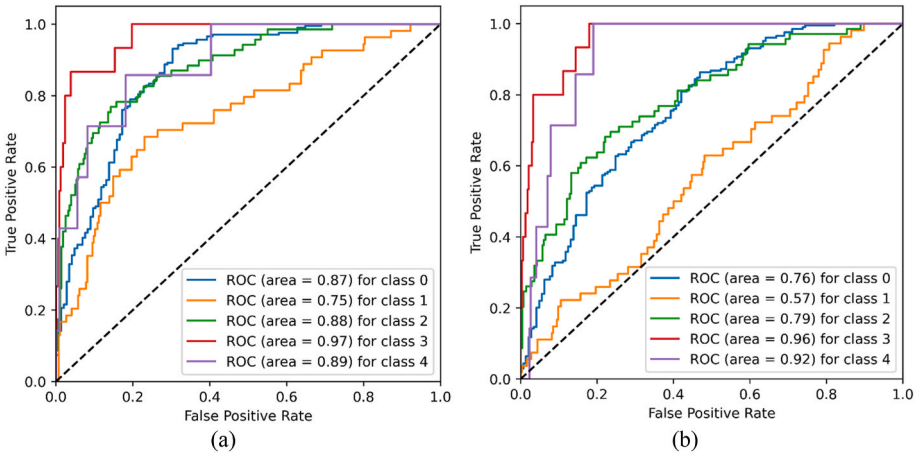

Fig. 20. ROC curve per class of (a) SCL method (b) without SCL end-to-end method for multiclass classification with Messidor-2 dataset.

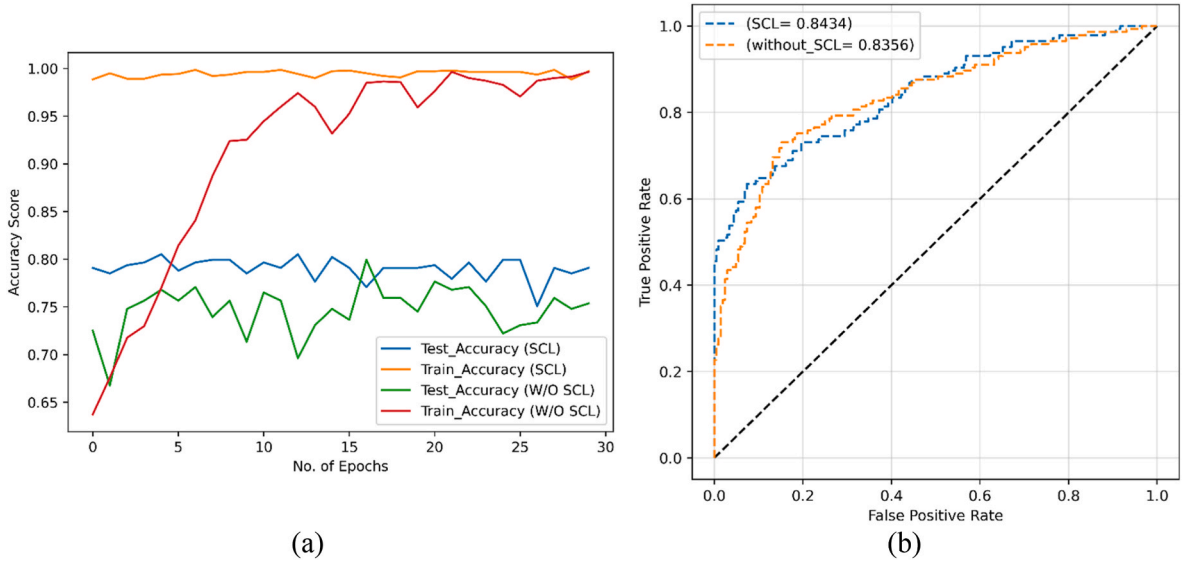

Fig. 21. (a) Training and testing accuracy and (b) ROC curve of SCL method and without SCL end-to-end method for binary classification with Messidor-2 dataset.

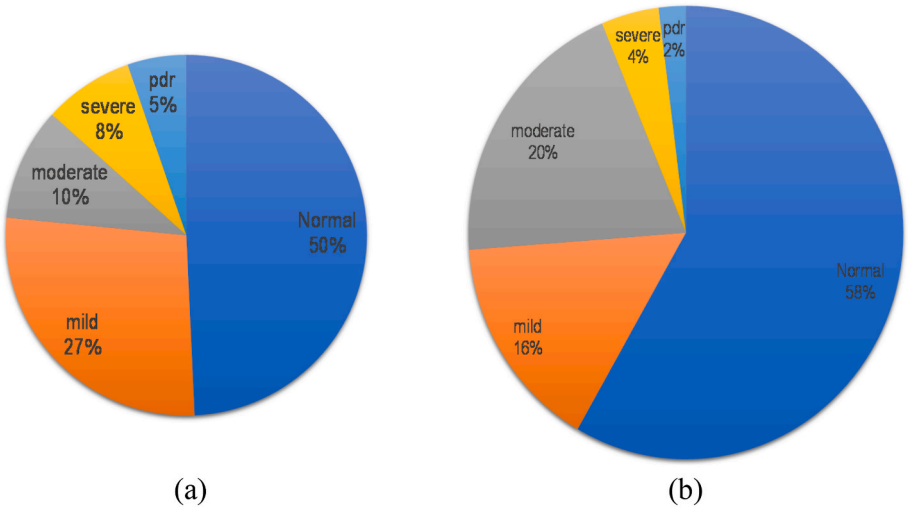

Fig. 22. The ratios of the number of images in datasets: (a) APTOS 2019 (b) Messidor-2.

Table 6
Comparison of overall model metrics with state-of-the-art methods for both binary and multiclass classifications.

Reference Number	No. of classes	Accuracy (%)	Precision (%)	Recall (%)
Binary classification				
[55]	2	97.82	98.00	98.00
[57]	2	96.10	–	–
Proposed	2	98.34	98.35	98.35
[56]	2	94.44	–	87.00
[58]	2	97.05	–	–
Proposed method	2	98.36	98.36	98.37
Multiclass classification				
[55]	5	80.96	–	–
[59]	5	77.90	**76.00**	**77.00**
[23]	5	82.18	–	–
Xception [60]	5	79.59	–	**82.35**
ResNet50 [60]	5	74.64	–	56.52
Inceptionv3 [60]	5	78.72	–	63.64
Proposed method	5	84.32	70.53	73.81
[58]	5	75.50	59.40	54.60
[61]	5	77.00	–	–
Proposed	5	84.36	70.51	73.84

Table 7
Comparison of class-wise model metrics with state-of-the-art methods for multiclass classification.

Reference Number	No DR	Mild	Moderate	Severe	PDR
Precision metric (%)					
[58]	93	59	64	32	49
Proposed method	99.62	61.33	76.43	50	81.82
Recall metric (%)					
[58]	97	36	73	27	40
Proposed method	97.79	82.14	80	51.72	40.91

used pre-trained CNN models to represent the FI. Spatial pooling techniques are described for obtaining the reduced versions of these representations without losing a lot of information. They used 80% of 3662 images for training and 20% (733) images for testing purposes. For binary classification, they achieved 97.82% accuracy, 98% precision, and 98% recall score. Pre-trained DenseNet121 with several modifications was proposed by Chaturvedi et al. [56], and for binary classification, they got an accuracy of 94.44% and a recall score of 87%. They used 15% (550) of 3662 images for testing purposes. A blended multi-modal fusion model was proposed by Bodapati et al. [57]. They extracted features from VGG16-fc1, fc2 layers and Xception convolutional layers, and later blended them using 1D and cross pooling to get better representation. They used 80% of 3662 images for training and 20% (733) testing purposes. They got an accuracy of 96.1% for the DR identification and 80.96% for severity classification. They did not show precision, recall scores. Kumar et al. [58] proposed a hybrid model composed of VGG16 and Capsule network and achieved an accuracy of 97.05% for the DR identification and 75.50% for the five stages classification. 15% of the 3662 images of APTOS dataset was used by them for the testing purpose. Table 6 shows that for the binary classification, the proposed model outperforms the existing models. It can accurately identify the presence of the DR in a FI.

Dondeti et al. [59] extracted features using Neural Architecture Search Network (NASNet) and projected them into lowed dimensional space using the t-SNE method, and using the v-SVM classifier, they got 77.90% accuracy, 76% precision, and 77% recall for five stages grading of the APTOS dataset. 80% of the 3662 images was used for training and the remaining 20% was used for the testing. Gangwar et al. [23] proposed a hybrid model pre-trained with Inception-Resnet-v2, and 82.18% accuracy was achieved for five class classification on APTOS dataset. Kassani et al. [60] used transfer learning models to extract the features from the FIs. Using the extracted features, they were able to classify the five stages of DR using a multilayer perceptron (MLP) neural network. Dekhil et al. [61] proposed a customized CNN model with five convolutional layers and obtained an accuracy of 77% for five stages grading of the APTOS 2019 dataset. They use 15% of the 3662 images for the testing purpose. It can be observed that for the multiclass classification (five classes), the proposed model outperforms the existing models.

From Table 6, it could be seen that for the multiclass classification, the proposed model achieved higher accuracy (84.36%) than the existing models. Though the overall precision and recall score of the model for the multiclass classification did not outperform some existing models but a high AUC score of 93.819% was obtained. This indicated that the model could distinguish the DR stages accurately. This higher AUC could be the result of using the supervised contrastive loss function, which attracted samples of the same class closer and pushed the samples of different classes apart in the projected embedding space [50]. Therefore, the model achieved state-of-the-art performance for the multiclass classification.

For advanced stages of the DR there is no known treatment. Diagnosis at the earlier or mild stage, will provide practitioner a chance to study the patient's glucose, lipid profile, and other risk factors. Then, imposing a strong control would reduce the progression of the DR to the later stages [54]. Only one study was found in the literature for the

multiclass classification with class-wise metric values. From Table 7 [58], attained a precision score of 93%, 59%, 64%, 32%, and 49% for the normal, mild, moderate, severe, and PDR stages, respectively. Again, for the recall metric, they achieved 97%, 36%, 73%, 27%, and 40% for the normal, mild, moderate, severe, and PDR stages, respectively. It can be seen that for the earlier stages especially normal, mild, and moderate stages, the model achieved higher precision and recall scores than the existing method. Therefore, the combination of higher overall accuracy, and higher class-wise precision and recall values than the existing methods and an overall high value of AUC for both the binary and multiclass classifications demonstrated a superior performance of the proposed model particularly for the DR detection at the earlier stages. With this information the health practitioners can devise a plan for effectively controlling the diabetes and possibly preventing the escalation of the DR in the later stages, which can result in vision loss.

6. Conclusions and future work

The proposed model for the DR grading using supervised contrastive loss function has achieved an overall accuracy, precision, recall, F1-score and AUC of 98.36%, 98.365%, 98.365%, 98.365% and 98.50% respectively during the binary classification for APTOS 2019 dataset and the results outperform the existing works. For the multiclass classification (APTOS 2019), the proposed model has achieved a higher accuracy of 84.364%, which again outperforms the existing methods. Furthermore, the proposed method has evidenced better precision and recall values for detecting the earlier stages of the DR signifying the robustness of the model. Furthermore, the model obtained an AUC score of 93.819%, indicating its capability to accurately distinguish among the classes. Ablation study conducted with Messidor-2 datasets also proved the superiority of the SCL over the conventional CNN model. Therefore, the experimental analysis demonstrated that the proposed model achieved state-of-the-art performance.

In near future, not only the severity levels of the DR will be detected but also localization of the various lesions will be conducted. Radiologists can be involved to interpret the annotations of the dataset. Further experiments will be conducted to validate the model using other datasets with larger batch sizes. Again, apps can be developed in order to facilitate the practical application of the model in the clinical applications.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

None.

Acknowledgments

The authors would like to thank Rajshahi University of Engineering and Technology (RUET) for supporting to conduct the research.

References

[1] I.D.F.D. Atlas, International Diabetes Federation — Facts & Figures, 2021.
[2] K. Boyd, American Academy of Ophthalmology - what Is Diabetic Retinopathy, 2021.
[3] R. Taylor, D. Batey, Handbook of Retinal Screening in Diabetes: Diagnosis and Management, John Wiley & Sons, 2012.
[4] W.L. Alyoubi, M.F. Abulkhair, W.M. Shalash, Diabetic retinopathy fundus image classification and lesions localization system using deep learning, Sensors 21 (2021) 3704.
[5] R.R.A. Bourne, G.A. Stevens, R.A. White, J.L. Smith, S.R. Flaxman, H. Price, J. B. Jonas, J. Keeffe, J. Leasher, K. Naidoo, others, Causes of vision loss worldwide, 1990–2010: a systematic analysis, Lancet Global Health 1 (2013) e339–e349.
[6] D.A. Salz, A.J. Witkin, Imaging in diabetic retinopathy, Middle East Afr, J. Ophthalmol. 22 (2015) 145.
[7] A. Sharafeldeen, M. Elsharkawy, F. Khalifa, A. Soliman, M. Ghazal, M. AlHalabi, M. Yaghi, M. Alrahmawy, S. Elmougy, H.S. Sandhu, Precise higher-order reflectivity and morphology models for early diagnosis of diabetic retinopathy using OCT images, Sci. Rep. 11 (2021) 1–16.
[8] H.S. Sandhu, M. Elmogy, A.T. Sharafeldeen, M. Elsharkawy, N. El-Adawy, A. Eltanboly, A. Shalaby, R. Keynton, A. El-Baz, Automated diagnosis of diabetic retinopathy using clinical biomarkers, optical coherence tomography, and optical coherence tomography angiography, Am. J. Ophthalmol. 216 (2020) 201–206.
[9] S.E. Moss, R. Klein, S.D. Kessler, K.A. Richie, Comparison between ophthalmoscopy and fundus photography in determining severity of diabetic retinopathy, Ophthalmology 92 (1985) 62–67.
[10] G.F. Elsayed, D. Krishnan, H. Mobahi, K. Regan, S. Bengio, Large margin deep networks for classification, ArXiv Prepr. ArXiv1803.05598 (2018) 850–860.
[11] Z. Zhang, M.R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, 32nd Conf. Neural Inf. Process. Syst. (2018) 8792–8802.
[12] M.M. Islam, H.-C. Yang, T.N. Poly, W.-S. Jian, Y.-C.J. Li, Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis, Comput. Methods Progr. Biomed. 191 (2020), 105320.
[13] L.K. Ramasamy, S.G. Padinjappurathu, S. Kadry, R. Damaševičius, Detection of diabetic retinopathy using a fusion of textural and ridgelet features of retinal images and sequential minimal optimization classifier, PeerJ Comput. Sci. 7 (2021).
[14] P.R. Asha, S. Karpagavalli, Diabetic retinal exudates detection using machine learning techniques, Int. Conf. Adv. Comput. Commun. Syst. (2015) 1–5.
[15] A. Ali, S. Qadri, W. Khan Mashwani, W. Kumam, P. Kumam, S. Naeem, A. Goktas, F. Jamal, C. Chesneau, S. Anam, others, Machine learning based automated segmentation and hybrid feature analysis for diabetic retinopathy classification using fundus image, Entropy 22 (2020) 567.
[16] S. Gayathri, V.P. Gopi, P. Palanisamy, A lightweight CNN for Diabetic Retinopathy classification from fundus images, Biomed. Signal Process Control 62 (2020), 102115.
[17] N. Sikder, M. Masud, A.K. Bairagi, A.S.M. Arif, A.-A. Nahid, H.A. Alhumyani, Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images, Symmetry 13 (2021) 670.
[18] M. Chetoui, M.A. Akhloufi, M. Kardouchi, Diabetic retinopathy detection using machine learning and texture features, IEEE Can. Conf. Electr. Comput. Eng. (2018) 1–4, https://doi.org/10.1109/CCECE.2018.8447809, 2018.
[19] S.M.A. Huda, I.J. Ila, S. Sarder, M. Shamsujjoha, M.N.Y. Ali, An improved approach for detection of diabetic retinopathy using feature importance and machine learning algorithms, Int. Conf. Smart Comput. Commun. (2019 7th) 1–5, https://doi.org/10.1109/ICSCC.2019.8843676, 2019.
[20] H. Liu, K. Yue, S. Cheng, C. Pan, J. Sun, W. Li, Hybrid model structure for diabetic retinopathy classification, J. Healthc. Eng. 2020 (2020).
[21] S. Sheikh, U. Qidwai, Smartphone-based diabetic retinopathy severity classification using convolution neural networks, in: Proc. SAI Intell. Syst. Conf., Springer, 2020, pp. 469–481.
[22] K. Xu, D. Feng, H. Mi, Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image, Molecules 22 (2017) 2054.
[23] A.K. Gangwar, V. Ravi, Diabetic retinopathy detection using transfer learning and deep learning, in: Evol. Comput. Intell., Springer, 2021, pp. 679–689.
[24] D.J. Hemanth, O. Deperlioglu, U. Kose, An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network, Neural Comput. Appl. 32 (2020) 707–721.
[25] S. Das, K. Kharbanda, M. Suchetha, R. Raman, E. Dhas, Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy, Biomed. Signal Process Control 68 (2021), 102600.
[26] R. Pires, S. Avila, J. Wainer, E. Valle, M.D. Abramoff, A. Rocha, A data-driven approach to referable diabetic retinopathy detection, Artif. Intell. Med. 96 (2019) 93–106.
[27] Y.-P. Liu, Z. Li, C. Xu, J. Li, R. Liang, Referable diabetic retinopathy identification from eye fundus images with weighted path for convolutional neural network, Artif. Intell. Med. 99 (2019), 101694.
[28] L. Math, R. Fatima, Adaptive machine learning classification for diabetic retinopathy, Multimed. Tool. Appl. 80 (2021) 5173–5186.
[29] X. Zeng, H. Chen, Y. Luo, W. Ye, Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network, IEEE Access 7 (2019) 30744–30753.
[30] W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, Z. Yi, Automated identification and grading system of diabetic retinopathy using deep neural networks, Knowl. Base Syst. 175 (2019) 12–25.
[31] H. Kaushik, D. Singh, M. Kaur, H. Alshazly, A. Zaguia, H. Hamam, Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models, IEEE Access 9 (2021), https://doi.org/10.1109/ACCESS.2021.3101142, 108276–108292.
[32] S. Maqsood, R. Damaševičius, R. Maskeliūnas, Hemorrhage detection based on 3D CNN deep learning framework and feature fusion for evaluating retinal abnormality in diabetic patients, Sensors 21 (2021) 3865.
[33] Y. Xu, Z. Zhou, X. Li, N. Zhang, M. Zhang, P. Wei, FFU-Net: feature fusion U-Net for lesion segmentation of diabetic retinopathy, BioMed Res. Int. 2021 (2021).
[34] M.K. Hasan, M.A. Alam, M.T.E. Elahi, S. Roy, R. Martí, DRNet: segmentation and localization of optic disc and fovea from diabetic retinopathy image, Artif. Intell. Med. 111 (2021), 102001.

[35] T. Nazir, A. Irtaza, J. Rashid, M. Nawaz, T. Mehmood, Diabetic retinopathy lesions detection using faster-RCNN from retinal images, Int. Conf. Smart Syst. Emerg. Technol. (2020 First) 38–42, https://doi.org/10.1109/SMART-TECH49988.2020.00025, 2020.
[36] N. Sambyal, P. Saini, R. Syal, V. Gupta, Modified U-Net architecture for semantic segmentation of diabetic retinopathy images, Biocybern. Biomed. Eng. 40 (2020) 1094–1109.
[37] G.T. Zago, R.V. Andreão, B. Dorizzi, E.O.T. Salles, Diabetic retinopathy detection using red lesion localization and convolutional neural networks, Comput. Biol. Med. 116 (2020), 103537.
[38] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, P.-A. Heng, CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading, IEEE Trans. Med. Imag. 39 (2019) 1483–1493.
[39] M.H. Mahmoud, S. Alamery, H. Fouad, A. Altinawi, A.E. Youssef, An automatic detection system of diabetic retinopathy using a hybrid inductive machine learning algorithm, Personal Ubiquitous Comput. (2021) 1–15.
[40] R. Afrin, P.C. Shill, Automatic lesions detection and classification of diabetic retinopathy using fuzzy logic, Int. Conf. Robot. Electr. Signal Process. Tech. (2019) 527–532.
[41] S. Lal, S.U. Rehman, J.H. Shah, T. Meraj, H.T. Rauf, R. Damaševičius, M. A. Mohammed, K.H. Abdulkareem, Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition, Sensors 21 (2021) 3922.
[42] Aravind Eye Hospital, APTOS 2019 blindness detection Kaggle (n.d.), https://www.kaggle.com/c/aptos2019-blindness-detection. (Accessed 11 June 2021).
[43] M. Tsighe Hagos, S. Kant, Transfer Learning Based Detection of Diabetic Retinopathy from Small Dataset, ArXiv E-Prints, 2019 arXiv–1905.
[44] M.D. Abràmoff, J.C. Folk, D.P. Han, J.D. Walker, D.F. Williams, S.R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, Automated analysis of retinal images for detection of referable diabetic retinopathy, JAMA Ophthalmol 131 (2013) 351–357.
[45] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, Feedback on a publicly distributed image database: the Messidor database, Image Anal. Stereol. 33 (2014) 231–234.
[46] R.A. Manju, G. Koshy, P. Simon, Improved method for enhancing dark images based on CLAHE and morphological reconstruction, Procedia Comput. Sci. 165 (2019) 391–398.
[47] S.A. Khan, S. Hussain, S. Yang, Contrast enhancement of low-contrast medical images using modified contrast limited adaptive histogram equalization, J. Med. Imaging Heal. Informatics. 10 (2020) 1795–1803.
[48] K. Zuiderveld, Contrast limited adaptive histogram equalization, Graph. Gems. (1994) 474–485.
[49] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J.B. Zimmerman, K. Zuiderveld, Adaptive histogram equalization and its variations, Comput. Vision, Graph, Image Process. 39 (1987) 355–368.
[50] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised Contrastive Learning, vol. 11362, 2020. ArXiv Prepr. ArXiv2004.
[51] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1251–1258.
[52] G. Hinton, S.T. Roweis, Stochastic Neighbor Embedding, NIPS, 2002, pp. 833–840.
[53] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, J. Mach. Learn. Technol. 2 (2011) 37–63.
[54] L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, H. Kong, R. Liu, X. Wang, X. Hou, Y. Liu, others, A deep learning system for detecting diabetic retinopathy across the disease spectrum, Nat. Commun. 12 (2021) 1–11.
[55] J.D. Bodapati, N.S. Shaik, V. Naralasetti, Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification, J. Ambient Intell. Hum. Comput. (2021) 1–15.
[56] S.S. Chaturvedi, K. Gupta, V. Ninawe, P.S. Prasad, Automated diabetic retinopathy grading using deep convolutional neural network, ArXiv Prepr. ArXiv2004.06334 (2020) 1–12.
[57] J.D. Bodapati, V. Naralasetti, S.N. Shareef, S. Hakak, M. Bilal, P.K.R. Maddikunta, O. Jo, Blended multi-modal deep convnet features for diabetic retinopathy severity prediction, Electronics 9 (2020) 914.
[58] G. Kumar, S. Chatterjee, C. Chattopadhyay, DRISTI: a hybrid deep neural network for diabetic retinopathy diagnosis, Signal, Image Video Process (2021) 1–8.
[59] V. Dondeti, J.D. Bodapati, S.N. Shareef, N. Veeranjaneyulu, Deep convolution features in non-linear embedding space for fundus image classification, Rev. d'Intelligence Artif. 34 (2020) 307–313.
[60] S.H. Kassani, P.H. Kassani, R. Khazaeinezhad, M.J. Wesolowski, K.A. Schneider, R. Deters, Diabetic retinopathy classification using a modified xception architecture, IEEE Int. Symp. Signal Process. Inf. Technol. (2019) 1–6.
[61] O. Dekhil, A. Naglah, M. Shaban, M. Ghazal, F. Taher, A. Elbaz, Deep learning based method for computer aided diagnosis of diabetic retinopathy, IEEE Int. Conf. Imaging Syst. Tech. (2019) 1–4.