Multimedia Tools and Applications (2024) 83:89837–89870
https://doi.org/10.1007/s11042-024-18968-5

Grading the severity of diabetic retinopathy using an ensemble of self-supervised pre-trained convolutional neural networks: ESSP-CNNs

Saeed Parsa[1] · **Toktam Khatibi**[2]

Received: 23 October 2023 / Revised: 22 December 2023 / Accepted: 13 March 2024 /
Published online: 2 April 2024

Abstract
Diabetic retinopathy (DR) is a common eye disorder that can lead to vision problems and blindness, necessitating accurate grading for effective treatment. While various artificial intelligence (AI) systems have been developed, surpassing human analysis in detecting DR, deep neural networks require large annotated datasets to learn the complex patterns and relationships necessary for grading, which are often limited in availability, to learn the intricate patterns and relationships required for accurate grading. However, such datasets are often limited in availability, requiring significant investments of human resources and time for the labeling process. To address these challenges, we propose ESSP-CNNs, a framework that harnesses popular CNN architectures (VGGNet, AlexNet, and ResNet). Our approach employs self-supervised learning, specifically the Bootstrap Your Own Latent (BYOL) technique, to pre-train neural networks on a vast unlabeled dataset. Additionally, we employ deep ensemble learning to construct a robust model for DR grading. Our methodology encompasses three main components: preprocessing fundus images, BYOL-based pre-training, and ensemble model construction. We conduct experiments and comparisons using the EyePACS and IDRiD datasets, with BYOL pre-training on EyePACS to enable the CNN models to acquire meaningful representations of fundus images, while IDRiD is used for severity grading. The performance of the proposed framework is further confirmed through thorough validation using the Messidor dataset. Through extensive experimentation on the IDRiD and Messidor datasets, ESSP-CNNs achieve notable accuracies of 71.84% and 75.42%, specificities of 88.76% and 87.13% along with AUC of 86.02% and 86.54%, respectively. The experimental results validate the effectiveness of our methodology in grading the severity of DR, with the ensemble model built from pre-trained CNNs yielding promising outcomes. Moreover, we compare our methodology against other state-of-the-art methods in DR grading, and our results demonstrate its satisfactory performance, surpassing previous alternatives in accurately assessing DR severity.

Keywords Diabetic retinopathy · Deep learning · Convolutional Neural Networks · Self-supervised learning · Ensemble learning · Color fundus images

Extended author information available on the last page of the article

1 Introduction

Diabetic retinopathy (DR) is a notable complication that arises from diabetes, and if left untreated, it can lead to vision impairment and blindness by damaging the retina [1]. Early diagnosis and accurate grading of the disease's severity are crucial steps to ensure proper treatment and prevent vision impairment [2]. The level of severity in DR is assessed by the formation of various lesions on the retina, which have a tendency to worsen gradually over time, ultimately contributing to the progression of the disease [3]. To accurately grade the severity of DR, it is essential to consider the number and type of these lesions, as they indicate the extent of damage caused to the retina [3]. Therefore, ophthalmologists and healthcare providers use various imaging techniques, such as optical coherence tomography (OCT) and fundus photography [4], to detect and assess these lesions and determine the appropriate course of treatment for the patient. Color fundus imaging, a widely used method in clinical research, allows for the examination of retinal structure and identification of abnormal features associated with DR. It is an affordable and easily accessible technique employed to monitor and grade the severity of DR [5, 6].

In regard to DR severity grading, various severity scales have been suggested and implemented in various countries due to the intricate nature of the disease [3]. The International Clinical Diabetic Retinopathy scale is the most commonly used method for grading DR severity [7]. This scale assigns a grade from 0 to 4 for DR based on the presence and severity of specific retinal abnormalities, as outlined in Table 1.

The provided image samples in Fig. 1 depict a selection of color fundus images obtained from the training set of the widely accessible IDRiD dataset [8]. Each image showcases a different DR grade based on the severity levels outlined in Table 1. The leftmost image illustrates a healthy retina, while the severity of the disease increases from left to right. The final image with grade 4 demonstrates various types of lesions visible on the retina, indicating the highest level of severity among the examples displayed.

According to the estimation of the International Diabetes Federation (IDF), the surge in diabetic patients globally has heightened the challenges associated with diagnosing DR, making it a significant burden on healthcare resources [9]. Beyond its sheer prevalence, DR presents formidable challenges to healthcare systems worldwide. The manual

Table 1 Assessing the severity of DR by evaluating various lesions using the International Clinical Diabetic Retinopathy scale

Grade	DR severity levels	Description of each stage
0	No DR	Absence of any abnormalities
1	Mild non proliferative DR	Exclusively microaneurysms
2	Moderate non proliferative DR	Beyond the presence of microaneurysms yet not reaching the severity of severe non-proliferative DR
3	Severe non proliferative DR	One or more of the following: ■ In each of the 4 quadrants, there is more than 20 intraretinal hemorrhage ■ 2 or more quadrants have evident venous beading ■ 1 or more quadrants have prominent intraretinal microvascular abnormalities but no indications of Proliferative DR
4	Proliferative DR	Neovascularization, pre-retinal hemorrhage, or a combination of the two

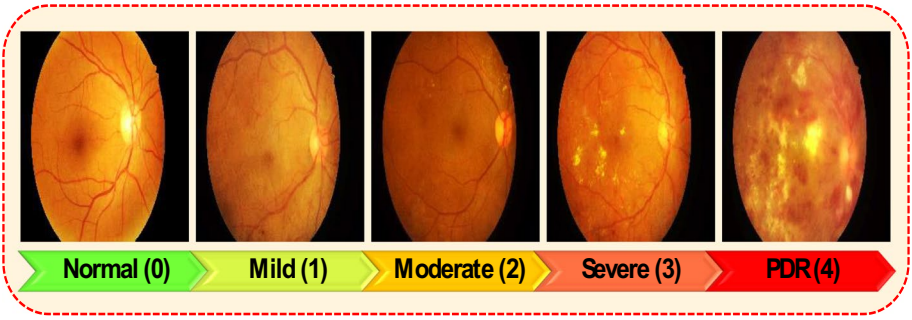

Fig. 1 Samples of IDRiD dataset images illustrate different severity grades of DR

grading of DR severity stands out as a particularly intricate and laborious task due to the multifaceted and nuanced effects of this disease on the retina [7]. This manual grading process not only demands substantial time and effort but also incurs significant costs [10]. Consequently, there exists an urgent and compelling need to innovate and implement an automated system capable of diagnosing DR and precisely grading its severity. Such a system would streamline the screening of high-risk patients, offering a valuable asset to clinicians in the accurate diagnosis of this debilitating disease.

The application of artificial intelligence (AI), notably through deep learning methods, has demonstrated promising outcomes in recognizing and discerning information within medical images [11]. With a specific focus on retinal images related to DR, these AI approaches have garnered considerable attention for their success in the diagnosis and classification of retinal conditions [3, 11]. While the last decade has witnessed a proliferation of research leveraging various deep learning algorithms for binary classification tasks related to DR (distinguishing healthy from diseased retinas) [3]., the nuanced task of grading the severity of different stages of DR has been comparatively less explored [3]. Therefore, a central motivation driving this study is to contribute to the field by automating the grading process and addressing the challenges associated with quantifying the severity of DR.

Deep Neural Networks (DNNs) have gained great popularity for grading the severity of DR using color fundus images [3, 12]. However, DNNs require large annotated datasets in order to be trained effectively and attain enhanced performance levels [13, 14]. In most of the previous studies, the same approaches have been followed to address the challenge of limited annotated data [3, 14]. For instance, to successfully train DNNs, data augmentation methods, which involve generating new data samples through a set of acceptable transformations, are applied [14]. Moreover, transfer learning methods, which leverage pre-trained neural networks formerly trained on large annotated datasets (such as the ImageNet [15]), are commonly used to improve the performance of models [3, 14]. however, pre-training a DNN on a large annotated dataset such as ImageNet provides successful performance in various domains of image classification, but this does not really solve the main problem of DNNs' dependence on substantial quantities of annotated data. Because as a fully supervised method, it requires a large amount of annotated data. In fact, in this approach, additional annotation data from another domain compensates for the lack of annotation in one domain.

Self-supervised learning presents a potentially effective alternative to conventional supervised pre-training methods, as it leverages unlabeled data to pre-train neural networks [13]. Self-supervised learning encompasses various techniques, among which Contrastive learning is one of the most widely used [16, 17]. By leveraging unlabeled data to create supervisory signals, this approach empowers machine learning (ML) models to learn without the need for labeled data, which can be costly or hard to obtain. To improve the performance of supervised classification, domain-specific representation is first acquired through contrastive learning before engaging in supervised learning. To be more precise, contrastive learning entails augmenting every image within the dataset by utilizing various methods such as cropping and altering color schemes. After augmenting the image data, the DL model creates a latent space depending on the similarity detected between the augmented images to capture the dataset features. The resulting model parameters obtained through contrastive learning can serve as a valuable starting point for training various models that use supervised learning techniques, ultimately leading to improved classification performance [17, 18].

The SimCLR framework is a simple contrastive learning technique used to learn representations of data by utilizing contrastive loss in the latent space via augmented views of the same image [17]. However, using random crops for augmenting images can result in some problems. To tackle this issue, bootstrap your own latent (BYOL), a novel framework for contrastive learning, which offers a more robust solution by eliminating the requirement for negative pairs and instead, bootstrapping on the learned representations [19]. While both contrastive learning frameworks have shown promising results in natural image processing, they may not be directly applicable to fundus images associated with DR. This is because the data augmentation techniques used in these frameworks may not be suitable for DR image augmentation. For example, color distortion techniques such as hue shifting or brightness adjustments, which are typically used in both frameworks, can significantly alter the appearance of images and change the color distribution of important features like blood vessels and lesions in the retina. This can make it more difficult for grading systems to accurately detect and classify these features, which could ultimately compromise the accuracy of the entire classification system. Therefore, in this study to introduce contrastive learning to DR fundus image classification particular image augmentation techniques are used to pre-train models for DR classification.

In this study, we proposed a novel methodology for grading DR that combines the advantages of contrastive learning and deep ensemble methods. Additionally, we present a pre-processing strategy based on CLAHE and Ben Graham approaches to enhance the quality of fundus images. The proposed method employs an ensemble of distinct convolutional neural network (CNN) models, which are pre-trained using the BYOL contrastive learning framework from unlabeled fundus images. Ensemble learning is a powerful ML technique that enhances the accuracy and reliability of predictions by combining the outcomes of multiple classification models into a single, robust classifier [20]. In other words, we utilized the BYOL contrastive learning framework to pre-train the three deep CNN models to develop an ensemble model for DR classification. The BYOL method allowed us to pre-train the CNN models on an extensive unlabeled dataset of DR without the need for labeled data, thereby enabling us to leverage the vast amount of unlabeled data available. We then fine-tuned the pre-trained models on labeled datasets for DR classification. During the fine-tuning process, the learned parameters of the previous model are transferred to the current CNN as initialization parameters. This approach allows us to build upon the knowledge gained from previous models, thereby improving the performance of the current model. To improve the performance of the CNN models, we employed an ensemble

method, which combines the predictions of the three fine-tuned models to obtain a more accurate and robust prediction. Our strategy leverages the strengths of each individual CNN model, while minimizing their weaknesses, leading to a better overall performance of the final model. By incorporating the strengths of contrastive learning and ensemble methods, our approach achieves superior accuracy in DR grading compared to conventional techniques.

The primary contributions of this paper can be outlined as follow:

- The CLAHE-Ben Graham preprocessing strategy is utilized to enhance the contrast and quality of the fundus image
- A self-supervised method is employed for pre-training models using a large unlabeled dataset to acquire valuable high-level representations that are beneficial for subsequent activities. Due to the considerable necessity of negative samples in numerous contrastive methods for achieving effectiveness, which can significantly hinder training efficiency, the contrastive learning framework BYOL is adopted. BYOL eliminates the requirement for negative examples. Furthermore, the data augmentation techniques are tailored to suit DR fundus image classification.
- An ensemble of pre-trained CNN models is developed which significantly improves the performance of our predictions, compared to the performance of any single model included in the ensemble.
- By leveraging unlabeled data, we have created a model that combines the benefits of deep ensemble learning and a self-supervised method. Our innovative approach has resulted in improved performance for DR grading when compared to conventional representative methods.

The succeeding sections of this paper are organized in the following format: Section 2 provides a concise analysis of past studies related to DR, while Section 3 presents the materials and methods considered in this study. Section 4 of this paper encompasses the presentation and analysis of the results, along with a thorough examination of their implications. Furthermore, a comprehensive discussion of the findings is presented, including a discussion of limitations and suggestions for future research directions. Finally, Section 5 provides key conclusions.

2 Literature review

This section presents a brief overview of the recent research on utilizing ML to automate the detection of DR through fundus photography. Many studies have previously dealt with the automatic detection of DR by employing traditional image processing and analysis techniques [12, 21]. Regarding this matter, a traditional sequence for the automatic detection of DR generally includes image preprocessing, extracting significant features by automatic and/or manual approaches, and ultimately classifying the image based on those features [12, 21]. Table 2 presents some pre-processing techniques utilized in previous studies that aim to make it easier to extract important features during the diagnosis in the subsequent stages. The objective of the pre-processing stage is to aid in this feature extraction process [12]. Some studies utilized conventional ML methods like naive Bayes (NB), support vector machine (SVM), random forests (RF), multilayer perceptron (MLP), and neural networks (NN), to identify DR [22]. However, currently, the most widely adopted

Table 2 A concise overview of previous relevant studies

Study	Task	Dataset	Modeling	Preprocessing
Huda et al. (2019) [26]	DR Detection	DIARETDB1	SVM, Decision Tree, K-Nearest Neighbor, Logistic Regression	Normalization, Standardization
Ramasamy et al. (2021) [27]	DR Detection	DIARETDB1 and KAGGLE	Feature extraction: grey-level co-occurrence and run length matrix, Ridgelet transform. Model: Sequential Minimal Optimization (SMO)	Filtering, Image conversion, Segmentation, Morphological operation
Mahmoud et al. (2021) [28]	DR Detection	CHASE	Feature extraction from segmented images and optimizing them. Model: hybrid inductive ML algorithm (HIMLA), multiple instance learning (MIL)	Contrast enhancement
Kaushik et al. (2021) [29]	DR Detection	EyePACS	Stacked generalization of three custom CNN	Luminosity normalization using gray world algorithm
Gangwar et al. (2021) [30]	DR Detection	Messidor-1 And APTOS	Inception-ResNet-v2	Getting coordinates of bounding box, image blurring, cropping, and resizing
BERBAR, M.A. (2022) [31]	DR Detection and Grading	Messidor-1, Messidor-2, Kaggle, and EyePACS	a binary classifier CNN, a Multi classifier CNN	Black corners removal, Median filter, Histogram match, Unsharp filter, CLAHE
Hardas et al. (2022) [32]	Retinal fundus image classification	DIARETDB1	Feature extraction: Gray-level co-occurrence matrix and object segmentation. Model: Gaussian mixture model, K-means, and SVM	Median filter, Grayscale conversion, Maximum a Posteriori algorithm, AHE, normalization, PCA
Zhang et al. (2022) [33]	DR Grading	EyePACS and APTOS	Source-Free Transfer Learning (SFTL)	Image resizing and normalization, Min-pooling filtering

Table 2 (continued)

Study	Task	Dataset	Modeling	Preprocessing
Islam et al. (2022) [22]	DR Detection and Grading	APTOS and Messidor-2	Supervised contrastive learning (SCL) and the Xception CNN model was deployed as the encoder with transfer learning	CLAHE
Luo et al. (2020) [11]	Retinal image classification	Messidor, DRIVE and DIARETDB1	Self-supervised Fuzzy Clustering Network (SFCN) by a feature learning module, reconstruction module, and a fuzzy self-supervision module	Image resizing
Kobat et al. (2022) [34]	Automated DR Detection	APTOS and a new custom dataset	Non-fixed-size patch division model, feature extraction with DenseNet20, feature selection by neighborhood component analysis, and classification by a cubic SVM	Image resizing, data augmentation/reduction
Qummar et al. (2019) [35]	DR Detection	Kaggle dataset	ensemble of five deep CNN models (Resnet50, Inceptionv3, Xception, Dense121, Dense169)	Image resizing, balancing the dataset
Sikder et al. (2021) [36]	Severity classification of DR	APTOS	Feature extraction using Histogram Features and Gray-Level Co-Occurrence then feature selection by Genetic Algorithm, classification using ensemble learning algorithm XGBoost	Noisy and duplicate image exclusion, Black border removal, resizing, Increasing sample sizes, Contrast enhancement, and Image tone mapping

approach for detecting and grading DR is by utilizing DNNs. With respect to the utilization of DNNs, conventional classification architectures like ResNet [23], VGG [24], or EfficientNet [25] are widely employed. Nonetheless, certain modified versions of these models have also been suggested.

Table 2 summarizes the research literature on the diagnosis of DR, indicating that previous studies have mainly focused on binary classification, and have achieved remarkable success in distinguishing between normal and DR images. However, the use of image processing techniques to calculate complex features in traditional ML methods can result in lengthy and intricate processing, which may not yield satisfactory results. Despite this, there are feasible ways to improve and advance ML models for the diagnosis of DR, especially in the grading of its severity, with great efficiency and without complex procedures. Recent literature suggests that deep descriptors can effectively represent retinal images, leading to a robust generalization of the model. To enhance the performance of DL methods, new hyperparameters can be explored, and efficient preprocessing techniques can be utilized. In this regard, there is significant potential for optimizing the performance of DL models in the diagnosis of DR. Hence, in this study, a novel methodology is devised for assessing the severity of DR by analyzing fundus images. The proposed model effectively leverages the benefits of self-supervised learning and ensemble learning techniques to significantly improve accuracy and efficiency in DR grading.

3 Methods

In this study, we propose an effective methodology for grading the severity of DR using color fundus images. Our approach employs self-supervised learning, specifically the BYOL technique, to pre-train neural networks on unlabeled images from a large dataset. To construct a robust model, we also employ deep ensemble learning techniques. Our methodology comprises two primary elements, as illustrated in Fig. 2. The first element uses the BYOL technique to pre-train the neural networks, and the second element utilizes these pre-trained networks to build the proposed ensemble model for DR grading. These elements have demonstrated promising outcomes in the field of general image classification. Once the pre-training and ensemble model-building phases are complete, the model becomes capable of analyzing color fundus images and predicting DR grading.

3.1 Dataset description

Initially, the neural networks are pre-trained using BYOL's proposed method on the public EyePACS dataset [37], where image labels are not utilized. Subsequently, for conducting experiments and making comparisons to perform severity grading for DR, this study mainly utilized the public IDRiD [8] and Messidor [38] datasets.

3.1.1 EyePACS

The EyePACS dataset is accessible to the public through Kaggle's data science competition. Comprising a vast collection of high-resolution retina images captured under diverse imaging conditions, the dataset includes both left and right fields for each subject. Images are systematically labeled with a subject ID and eye designation (e.g., 1_left.jpeg represents the left eye of patient ID 1). A clinician has assessed the presence of Diabetic

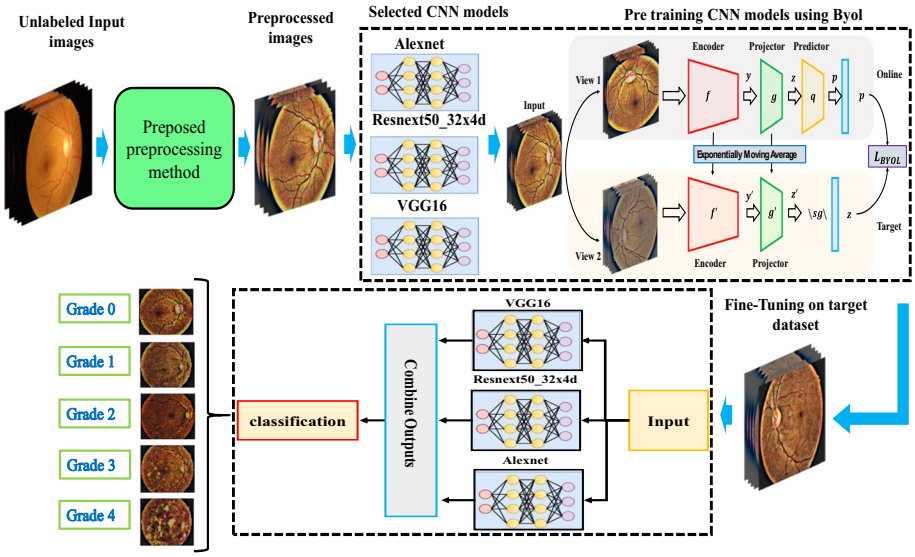

Fig. 2 The proposed ESSP-CNNs methodology involves using an ensemble of DL models along with a contrastive learning technique for the grading of DR. The approach leverages the strengths of both techniques to improve the accuracy of DR grading

Retinopathy (DR) on a scale ranging from 0 to 4, where 0 indicates no DR, and 4 indicates Proliferative DR. Recognizing the inherent challenges of real-world data, the dataset exhibits variations such as artifacts, focus issues, and exposure irregularities.

Given the extensive size of the EyePACS dataset, our study focused exclusively on its training set. This training subset encompasses 35126 fundus images captured using a variety of cameras. The primary aim of the dataset is to serve as a robust benchmark for the development of advanced algorithms capable of handling the complexities associated with real-world data, including images with inherent noise in both content and labels. The distribution of images across DR severity grades reveals a notable imbalance: 25802 images in class normal, 2438 in class mild, 5288 in class moderate, 872 in class severe, and 708 in class PDR. Figure 3 illustrates the frequency distribution of EyePACS training dataset classes at five DR severity levels, emphasizing the pronounced class imbalance.

Importantly, this dataset serves as the unlabeled dataset for the pre-training of neural networks using the Bootstrap Your Own Latent (BYOL) method. This strategy leverages the large, diverse set of unlabeled data within the EyePACS dataset to pre-train neural networks effectively, contributing to the subsequent success of the proposed methodology.

3.1.2 IDRiD and Messidor

The IDRiD dataset, representing the Indian Diabetic Retinopathy Image Dataset, is a significant resource for medical imaging in the Indian context. Notably, it provides detailed annotations of diabetic retinopathy lesions and normal retinal structures at a pixel level within its 516 fundus images. Originating from clinical examinations in Nanded, Maharashtra, India, this dataset serves as a crucial tool for developing and evaluating advanced image analysis algorithms focused on early diabetic retinopathy detection. Structured into three components, the dataset encompasses segmentation, disease grading, and

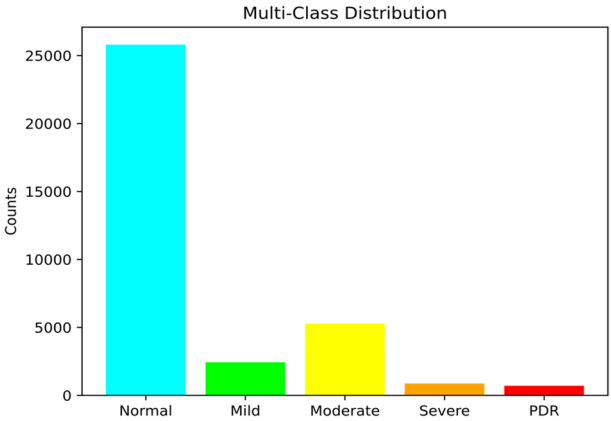

Fig. 3 Frequency of EyePACS training dataset in the five different leves

localization. The segmentation section includes 81 original color fundus images with corresponding ground truth images for lesions and the optic disc. Disease grading comprises 516 images categorized into a training set (413 images) and a test set (103 images), with accompanying ground truth labels for severity grades in CSV format. The localization component involves 516 images with ground truth labels for the optic disc center and fovea center locations in CSV format.

These images were captured using a Kowa VX-10 alpha fundus camera under pupil dilation, exhibiting a resolution of 4288×2848 pixels, stored in JPG format with an approximate size of 800 KB each. In this study, the Disease grading component is specifically utilized. Classified into five distinct severity grades, the accuracy of these images has been assessed by two medical professionals based on the ICDR severity scale. The dataset's partitioning into a predefined training set comprising 413 images and a separate test set comprising 103 images remains consistent across all experiments conducted in this study. It's important to note that the IDRiD dataset is also characterized by imbalances, further reflecting the real-world challenges in diabetic retinopathy data.

The Messidor dataset, a key component of the Messidor project, comprises 1200 fundus images captured by three French ophthalmology departments. These images, obtained using a Topcon TRC NW6 non-mydriatic retinography with a 3CCD camera, play a vital role in evaluating segmentation algorithms and image database management tools for diabetic retinopathy.

Detailed insights into the Messidor dataset reveal that the 1200 fundus color images were captured with 8-bit color depth per plane, varying resolutions of 1440×960, 2240×1488, or 2304×1536 pixels. Notably, 800 images were taken with pupil dilation (using one drop of Tropicamide at 0.5%), while 400 were captured without dilation. The dataset is organized into three sets, corresponding to ophthalmologic departments, and further divided into four zipped subsets containing 100 images each in TIFF format. Accompanied by an Excel file, these subsets provide medical diagnoses for each image, encompassing Retinopathy grade and Risk of macular edema. Retinopathy grades range from 0 (Normal) to 3, considering microaneurysms, hemorrhages, and neovascularization. Emphasizing privacy regulations, the Messidor dataset stands as a valuable resource for studies in computer-assisted diagnosis of diabetic retinopathy.

Since a predefined split was unavailable, the image samples were divided into a test set and a training set for the purpose of experimentation. For DR severity grading, the test set was determined by considering 20% of the 1200 samples, and the remaining 80% of the samples were assigned to the training set. This division enabled the model to be trained on the majority of the data and evaluated on a separate set of images. The data distribution for both the IDRiD and Messidor datasets is presented in Table 3, illustrating how the data is distributed among various categories.

3.2 Image preprocessing

Performing pre-processing on fundus images is crucial in order to improve the quality of images. This step is essential to maintain image consistency and enhance image features, which ultimately leads to better analysis results. Neglecting pre-processing can result in the use of low-quality fundus images, which may compromise the accuracy of the model. In that regard, an effective technique is presented for the proper pre-processing of fundus images. The following detailed steps describe the preprocessing techniques employed:

3.2.1 Black region removal

The first stage of the preprocessing phase is to remove the black regions present in the fundus images. Typically, in fundus images, the circular retinal image is surrounded by black areas that lack any useful information. Since these regions can affect the model's output, it's critical to remove them from the image to obtain precise and accurate results. To achieve this, a custom function is implemented, which adapts its behavior based on the dimensionality of the input image:

- For grayscale images, the function identifies pixels with intensity values greater than a specified tolerance threshold. The resulting mask is then applied to retain only the relevant regions of the image.
- In the case of RGB images, the function first converts the image to grayscale. It then performs the same process as described for grayscale images on each channel (R, G, B). The channels are subsequently stacked together to form the final preprocessed RGB image.

Table 3 Distribution of the datasets (IDRiD and Messidor)

DR stage	Number of training images		Number of test images	
	IDRiD	Messidor	IDRiD	Messidor
No DR (0)	132	428	36	118
Mild (1)	22	128	3	25
Moderate (2)	137	204	31	43
Severe (3)	72	200	21	54
PDR (4)	50	-	12	-
total	413	960	103	240

This careful approach to black region removal ensures that the circular retinal image is preserved while eliminating undesired background artifacts.

3.2.2 Uniform resizing and ben graham method

In the second stage of the preprocessing, all images are resized to a uniform size. The images size is various in datasets for instance the Messidor dataset comprises images of varying dimensions, such as 2240×1488, 2304×1536, and 1440×960 pixels. To minimize memory usage during the learning process and to ensure uniformity in image size for training the CNN model, all images are resized to 256×256 pixels. This ensures consistency in the dataset and optimal conditions for the model's training. Subsequently, the image quality is enhanced using the Ben Graham method, which was introduced by the winner of the DR competition on Kaggle [39]. This pre-processing technique involves three steps. Firstly, all images are rescaled to a standard radius of either 300 or 500 pixels. Secondly, the local average color is subtracted from the images, and it is important to note that 50% grey is mapped to the local average during this process. Lastly, the images are clipped to a size of 90% to eliminate any boundary effects.

3.2.3 Contrastive limited adaptive histogram equalization (CLAHE)

Once the Ben Graham method has been applied, the images are further improved by utilizing the Contrastive Limited Adaptive Histogram Equalization (CLAHE) [40] technique to improve their visibility and contrast levels. CLAHE is an enhanced version of AHE, which addresses the issue of noise amplification by limiting the degree to which contrast is amplified in the image. In order to implement the CLAHE method, the initial step involves converting the color space of the fundus image from BGR to LAB. Following this, the image is divided into three components, and the first component undergoes CLAHE processing with a clip limit of 2 and a tile grid size of 8×8. Once this is completed, the image components are combined back together and the color space is converted back from LAB to BGR. Ben Graham's approach is employed before CLAHE to ensure that any undesired noise present in the images is not amplified by the CLAHE method.

3.2.4 Denoising

In the final stage of the preprocessing step, a denoising method is applied to reduce noise in the images while preserving their edges and texture details. This denoising method uses a non-local means algorithm that compares each pixel in the image to a patch of pixels surrounding it, which is called a search window. The central pixel is then replaced by a weighted average of the values of the pixels within the search window, where the weight of each pixel is determined by its similarity to the central pixel. The impact of pre-processing can be observed in Fig. 4.

The enhancement of fundus images is clearly evident in the output images, as illustrated in Fig. 4, showcasing the effectiveness of the preprocessing algorithm (Algorithm 1) used to enhance the visibility and quality of the images. Notably, this enhancement results in improved clarity and increased recognizability of image details.

Input: Raw Image
Output: Enhanced Image
1) Read the input image and store it in a variable called ***Image***.
2) Remove any black regions in the ***Image*** and store the result in a variable called ***crImage***.
3) Resize ***crImage*** to a size of 256x256x3 and save it in a new variable called ***reImage***.
4) Apply the Ben-Graham approach to ***reImage***:
   a) Use Gaussian blur with parameters Gaussian Kernel Size= (0,0) and SigmaX= 6 to blur ***reImage***. Store the result in a variable called ***bluImage***.
   b) Blend ***bluImage*** with ***reImage*** using alpha = 4.5, beta = -4, and gamma = 30. Save the result in a variable called ***bgImage***.
5) Create a CLAHE with parameters Clip Limit = 2 and Tile-Grid-Size = (8,8)
6) Use CLAHE to enhance ***bgImage*** and store the result in a variable called ***enImage***.
7) Apply a denoising method to the ***enImage***.

3.3 Proposed framework (ESSP-CNNs)

We adopted the most prevalent method that is commonly employed in the literature to determine the severity of DR [3]. To determine the severity of DR, we employ a multi-class classification method, in which we predict the probability of multiple distinct classes. These classes represent different grades of the disease, with the number of grades varying depending on the dataset used: 4 for Messidor and 5 for IDRiD.

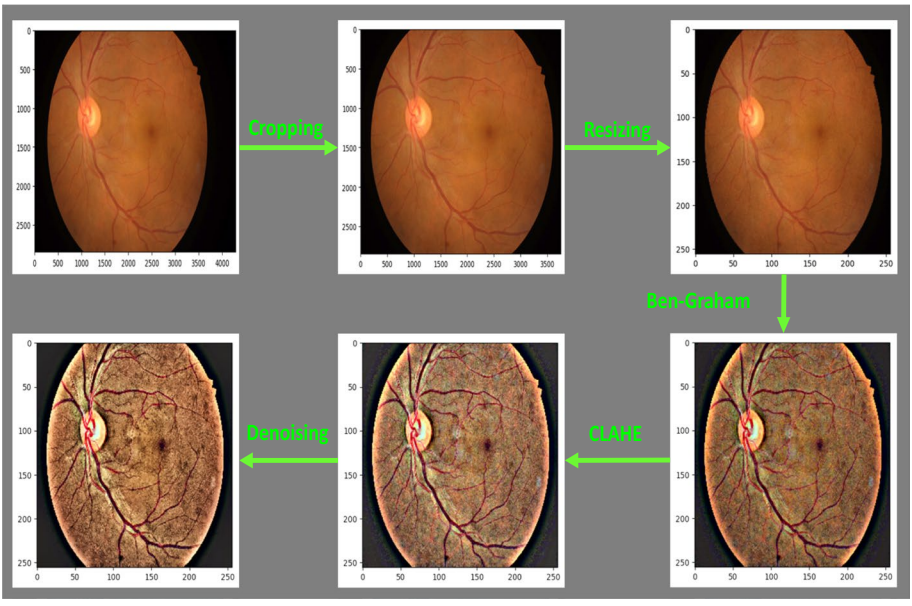

Fig. 4 The proposed method for pre-processing aim to improve the overall quality of retinal fundus images

Figure 2 illustrates the architecture of the proposed framework. As depicted, the unlabeled fundus images undergo pre-processing, after which the desired neural networks are pre-trained using the BYOL method. Subsequently, the pre-trained networks are fine-tuned using small labeled datasets. Finally, an ensemble model is constructed using these networks, which has shown promising results in DR grading.

3.3.1 CNN models

In this study, we employed various CNN architectures, each possessing distinct capabilities, including:

I. **AlexNet** [41]: AlexNet is a prominent CNN architecture that achieved victory in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). It is renowned for its innovative architecture comprising interconnected and stacked layers. The architecture consists of 5 convolutional layers, which extract features from the input images, followed by 3 fully connected layers responsible for classification. Throughout the network, several max-pooling layers are inserted to down-sample the feature maps and enhance spatial invariance. To accelerate the training procedure, each convolutional layer and a fully connected layer in AlexNet employ rectified linear unit activation, which introduces nonlinearity and helps in speeding up the learning process. This combination of architectural elements and techniques contributed to the success of AlexNet in image classification tasks.

II. **VGGNet** [24]: VGGNet, a widely recognized CNN, achieved an impressive second place in the renowned ILSVRC competition of 2014. The VGG-19 and VGG-16 models are the most successful VGG models and have 19 and 16 weight layers, respectively, which are available as pre-trained models. These models are notably deep, comprising 16 convolutional and fully connected layers. The convolutional layers use very small (3×3x3) filters and are very similar to each other. In contrast to AlexNet, which applies a pooling layer after each convolutional layer, VGGNet integrates a pooling layer after a series of three or two consecutive convolutional layers. In this study, the VGG-16 model is employed.

III. **ResNet** [23]: ResNet, another well-known CNN, gained prominence as the victor of the ILSVRC competition in 2015. Notably, ResNet surpasses both AlexNet and VGGNet in terms of depth, being approximately twenty times and eight times deeper, respectively. The distinguishing breakthrough of ResNet lies in its utilization of residual (RES) layers, which imbue it with a unique "network-in-network" structure. This network integrates batch normalization and skip connections into its design, and rather than employing fully connected layers at the network's end, it adopts global average pooling. Unlike other CNNs, ResNet explicitly formulates its layers to learn residual functions based on the input of each layer, rather than learning functions without any reference. This formulation makes the network much deeper than VGGNet while maintaining a smaller model size, making it easier to optimize. The ResNet architecture has multiple versions, and in our study, we utilized the ResNeXt50_32×4d model. This model has a cardinality of 32, which enables it to handle complex visual features in images. We chose this model due to its small memory footprint and its effectiveness in handling complex visual features.

We select these architectures based on their established reputation and significant contributions to the field of computer vision. VGG16's deep architecture and excellent performance on image recognition tasks, AlexNet's pioneering use of deep learning techniques and influential impact on the field, and ResNet50's introduction of residual learning for training very deep networks made them compelling choices for our research, ensuring robust and state-of-the-art performance [23, 24, 42].

3.3.2 Pretraining with BYOL

The proposed framework is characterized by several pivotal components, notably involving the pre-training of three mentioned CNN models through the application of the BYOL method [19] on an extensive, unlabeled dataset of retinal fundus images. BYOL, a distinct form of contrastive learning, places a strong emphasis on generating representations through the identification of positive pairs.

Addressing limited annotated data with BYOL The limited availability of annotated data poses a significant challenge in training accurate AI systems for DR grading. BYOL proves to be a valuable solution to this challenge by leveraging unlabeled data for pre-training. Contrastive learning methods, such as BYOL, excel in learning meaningful representations from positive pairs, enabling the models to capture intricate patterns in unlabeled data. This approach significantly reduces the reliance on annotated datasets, making it particularly suitable for medical imaging tasks where labeled data is often scarce.

Data source and characteristics The dataset selected for BYOL pre-training is drawn from the training set of the EyePACS dataset, a widely recognized repository of retinal fundus images. This dataset was chosen deliberately to leverage its diversity and relevance to the domain of retinal pathology. The dataset encompasses a rich spectrum of fundus images, capturing variations in retinal conditions, image quality, and anatomical structures. In our methodology, pre-training is specifically conducted on the EyePACS dataset, employing self-supervised learning techniques that operate without the need for annotated image labels. The principal objective of this pre-training phase is to derive a meaningful representation of input images, with the ultimate aim of applying these representations to downstream tasks such as DR classification.

Modification of BYOL for fundus image processing While BYOL is generally well-suited for image representation learning, its direct application to DR fundus image processing is constrained by the unique characteristics of these images and their sensitivity to the augmentation techniques employed in this approach. Consequently, specific modifications have been introduced to the augmentation techniques, tailoring them to the distinct requirements of fundus image processing.

In addressing this, we have introduced a novel augmentation technique specifically tailored for fundus images in this study. This involves the application of distinct augmentations to the two views of the fundus image utilized within the BYOL framework. Our method incorporates a range of techniques, including resizing, horizontal or vertical flipping, random Gaussian noise injection, and random contrast adjustments. Importantly, the novel image augmentation strategies, as visually represented in Fig. 5, are designed to avoid significant alterations in the color distribution of critical features such as blood vessels and lesions in the retina. These adjustments not only preserve essential visual

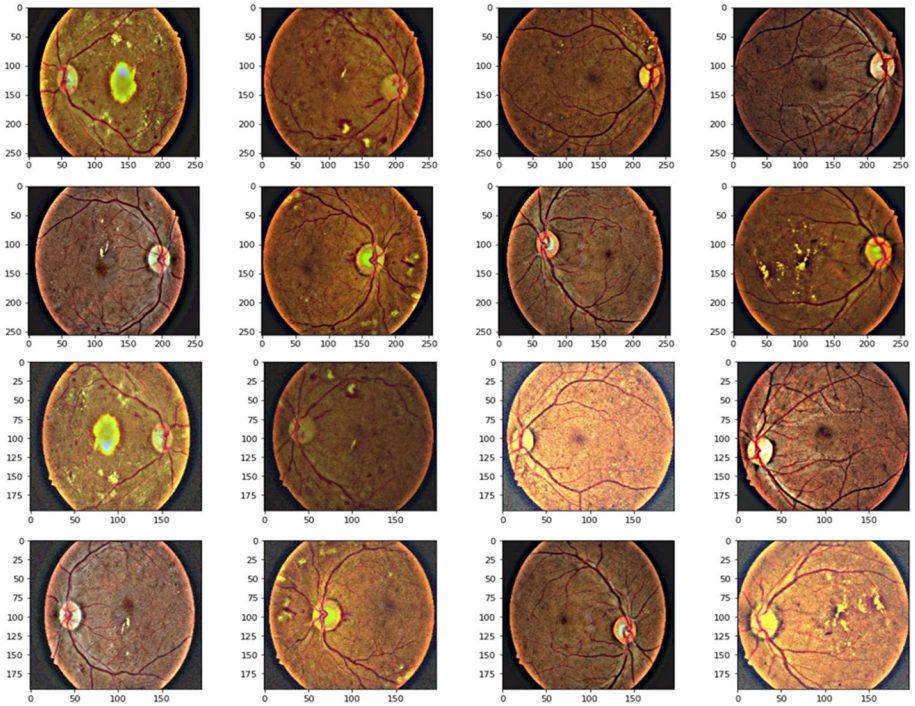

Fig. 5 Examples of fundus images before and after applying the augmentation techniques in our proposed framework

characteristics but also enhance the model's capacity to generalize effectively to previously unseen data.

BYOL architecture In the BYOL architecture, depicted in Fig. 6, the learning process relies on the interaction between two key components: the online network and the target network. Online and target networks work together to extract meaningful information from unlabeled data. Despite having the same structure, these networks employ different weights.

The online network consists of an encoder (f) responsible for generating a representation (y) from a slightly augmented view of a fundus image. This representation is then passed through a projector (g) to produce a projection (z), which is further utilized by a predictor module (q) to make predictions (p). Concurrently, the target encoder (f') processes a more strongly augmented view of the same image, generating a representation (y') and a corresponding projection (z') that serves as a reference. To facilitate learning, a stop-gradient function (sg) is employed. Once the projections from the target network (z_ξ) and the L2-normalized predictions from the online network (p_θ) are obtained, both components are used to calculate the loss using the provided equation, as outlined in Eq. (1).

$$L_{\theta,\xi} \triangleq \|p_\theta - z_\xi\|_2^2 = 2 - 2 \bullet \frac{\langle p, z' \rangle}{\|p\|_2 \bullet \|z'\|_2} \tag{1}$$

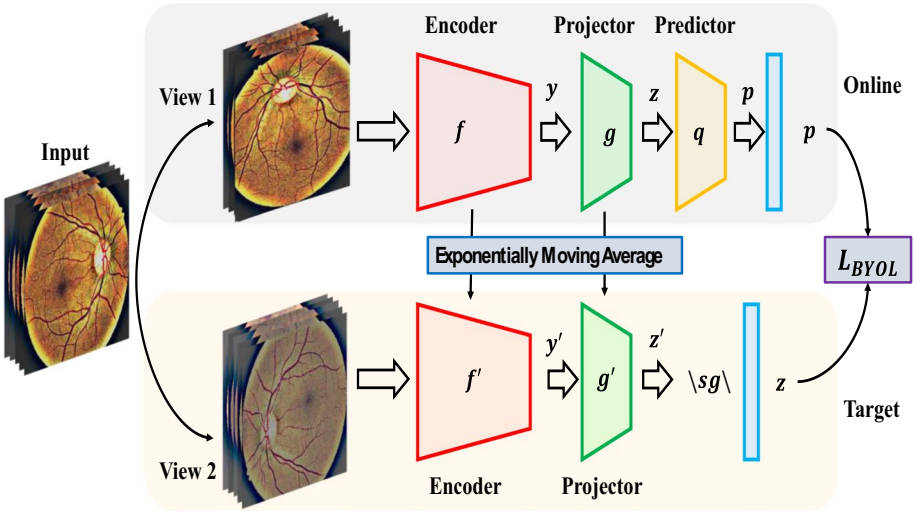

Fig. 6 The structure of the BYOL approach for pretraining networks

As mentioned earlier, to compute loss $L_{\theta,\xi}$, the slightly augmented version of fundus images is inputted to the online encoder, while the strongly augmented version of the same images is provided to the target encoder. To ensure symmetry, the strongly augmented version of fundus images is fed to the online encoder, and the slightly augmented versions are fed to the target encoder, resulting in the symmetrical loss $L'_{\theta,\xi}$. The final BYOL loss is obtained by adding $L_{\theta,\xi}$ and $L'_{\theta,\xi}$ together. During each iteration, the BYOL loss is minimized by adjusting the parameters of the online network (θ). Simultaneously, the parameters of the target network (ξ) are updated through an exponential moving average of θ. This update is performed using the formula $\xi \leftarrow \tau\xi + (1-\tau)\theta$, where τ represents the decay rate of the target.

Pretraining process In the pretraining phase using the EyePACS dataset, three distinct CNN models—AlexNet, ResNeXt50, and VGG16—are pretrained utilizing the BYOL method. This self-supervised learning technique involves predicting target features from augmented image pairs. The augmentations include resizing to model-specific dimensions (196×196 for AlexNet, 144×144 for ResNeXt50 and VGG16), random vertical and horizontal flips, random Gaussian noise with a probability of 0.5, and random contrast adjustments with a probability of 0.6. These operations ensure diverse representations while accounting for the sensitivity of fundus images. The BYOL training class is implemented using PyTorch Lightning, employing an MLP with a single hidden layer containing 4096 units for both the projector and predictor networks, each with an output dimension of 256.

The pretraining involves training the models with varying image sizes and durations: AlexNet (224×224, 80 epochs), ResNeXt50 (184×184, 50 epochs), and VGG16 (164×164, 50 epochs). A consistent batch size of 128, a learning rate of 3e-4, and a weight decay of 1e-6 are applied to all models. The BYOL approach facilitates the development of meaningful image representations without relying on explicit supervision. This critical pretraining step enhances the models' capacity to capture complex features in retinal fundus images, setting the stage for subsequent tasks related to diabetic retinopathy classification.

More details on the experimental results, including the pre-training process, performance metrics, and comparisons, are provided in the Experimental Results section.

3.3.3 Full fine-tuning CNN models

Fine-tuning is the process of training a pre-trained CNN model on a smaller dataset to adapt it to a new task. It involves adjusting the weights of the pre-trained model to better fit the new task with the new dataset in a supervised manner. Fine-tuning can improve model performance by leveraging pre-learned features from a large dataset, especially when the new dataset is small [43].

After pre-training the CNN models on the EyePACS dataset, the next step is to fine-tune them on two smaller datasets, IDRiD and Messidor, to adapt the learned features to the specific task of classifying DR. In this case, full fine-tuning is employed to train the CNN models on annotated images from these datasets, which involves fine-tuning all network layers of the pre-trained model. By fine-tuning all layers, the model can learn new features that are more relevant to the new task of DR classification, thereby improving its performance. Studies have shown that fine-tuning a pre-trained CNN model on target data can significantly enhance its performance [44], especially when the pre-trained model was initially trained on source data. This finding highlights the potential benefits of leveraging pre-trained models for related tasks and further fine-tuning them on target data to improve their adaptability to specific applications.

3.3.4 Ensemble design

Ensemble learning is a powerful ML method that aims to enhance predictive performance by combining multiple models [43, 45]. The ensemble model presented in this study utilizes AlexNet, ResNeXt50, and VGG16 models which in previous sections are pre-trained in a self-supervised manner and fine-tuned on specific datasets. The motivation behind employing an ensemble of these models lies in harnessing the complementary strengths of each individual architecture. Each CNN, having undergone self-supervised learning and fine-tuning, captures unique aspects and features of the input data related to diabetic retinopathy.

To provide a more nuanced understanding of the ensemble's decision-making process, we delve into the collaborative efforts of the individual CNN models. Each model contributes its distinctive insights by processing the input image independently. While the modified linear layers ensure compatibility for fusion, it is crucial to highlight that the ensemble model's strength lies in its ability to weigh and combine the contributions from each constituent CNN. The concatenated feature representations offer a comprehensive view of the input data, allowing the ensemble model to discern complex patterns and nuances associated with diabetic retinopathy.

To construct the ensemble model, modifications involve removing the last linear layer of each pre-trained CNN and replacing it with an identity function. This ensures the preservation of feature representations without further transformations. Concatenating these preserved outputs forms a fused representation, which undergoes classification through a new linear layer, the classifier. The classifier adapts to map the concatenated features to the corresponding class labels, applying the ReLU activation function to the feature vector before classification. During training, the parameters of pre-trained

models are frozen to safeguard the valuable features learned during self-supervised pre-training, and only the new linear layer undergoes updates to minimize computational overhead.

Figure 7 illustrates the proposed ensemble model, showcasing the collaboration of three self-supervised pre-trained CNNs that have undergone fine-tuning for fundus image classification. The inference process involves the parallel processing of the input image by each individual CNN, contributing to the concatenated feature representation. This fused representation undergoes classification through the new linear layer, culminating in the final prediction.

By combining the strengths of multiple CNNs and allowing them to contribute to the final prediction, the ensemble model excels in improving overall accuracy and generalization performance compared to using a single model. Additionally, the strategic freezing of pre-trained models' parameters helps maintain the quality of their learned representations, while selective training of the new linear layer optimizes computational efficiency.

3.4 Evaluation metrics

In order to evaluate the performance of our proposed approach, we utilize a comprehensive range of evaluation metrics. These metrics encompass F1-score, precision, accuracy, recall, and the area under the receiver operating characteristic (ROC) curve. These metrics provide a thorough evaluation of the performance of our method and enable us to gain insights into its effectiveness and reliability.

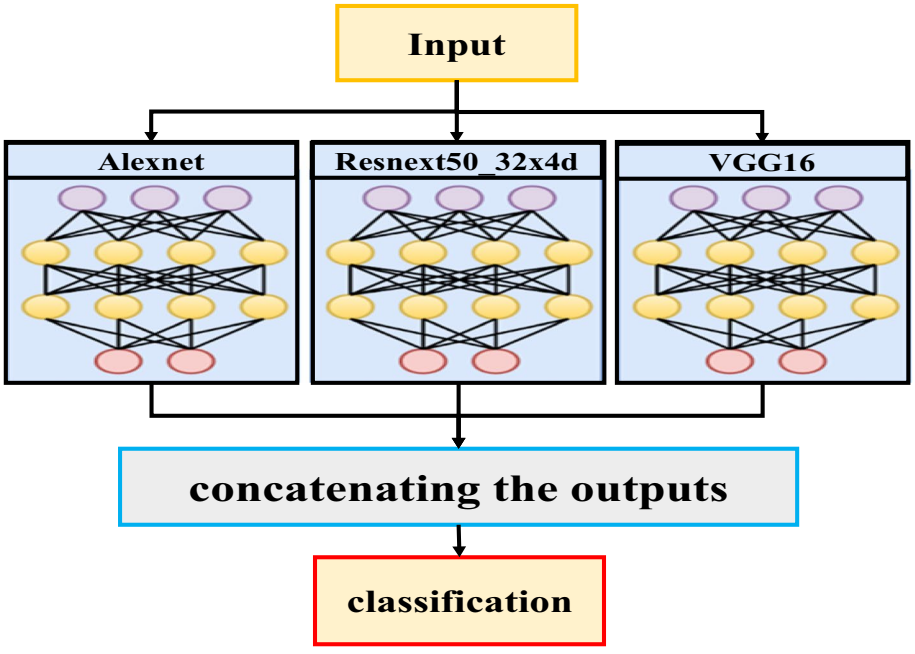

Fig. 7 The ensemble model in this study consists of three self-supervised pre-trained CNNs that have undergone fine-tuning for fundus image classification

Accuracy measures the proportion of correctly identified cases among all the detected cases (as depicted in Eq. (2)). This metric enables us to assess the algorithm's classification performance and determine how effectively it is able to accurately categorize the data [46].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

In this regard, TN (true negative), FN (false negative), FP (false positive), and TP (true positive) represent specific outcomes in the classification process.

Equation (3) defines precision, which is a fundamental metric that calculates the ratio of true positives out of all predicted positive instances [46]. In this study, weighted average precision (Eq. (4)) is utilized to evaluate the model's performance, considering class imbalances and providing a balanced measure of positive prediction accuracy across all classes.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Precision_w = \frac{\sum_{i=0}^{k} \frac{TP_i}{TP_i + FP_i} * N_i}{\sum_{i=0}^{k} N_i} \tag{4}$$

Recall, also known as sensitivity and defined by Eq. (5), measures the model's capability to correctly detect a significant portion of true positive cases [46]. In this study, the weighted recall (Eq. (6) is utilized to evaluate the model's performance accurately detecting every patient affected by DR, considering class imbalances, and offering a balanced assessment across all classes.

$$Recall = \frac{TP}{TP + FN} = Sensitivity \tag{5}$$

$$Recall_w = \frac{\sum_{i=0}^{k} \frac{TP_i}{TP_i + FN_i} * N_i}{\sum_{i=0}^{k} N_i} \tag{6}$$

The weighted F1-score, defined by Eq. (7), is derived from the harmonic mean of the weighted Recall and weighted Precision metrics [46]. It serves as a balanced measure that combines the strengths of both metrics, providing a comprehensive evaluation of the model's performance.

$$F1 - Score = \frac{2 \times Precision_w \times Recall_w}{Precision_w + Recall_w} \tag{7}$$

Specificity, as defined by Eq. (8), quantifies the accuracy of a classifier in correctly identifying actual negative instances. To assess the overall performance of the model in this study, the evaluation employs weighted specificity, as indicated by Eq. (9).

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Specificity_w = \frac{\sum_{i=0}^{k} \frac{TN_i}{TN_i + FP_i} * N_i}{\sum_{i=0}^{k} N_i} \tag{9}$$

4 Experimental result

This section provides a comprehensive account of the experimental procedures and findings, encompassing detailed descriptions and insightful analyses of the results.

4.1 Pre-training the networks

To pretrain the AlexNet, ResNeXt50, and VGG16 models, we utilize a comprehensive dataset of retinal fundus images called EyePACS. This dataset contains a diverse range of retina images exhibiting various pathological features like exudates, hemorrhages, and neovascularization, which are indicative of DR. In the pretraining process, we employ self-supervised learning, meaning image labels aren't used. The primary objective of pretraining is to acquire a meaningful representation of the input images, which could subsequently be utilized for downstream tasks like DR classification. For this purpose, we employ the BYOL method. This approach involves training a CNN model to predict a set of target features derived from a pair of augmented images. The target features are generated by a target encoder that continually updates its weights using a moving average of the main encoder's weights. This process facilitates the learning of effective representations from unlabeled data.

To implement BYOL first, a custom PyTorch module is developed to apply a series of augmentation functions to the input image. Due to the sensitivity of fundus images, specific augmentation operations are carefully selected to ensure their suitability for this particular domain. The Kornia library, which offers a wide range of fully differentiable computer vision operations, is employed for this purpose. The sequence of augmentation operations for pretraining each model is as follows:

- Resizing: The image is resized to a specific size based on the model being pretrained. In this case, we use a size of 196×196 for AlexNet, 144×144 for both ResNeXt50, and VGG16.
- Random Vertical and Horizontal Flips: The image undergoes random flips in both the vertical and horizontal directions, with a probability parameter set to 1. This means that the flips are always applied. These operations augment the dataset by introducing variations in the orientation of the fundus images.
- Random Gaussian Noise: A probability of 0.5 is assigned for adding Gaussian noise to the image, with a standard deviation of 0.05 and a mean of 0.1. This step introduces random variations to the pixel values of the image, simulating real-world noise that can be present in fundus images.
- Random Contrast Adjustments: Random contrast adjustments are applied to the image with a probability of 0.6. The contrast values are randomly selected from the range of 0.5 to 2.0. This operation enhances or reduces the difference between the pixel intensities, thereby altering the image's contrast.

These augmentation operations are consistent across all models during the pretraining stage, ensuring a unified approach.

In the next step, we develop an Encoder module, which plays a crucial role in extracting meaningful features from the base model and projecting them into a lower-dimensional latent space. To achieve this, we utilize a wrapper class in PyTorch, enabling seamless integration of BYOL with any model. The Encoder module consists of two key components:

- Feature Extractor: This component collects the outputs from one of the last layers of the base model. By capturing the representations generated at this stage, we can leverage high-level features that encode important information about the input data.
- Projector: This component comprises a linear layer responsible for projecting the extracted features into a space with lower dimensions.

Finally, we implement the BYOL training class using the PyTorch Lightning library, which provides a convenient framework for DL projects in PyTorch. The same architecture is used for both the projector and predictor networks. These networks consist of a Multi-Layer Perceptron (MLP) with a single hidden layer containing 4096 hidden units. The output dimensions of both the projector and predictor are set to 256.

In the pre-training phase, we train the CNN models using different image sizes and training durations based on the depth and memory requirements of each model. The AlexNet model is trained on pre-processed images of size 224×224 for 80 epochs. The ResNeXt50 model and VGG16 model are trained for 50 epochs with image sizes of 184×184 and 164×164, respectively. Throughout the training process, we utilize a batch size of 128 for all models. The learning rate is set to 3e-4, and a weight decay of 1e-6 is applied during the training process for all models. During pre-training, each CNN model is trained on the EyePACS dataset using BYOL to learn a meaningful representation of the input images. The CNN models learn to extract features that capture important characteristics of the retinal fundus images, such as the presence of exudates or hemorrhages, without relying on any explicit supervision. By leveraging the EyePACS dataset and employing the BYOL method, we aim to enhance the models' ability to capture and understand the complex features and characteristics present in retinal fundus images. This pretraining step plays a crucial role in preparing the models for subsequent tasks related to DR classification and analysis.

4.2 Experiments on IDRiD dataset

After pre-training the three CNN models using the BYOL technique on the EyePACS dataset, the subsequent phase involved fully fine-tuning them on the IDRiD dataset to construct the proposed ensemble model for DR classification. The full fine-tuning step entailed updating the weights of all layers in the pre-trained CNN models by leveraging labeled images from the IDRiD dataset. This crucial process facilitated the models' adaptation of learned features specifically tailored to the task of DR classification. By fine-tuning with the parameters outlined in Table 4, the models were optimized to deliver enhanced performance on the IDRiD dataset, thus improving their ability to accurately classify DR.

During the fine-tuning process, the CNN models underwent training to discriminate between different severity levels of DR using the labeled images from the IDRiD dataset. This fine-tuning on the IDRiD dataset provided an opportunity for the models to further refine their learned features, enabling them to better capture the specific characteristics

Table 4 Some hyperparameter values of the CNN models in the fine-tuning process on the IDRiD dataset

	Image size	Batch size	Learning rate	Number of epochs	Optimizer
AlexNet	256×256	24	0.00001	80	Adamax
ResNeXt50	256×256	18	0.00001	50	Adamax
VGG16	256×256	24	0.00001	80	Adamax

associated with the DR disease. Upon fully fine-tuning the pre-trained CNN models on the IDRiD dataset, an ensemble model (ESSP-CNNS) is constructed to enhance the accuracy of DR classification. By combining the refined models' learned features and decision-making capabilities, the ensemble model improves the reliability and robustness of the DR classification system.

The ESSP-CNNS model proposed in this study is trained using 80% of the IDRiD dataset for 80 epochs. During the training process, a batch size of 22, a learning rate of 1e-5, and the Adamax optimizer are utilized. The primary objective of the model is to classify DR into five severity grades. To assess the overall performance of the proposed model, the remaining 20% of the IDRiD dataset is reserved for testing.

Furthermore, to assess and compare the performance of individual AlexNet, ResNeXt50, and VGG16 models utilizing Imagenet weights with our ESSP-CNNS model, we conducted an evaluation on the IDRiD dataset using these selected CNNs. The evaluation includes analyzing the confusion matrix, as depicted in Fig. 8. The confusion matrix provides valuable insights into the classification performance of both the ESSP-CNNS model and the individual CNN models. This comprehensive evaluation allows for a thorough assessment of the effectiveness of the ESSP-CNNS model in accurately classifying different severity grades of DR in the IDRiD dataset, while also comparing its performance to that of the other models.

Based on the analysis of the confusion matrices (Fig. 8), it is evident that the ESSP-CNNS model outperformed the end-to-end AlexNet, ResNeXt50, and VGG16 models across almost every class of DR. These results clearly demonstrate the superior performance of the ESSP-CNNS model in accurately classifying the different severity levels of DR.

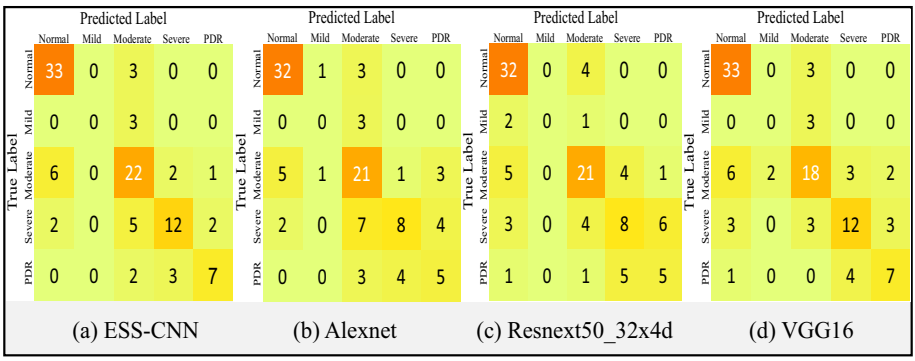

Fig. 8 Confusion matrices of the ESSP-CNNS model, and each individual CNN model in DR grading on IDRiD

Table 5 The results of the ESSP-CNNS model and each individual CNN model on the IDRiD dataset

Method	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
AlexNet (ImageNet)	64.08	87.46	63.16	64.08	62.86	80.62
ResNeXt50 (ImageNet)	64.08	86.94	60.85	64.08	62.14	83.83
VGG16 (ImageNet)	67.96	88.58	66.56	67.96	66.91	84.39
ESSP-CNNS (Proposed)	**71.84**	**88.76**	**69.60**	**71.84**	**70.31**	**86.02**

Table 6 Some hyperparameter values of the CNN models in the fine-tuning process on the Messidor dataset

	Image size	Batch size	Learning rate	Number of epochs	Optimizer
AlexNet	256×256	22	0.00001	50	Adamax
ResNeXt50	256×256	24	0.00001	20	Adamax
VGG16	256×256	26	0.00002	30	Adamax

Additionally, Table 5 presents the evaluation metrics for DR classification, which include the ESSP-CNNS model along with the aforementioned CNN models. Remarkably, the ESSP-CNNS model demonstrates impressive results, achieving higher overall specificity, precision, recall, and F1-score values of 88.76% 69.60%, 71.84%, and 70.31%, respectively. These metrics highlight the high accuracy and effectiveness of the ESSP-CNNS model in accurately classifying DR, surpassing the performance of each individual CNN model.

4.3 Experiments on Messidor dataset

To provide further evidence of the superior performance and effectiveness of the ESSP-CNNS model over end-to-end CNN methods, an additional experiment is conducted on the Messidor dataset. Following the pre-training of three CNN models using BYOL on the EyePACS dataset, the subsequent phase involved fully fine-tuning them on Messidor to construct the proposed ensemble model for DR classification. After the BYOL pre-training, the fine-tuning step included updating the weights of all layers in the pre-trained CNN models by utilizing labeled images from the Messidor dataset. This process allowed the models to adapt their learned features to the specifics of DR classification on the Messidor dataset.

For the fine-tuning process, we adopted a similar approach as with the IDRiD dataset. The Messidor dataset was split into two sets: 80% for training and 20% for testing. Unlike the pre-training phase, no data augmentation techniques were applied during the fine-tuning process. The use of only the original preprocessed data ensured that the models learned from the authentic distribution of Messidor images. A summary of the utilized parameters in the fine-tuning process is presented in Table 6.

Subsequently, the ESSP-CNNS model is constructed after the fine-tuning process. The proposed ESSP-CNNS model is trained using 80% of the Messidor dataset for 80 epochs.

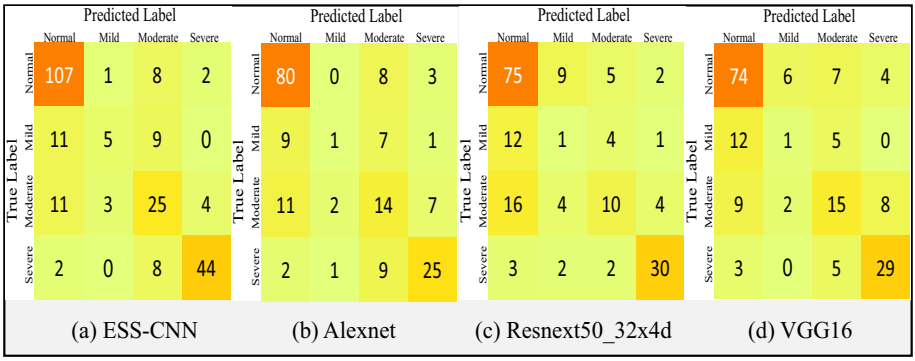

Fig. 9 Confusion matrices of the ESSP-CNNS model, and each individual CNN model in DR grading on Messidor

Table 7 The results of the ESSP-CNNS model and each individual CNN model on the Messidor dataset

Method	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
AlexNet (ImageNet)	66.67	82.97	63.39	66.67	64.25	77.45
ResNeXt50 (ImageNet)	64.44	79.45	62.06	64.44	62.62	77.65
VGG16 (ImageNet)	66.11	82.25	62.68	66.11	64.20	80.66
ESSP-CNNS (Proposed)	**75.42**	**87.13**	**74.70**	**75.42**	**73.99**	**86.54**

The training process involves a batch size of 24, a learning rate of 1e-5, and the Adamax optimizer. The objective of the model is to classify DR into four severity grades. To evaluate the overall performance of the proposed model, the remaining 20% of the Messidor dataset is utilized for testing. The robustness of the ESSP-CNNS model and the individual CNN models on the Messidor dataset is assessed by examining a confusion matrix, as illustrated in Fig. 9.

By analyzing the obtained confusion matrices from the Messidor dataset (Fig. 9), it is evident that the ESSP-CNNS model once again demonstrated superior performance compared to the end-to-end CNN models. In line with the previous findings on the IDRiD dataset, the evaluation outcomes of the ESSP-CNNS model on the Messidor dataset consistently exhibited superior performance compared to the end-to-end method employing the same CNN models (Table 3). The proposed ESSP-CNNS model achieved an impressive test accuracy of 75.42%, demonstrating a significant 8.75% improvement over the highest accuracy attained by the CNN models (66.67%) in DR classification. Additionally, the ESSP-CNNS model exhibited enhanced overall F1-score, recall, specificity, and precision values compared to the end-to-end model for grading the severity of DR on the Messidor dataset (Table 7). This consistent superiority of the ESSP-CNNS model reinforces its effectiveness and reliability in achieving higher accuracy and better classification results across different datasets and also accurately classifying and grading DR.

Additionally, the effectiveness of the ESSP-CNNS model in DR grading on both the IDRiD and Messidor datasets is assessed by examining the ROC curves. The ROC analysis

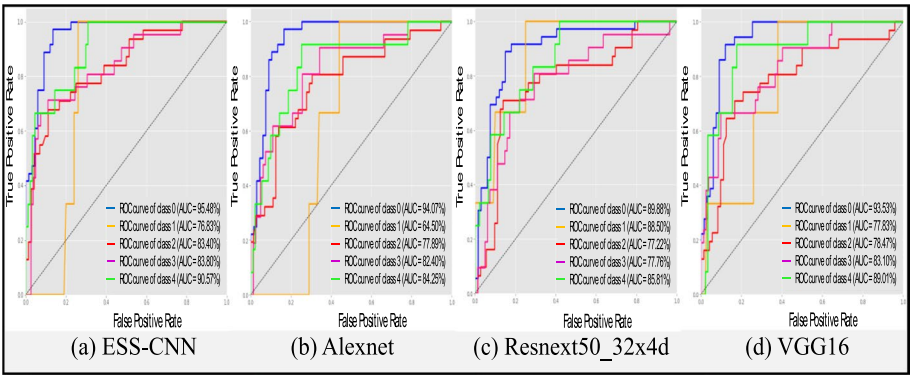

Fig. 10 The ROC curves for individual classes in the DR grading task using the IDRiD dataset are compared between the ESSP-CNNS model and the CNN models

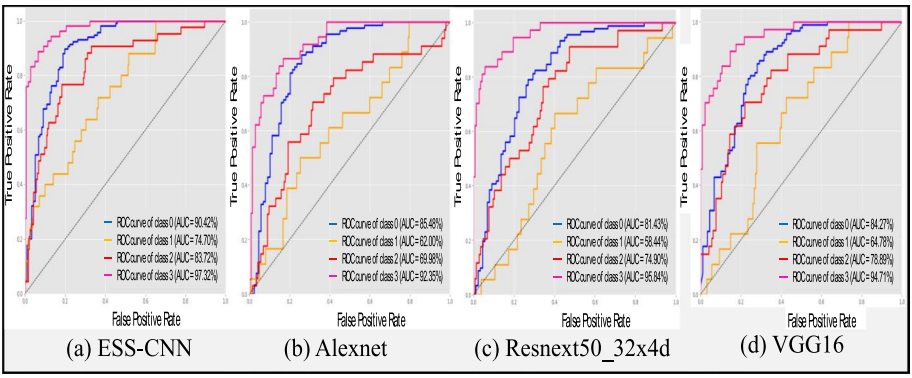

Fig. 11 The ROC curves for individual classes in the DR grading task using the Messidor dataset are compared between the ESSP-CNNS model and the CNN models

serves as a measure of the classifier's capability to distinguish among different classes, with the area under the curve (AUC) score summarizing its performance. The ROC curves for the ESSP-CNNS model and end-to-end CNN models further confirm the effectiveness of the proposed model. On the IDRiD dataset, the ESSP-CNNS model achieves an overall AUC score of 86.02%, surpassing the AUC scores of the AlexNet (80.62%), ResNeXt50 (83.83%), and VGG16 (84.39%) models. Similarly, on the Messidor dataset, the ESSP-CNNS model obtains the highest overall AUC score of 86.54%, outperforming the AlexNet (77.45%), ResNeXt50 (77.65%), and VGG16 (80.66%) models.

Figure 10 illustrates the ROC curves for the various models on the IDRiD dataset, while Fig. 11 showcases the ROC curves for the same models on the Messidor dataset. These curves provide a visual representation of the model's performance in terms of the true positive rate versus the false positive rate. These figures clearly demonstrate that the ESSP-CNNS model consistently exhibits improved class-wise AUC scores compared to the end-to-end approaches across nearly all classes of DR. This improvement in AUC scores highlights the enhanced performance of the ESSP-CNNS model in accurately classifying and grading DR.

The obtained results demonstrate that pre-training CNN models with BYOL on large, unlabeled datasets (in this case we use the EyePACS dataset) can lead to significant improvements in performance on downstream tasks such as DR classification. This highlights the effectiveness and potential of self-supervised learning methods such as BYOL in practical applications of medical image analysis. Furthermore, the ensemble model employed a combination of predictions from three distinct fine-tuned CNN models to enhance the accuracy of DR classification. By leveraging the diverse capabilities of these three CNN models, the ensemble model achieved more precise and robust predictions compared to any individual model. Additionally, this approach introduced diversity, thereby mitigating the risk of overfitting the training data.

Experimental results show that the ESSP-CNNS model outperformed the individual CNN models on DR classification tasks, achieving higher accuracy and better generalization. This highlights the effectiveness of self-supervised pretraining and combining multiple CNN models using an ensemble approach for improving the accuracy and robustness of medical image analysis tasks such as DR classification.

4.4 Performance comparison with existing work

In order to evaluate the effectiveness and performance of our model, it is crucial to perform a comprehensive comparison with existing studies in the field. This allows us to provide context to our results, validate the novelty of our approach, and gain valuable insights into the prevailing state-of-the-art techniques employed in the grading of DR. Within this subsection, we undertake a comparative analysis, contrasting the results of our model with the findings of pertinent studies in the literature. Our proposed methodology, referred to as ESSP-CNNS, entails the utilization of self-supervised pre-training of the three networks, followed by fine-tuning them to establish an ensemble model for DR grading. The comparison is executed on each of the Messidor and IDRiD datasets, employing the exact same evaluation metrics employed in previous research studies.

4.4.1 IDRiD dataset comparison

The presented ESSP-CNNS methodology is compared against various contemporary state-of-the-art approaches on the IDRiD dataset, as outlined in Table 8. The CF-DRNet [47] is a hierarchical Coarse-to-fine network based on CNNs, designed for grading five stages of DR severity, and involves Coarse and Fine Subnetworks for two-class and multiclass classification of DR respectively. SUNet [48] integrates multi-task learning and a feature blending block to effectively perform simultaneous grading of both diabetic macular edema (DME) and DR, resulting in improved performance on both tasks. CANet [49], an innovative attention network designed for cross-disease analysis, incorporates disease-dependent and disease-specific attention modules to collectively assess both DME and DR, leveraging only image-level supervision. MIE-DR [50] introduces a unique pre-training approach centered around self-supervised learning, utilizing multimodal visual data to significantly enhance the accuracy of DR grading. The SKD [51] employs self-knowledge distillation, CAM-Attention, and a Mimicking Module to achieve efficient and precise DR grading with only image-level supervision.

ESSP-CNNS distinguishes itself as a leading model in DR classification, surpassing the latest benchmarks with a remarkable 3.88% accuracy improvement over the closest competitor, SKD, and outperforming other notable methods introduced in 2022 and 2023. The

Table 8 The performance of the ESSP-CNNS model on the IDRiD dataset is compared to state-of-the-art approaches, with the best result highlighted in bold

Method	Year	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
CF-DRNet [47]	2020	56.19	-	-	-	-
PLDMLT [52]	2023	57.69				
Fine Network [47]	2020	58.33	-	-	-	-
SUNet [48]	2020	65.05	-	-	-	-
CANet [49]	2019	65.10	-	-	-	-
MIE-DR [50]	2022	65.05 ± 1.19	-	-	-	**93.00**
Multiscale features [53]	2022	66.35				
HA-Net [54]	2022	66.41	-	-	-	**-**
SKD [51]	2020	67.96	-	-	-	-
ESSP-CNNS (Proposed)		**71.84**	**69.60**	**71.84**	**70.31**	86.02

precision and recall metrics further highlight its balanced and reliable performance, aligning effectively with evolving DR grading standards. With a precision rate of 69.60% and a recall rate of 71.84%, ESSP-CNNS achieves a commendable equilibrium crucial for accurate DR assessment. The F1-score of 70.31% reinforces its capability to simultaneously minimize false positives and capture true positive cases. Notably, the model's substantial AUC score of 86.02% underscores its proficiency in distinguishing between diverse DR severity levels, solidifying ESSP-CNNS as an advanced and competitive choice within the latest developments in DR classification research.

4.4.2 Messidor dataset comparison

Furthermore, to conduct a thorough assessment of the ESSP-CNNS model's performance, we extend its evaluation to other datasets to assess its generalizability for DR grading. Following that, we conduct a comparison between our approach and other existing models for DR grading on the Messidor dataset, as documented in Table 9. Once again, we assess the

Table 9 The performance of the ESSP-CNNS model on the Messidor dataset is compared to state-of-the-art approaches, with the best result highlighted in bold. The asterisk (*) indicates that the result is obtained from reference [58]

Method	Year	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
GLCM/RF [56]	2016	45.91	-	-	-	52.00
GLCM/SVM [56]	2016	47.00	-	-	-	60.00
Fractal-based [55]	2019	48.30	-	-	-	-
SKD *	2020	60.80	-	-	-	-
CANet *	2019	68.00	-	-	-	-
Expert [57]	2015	68.10	-	-	-	-
WAD-Net [58]	2022	71.20	0.716	0.713	0.713	0.808
Hybrid model [30]	2021	72.33	-	-	-	-
DR framework [59]	2022	73.60	-	-	-	-
ESSP-CNNS (Proposed)		**75.42**	**74.70**	**75.42**	**73.99**	**86.54**

effectiveness of our proposed method by comparing it with two DL approaches, CANet [49] and SKD [51], on the Messidor dataset. It is important to highlight that these models are trained exclusively using image-level supervision. Additionally, we assess the effectiveness of conventional ML techniques, such as utilizing Fractal-based features [55] and combining GLCM features with various classifiers [56]. Moreover, we conduct a comparative analysis between our model and an expert [57]. Lastly, we conduct a comparison between ESSP-CNNS and WAD-Net [58], a comprehensive framework for weakly-supervised domain adaptation in grading severity of DR. WAD-Net effectively addresses the difficulties associated with domain adaptation and interpretable multi-instance learning.

In this comparative analysis on Messidor dataset, ESSP-CNNS asserts its superiority in DR classification, achieving an impressive accuracy of 75.42%, surpassing all other methods in the table. Noteworthy is its balanced performance, with a precision rate of 74.70% and a recall rate of 75.42%, crucial for accurate DR grading. The F1-score of 73.99% and an AUC of 86.54% further underscore its proficiency in distinguishing between various DR severity levels. WAD-Net, a 2022 model, also demonstrates competitive performance with an accuracy of 71.20%. Additionally, the DR framework, a late 2022 model achieving an accuracy of 73.60%, stands as another notable contribution to the evolving landscape of DR classification methodologies. These specific results emphasize the robustness of ESSP-CNNS, showcasing notable improvements over the benchmark methods.

In summary, our comprehensive comparative analysis underscores the effectiveness and novelty of the ESSP-CNNS model, showcasing its potential for accurate and interpretable DR grading across diverse datasets.

4.5 Discussion

In this study, we proposed a methodology for grading the severity of DR using color fundus images. Our approach utilized self-supervised learning, specifically the BYOL (Bootstrap Your Own Latent) technique, for pre-training neural networks on a large dataset of unlabeled images. We also employed deep ensemble learning techniques to construct a robust model for DR grading. Our methodology consisted of two main components: pre-training the neural networks using BYOL and building an ensemble model for DR grading. The use of self-supervised learning techniques, such as BYOL, has shown promising outcomes in various image classification tasks. By pre-training the neural networks on the EyePACS dataset without using its labels, which consist of 35,126 fundus images, our methodology aimed to learn meaningful representations of the input images. This pre-training phase was crucial for developing a robust model capable of accurately analyzing color fundus images and predicting DR grading. To evaluate the performance of our methodology, we utilized two public datasets: IDRiD and Messidor. The IDRiD dataset contains 516 fundus images categorized into five distinct severity levels, while the Messidor dataset comprises 1200 images categorized into four grades of severity. These datasets provided a diverse range of examples for training and testing our model. Pre-processing of fundus images played a vital role in improving the quality of images and enhancing the performance of our model. Our pre-processing technique involved several steps, to enhance visibility and contrast levels. The effectiveness of our pre-processing techniques was demonstrated through enhanced visibility and improved image quality, as shown in Fig. 6. The proposed framework, named ESSP-CNNs (Ensemble of Self-Supervised Pretrained CNNs), utilized multiple CNN architectures, including AlexNet, VGG16, and ResNeXt50. These architectures were selected based on their established reputation and consistent success in various

medical image classification scenarios. The ensemble model constructed using these CNNs demonstrated promising results in DR grading. The pre-training of the CNN models using the BYOL method on the EyePACS dataset was a crucial step in our methodology. BYOL is a form of contrastive learning that emphasizes the creation of meaningful representations using positive pairs. However, we made specific modifications to the augmentation techniques used in BYOL to make them suitable for fundus image processing. The pre-training process aimed to obtain a meaningful representation of the input images that could be applied to downstream tasks, such as DR classification. The BYOL framework involved the interaction between online and target networks, utilizing encoder, projector, and predictor components. The loss function was calculated based on the obtained projections of the target projector and predictions of the online predictor module. Introducing the predictor module into the online network and utilizing the moving average of online network parameters as the target network, facilitated the generation of informative representations. Our proposed methodology and framework have several strengths. The utilization of self-supervised learning and deep ensemble techniques allowed us to leverage the strengths of these methods and improve the accuracy of DR grading. The pre-training phase using BYOL on a large dataset of unlabeled images helped develop meaningful representations and enhance the performance of the model. The ensemble of CNN models further improved the robustness and generalization capabilities of our approach.

5 Limitations

However, there are some limitations to consider. Firstly, the datasets used in this study suffer from class imbalance, where the number of samples in each class is uneven. This can affect the performance of the classification models, as they may prioritize the majority class and struggle to accurately identify the minority class. Additionally, errors in grading can introduce inaccuracies in the ground truth labels, leading to potential biases in the training and evaluation process. These factors need to be carefully addressed to ensure the reliability and generalizability of our findings. Furthermore, the computational resources available for this study were limited, which impacted the scope and scale of the experiments conducted. The training and evaluation processes could have benefited from more extensive computational power, allowing for larger model architectures, longer training times, and more thorough hyperparameter optimization. Additionally, the use of more advanced techniques such as ensembling or transfer learning could have been explored given sufficient computational resources. Overcoming these resource restrictions would enable more comprehensive investigations and potentially lead to improved performance and insights.

6 Future work

In future research, addressing the limitations identified in this study should be prioritized. Firstly, efforts should be made to address the issues related to imbalanced datasets. Techniques such as under-sampling the majority class, oversampling the minority class, or utilizing data augmentation methods can help mitigate the impact of class imbalance on model performance. Moreover, gathering larger and more diverse datasets can provide a better representation of real-world scenarios and further improve the generalizability of the proposed methodology. To enhance the accuracy and reliability of the grading process, future work should focus on

refining the ground truth labels and minimizing errors in grading. This can be achieved through the involvement of multiple experts for label verification or by utilizing consensus-based grading systems. Implementing quality control measures during the data collection and labeling processes can also contribute to reducing the impact of grading errors on the overall performance of the models. Lastly, future studies should aim to overcome resource restrictions by leveraging high-performance computing resources or cloud-based solutions. This would enable researchers to explore larger and more complex models, conduct more extensive experiments, and thoroughly investigate the potential benefits of advanced techniques such as ensembling or transfer learning. Additionally, investigating the use of alternative architectures or optimizing existing models to achieve better trade-offs between computational requirements and performance can be valuable avenues for future exploration.

7 Conclusions

In conclusion, our study presents an effective methodology for grading the severity of DR using color fundus images. The combination of self-supervised learning and deep ensemble techniques, along with robust preprocessing and well-established CNN architectures, contribute to the accuracy and reliability of the grading process. The proposed methodology encompassed three main components: fundus images preprocessing with proposed method, network pre-training with BYOL and the creation of an ensemble model. Through comprehensive experiments on the Messidor and IDRiD datasets, the effectiveness of the approach was demonstrated, achieving highly accurate predictions of DR severity grades.

The framework, ESSP-CNNs, harnessed well-established CNN architectures such as AlexNet, VGG16, and ResNeXt50 for the task of DR severity grading. The findings of this study hold great promise, offering a potential solution for automating the grading of DR and providing valuable support to healthcare professionals in diagnosing and managing this sight-threatening condition. By leveraging the power of deep learning and self-supervised learning techniques, the proposed methodology opens new avenues for enhancing the efficiency and accuracy of DR diagnosis in clinical settings.

Data availability We have used public datasets and they are available on the websites.

References

1. Saleh E et al (2018) Learning ensemble classifiers for diabetic retinopathy assessment. Artif Intell Med 85:50–63
2. Heng L et al (2013) Diabetic retinopathy: pathogenesis, clinical grading, management and future developments. Diabet Med 30(6):640–650
3. Tsiknakis N et al (2021) Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. Comput Biol Med 135:104599
4. Shanthi T, Sabeenian R (2019) Modified Alexnet architecture for classification of diabetic retinopathy images. Comput Electr Eng 76:56–64
5. Moss SE et al (1985) Comparison between ophthalmoscopy and fundus photography in determining severity of diabetic retinopathy. Ophthalmology 92(1):62–67
6. Lim G et al (2020) Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review. Eye Vision 7(1):1–13
7. Krause J et al (2018) Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 125(8):1264–1272

8. Porwal P et al (2020) Idrid: Diabetic retinopathy–segmentation and grading challenge. Med Image Anal 59:101561
9. Teo ZL et al (2021) Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology 128(11):1580–1591
10. Bodapati JD et al (2020) Blended multi-modal deep convnet features for diabetic retinopathy severity prediction. Electronics 9(6):914
11. Luo Y et al (2020) Retinal image classification by self-supervised fuzzy clustering network. IEEE Access 8:92352–92362
12. Stolte S, Fang R (2020) A survey on medical image analysis in diabetic retinopathy. Med Image Anal 64:101742
13. Jing L, Tian Y (2020) Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans Pattern Anal Mach Intell 43(11):4037–4058
14. Litjens G et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
15. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
16. Hu X et al (2021) Contrastive learning based on transformer for hyperspectral image classification. Appl Sci 11(18):8670
17. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR
18. Higashi T et al (2023) Estimation of degradation degree in road infrastructure based on multimodal ABN using contrastive learning. Sensors 23(3):1657
19. Grill J-B et al (2020) Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 33:21271–21284
20. Fraz MM et al (2012) An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Trans Biomed Eng 59(9):2538–2548
21. Salamat N, Missen MMS, Rashid A (2019) Diabetic retinopathy techniques in retinal images: A review. Artif Intell Med 97:168–188
22. Islam MR et al (2022) Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. Comput Biol Med 146:105602
23. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
24. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
25. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR, vol 97, pp 6105–6114
26. Huda SA, Ila IJ, Sarder S, Shamsujjoha M (2019) An improved approach for detection of diabetic retinopathy using feature importance and machine learning algorithms. In: 2019 7th International Conference on Smart Computing & Communications (ICSCC). IEEE
27. Ramasamy LK et al (2021) Detection of diabetic retinopathy using a fusion of textural and ridgelet features of retinal images and sequential minimal optimization classifier. PeerJ Comput Sci 7:e456
28. Mahmoud MH, Alamery S, Fouad H, Altinawi A (2021) An automatic detection system of diabetic retinopathy using a hybrid inductive machine learning algorithm. Personal Ubiquit Comput 1–15
29. Kaushik H et al (2021) Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models. IEEE Access 9:108276–108292
30. Gangwar AK, Ravi V (2021) Diabetic retinopathy detection using transfer learning and deep learning. in Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1. Springer
31. Berbar MA (2022) Diabetic retinopathy detection and grading using deep learning. Menoufia J Electron Eng Res 31(2):11–20
32. Hardas M et al (2022) Retinal fundus image classification for diabetic retinopathy using SVM predictions. Physical Eng Sci Med 45(3):781–791
33. Zhang C, Lei T, Chen P (2022) Diabetic retinopathy grading by a source-free transfer learning approach. Biomed Signal Process Control 73:103423
34. Kobat SG et al (2022) Automated diabetic retinopathy detection using horizontal and vertical patch division-based pre-trained DenseNET with digital fundus images. Diagnostics 12(8):1975
35. Qummar S et al (2019) A deep learning ensemble approach for diabetic retinopathy detection. Ieee Access 7:150530–150539
36. Sikder N et al (2021) Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. Symmetry 13(4):670

37. Cuadros J, Bresnick G (2009) EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. J Diabetes Sci Technol 3(3):509–516
38. Decencière E et al (2014) Feedback on a publicly distributed image database: the Messidor database. Image Anal Stereol 33(3):231–234
39. APTOS 2019 Blindness Detection. Available online: https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed 1 May 2022
40. Manju R, Koshy G, Simon P (2019) Improved method for enhancing dark images based on CLAHE and morphological reconstruction. Procedia Comput Sci 165:391–398
41. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inform Process Syst 25 (NIPS 2012).
42. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
43. Kumar A et al (2016) An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J Biomed Health Inform 21(1):31–40
44. Girshick R, Donahue J, Darrell T, Malik J, UC Berkeley, ICSI (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. ICCS
45. Chen Y et al (2019) Deep learning ensemble for hyperspectral image classification. IEEE J Sel Top Appl Earth Observ Remote Sens 12(6):1882–1897
46. Powers DM (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061
47. Wu Z et al (2020) Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network. Artif Intell Med 108:101936
48. Tu Z, Gao S, Zhou K, Chen X, Fu H, Gu Z, Cheng J, Yu Z, Liu J (2020) SUNet: a lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE
49. Li X et al (2019) CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE Trans Med Imaging 39(5):1483–1493
50. Hervella AS et al (2022) Multimodal image encoding pre-training for diabetic retinopathy grading. Comput Biol Med 143:105302
51. Luo L, Xue D, Feng X (2020) Automatic diabetic retinopathy grading via self-knowledge distillation. Electronics 9(9):1337
52. Liu H, Huang C (2023) PLDMLT: multi-task learning of diabetic retinopathy using the pixel-level labeled fundus images. Comput Mater Contin 76(2):1745–1761. https://doi.org/10.32604/cmc.2023.040710
53. Muhammad J, Aliyu HO, Bashir SA, Mohammed DA (2022) Diabetes retinopathy severity grading using multi-scale image pyramid techniques. http://repository.futminna.edu.ng:8080/jspui/handle/123456789/19076
54. Shaik NS, Cherukuri TK (2022) Hinge attention network: A joint model for diabetic retinopathy severity grading. Appl Intell 52(13):15105–15121
55. Alzami F, Megantara RA, Fanani AZ (2019) Diabetic retinopathy grade classification based on fractal analysis and random forest. In: 2019 International Seminar on Application for Technology of Information and Communication (iSemantic). IEEE, pp 272–276. https://doi.org/10.1109/ISEMANTIC.2019.8884217
56. Labhade JD, Chouthmol L, Deshmukh S (2016) Diabetic retinopathy detection using soft computing techniques. In: 2016 international conference on automatic control and dynamic optimization techniques (ICACDOT). IEEE, pp 175–178. https://doi.org/10.1109/ICACDOT.2016.7877573
57. Seoud L, Chelbi J, Cheriet F (2015) Automatic grading of diabetic retinopathy on a public database. In: Ophthalmic Medical Image Analysis International Workshop. University of Iowa. MICCAI, Munich. https://doi.org/10.17077/omia.1032
58. Cao P et al (2022) Collaborative learning of weakly-supervised domain adaptation for diabetic retinopathy grading on retinal images. Comput Biol Med 144:105341
59. Hou Q et al (2022) Image quality assessment guided collaborative learning of image enhancement and classification for diabetic retinopathy grading. IEEE J Biomed Health Inform 27(3):1455–1466

Authors and Affiliations

Saeed Parsa[1] · Toktam Khatibi[2]

✉ Toktam Khatibi
toktam.khatibi@modares.ac.ir

Saeed Parsa
s_parsa@modares.ac.ir

[1] Department of Industrial Engineering, School of Industrial and Systems Engineering, Tarbiat Modares University, Tehran 14117-13114, Iran

[2] School of Industrial and Systems Engineering, Tarbiat Modares University, Tehran 14117-13114, Iran