mathematics

MDPI

Article

Computer-Aided Diagnosis of Diabetic Retinopathy Lesions Based on Knowledge Distillation in Fundus Images

Ernesto Moya-Albor [1,*,†], Alberto Lopez-Figueroa [1,†], Sebastian Jacome-Herrera [1,†], Diego Renza [2,†] and Jorge Brieva [1,*,†]

[1] Facultad de Ingeniería, Universidad Panamericana, Augusto Rodin 498, Ciudad de México 03920, Mexico; 0234621@up.edu.mx (A.L.-F.); 0237105@up.edu.mx (S.J.-H.)
[2] Facultad de Ingeniería, Universidad Militar Nueva Granada, Carrera 11 101-80, Bogota 110111, Colombia; diego.renza@unimilitar.edu.co
* Correspondence: emoya@up.edu.mx (E.M.-A.); jbrieva@up.edu.mx (J.B.); Tel.: +52-55-5482-1600 (ext. 5210) (E.M.-A.)
† These authors contributed equally to this work.

Abstract: At present, the early diagnosis of diabetic retinopathy (DR), a possible complication of diabetes due to elevated glucose concentrations in the blood, is usually performed by specialists using a manual inspection of high-resolution fundus images based on lesion screening, leading to problems such as high work-intensity and accessibility only in specialized health centers. To support the diagnosis of DR, we propose a deep learning-based (DL) DR lesion classification method through a knowledge distillation (KD) strategy. First, we use the pre-trained DL architecture, Inception-v3, as a teacher model to distill the dataset. Then, a student model, also using the Inception-v3 model, is trained on the distilled dataset to match the performance of the teacher model. In addition, a new combination of Kullback–Leibler (KL) divergence and categorical cross-entropy (CCE) loss is used to measure the difference between the teacher and student models. This combined metric encourages the student model to mimic the predictions of the teacher model. Finally, the trained student model is evaluated on a validation dataset to assess its performance and compare it with both the teacher model and another competitive DL model. Experiments are conducted on the two datasets, corresponding to an imbalanced and a balanced dataset. Two baseline models (Inception-v3 and YOLOv8) are evaluated for reference, obtaining a maximum training accuracy of 66.75% and 90.90%, respectively, and a maximum validation accuracy of 35.94% and 81.52%, both for the imbalanced dataset. On the other hand, the proposed DR classification model achieves an average training accuracy of 99.01% and an average validation accuracy of 97.30%, overcoming the baseline models and other state-of-the-art works. Experimental results show that the proposed model achieves competitive results in DR lesion detection and classification tasks, assisting in the early diagnosis of diabetic retinopathy.

Keywords: knowledge distillation; fundus images; diabetic retinopathy; lesion classification; convolutional neural networks; Inception-v3; Kullback–Leibler divergence; categorical cross-entropy; computer-aided diagnosis disease; health

MSC: 68U10; 68T05; 92C55

Citation: Moya-Albor, E.; Lopez-Figueroa, A.; Jacome-Herrera, S.; Renza, D.; Brieva, J. Computer-Aided Diagnosis of Diabetic Retinopathy Lesions Based on Knowledge Distillation in Fundus Images. *Mathematics* **2024**, *12*, 2543. https://doi.org/10.3390/math12162543

Academic Editor: Yongmin Li

Received: 13 July 2024
Revised: 10 August 2024
Accepted: 13 August 2024
Published: 17 August 2024

1. Introduction

Diabetes is a chronic disease that appears when the organism cannot produce enough insulin or it cannot effectively be used. There are two types of diabetes: type 1, which is not preventable and requires the patient to be provided with external insulin, and type 2, which is the most common (90% of patients) and is preventable [1].

According to the 2021 Diabetes Atlas of the International Diabetes Federation (IDF) [1], 537 million people worldwide, between 20–79 years, have diabetes, and by 2030, the

number could increase to 643 million. Thus, following this trend, by 2045 this will be 783 million people (between 20–79 years). Recently, Professor Andrew Boulton, President of the IDF, declared that diabetes is a "pandemic of unprecedented magnitude", because an it is estimated that 10.5% (around half a billion persons) of the worldwide adult population lives with diabetes [2].

On the other hand, a possible complication of diabetes is diabetic retinopathy (DR), which is caused by elevated glucose concentrations in the blood, resulting in the damage of blood vessels and the abnormal growth of new vessels in the retina [3]. This disease occurs among 30–40% of diabetic patients, making the rise in diabetes prevalence parallel to DR. A recent global analysis estimates that the worldwide prevalence of DR is around 103 million patients, and it will rise to 161 million by 2045 [2].

Diabetic retinopathy has been studied for a long time. For example, Porta et al. [4] presented the history of diabetic eye disease as follows: in 1876, Wilhelm Manz described some pathologies associated with diabetic retinopathy, for example, fibrovascular deterioration of the optic disc, adhesions in the vitreous and retina, detachment of the retina, and hemorrhages in the vitreous. In 1877, Mackenzie described the finding of microaneurysms with vitreous and retinal bleeding. In 1944, Ballantyne and Lowenstein were the first to use the term "diabetic retinopathy". On the other hand, in 2017, Ometto et al. [5] studied the effects of other factors unrelated to DR, how lesions are distributed for discrepancies between decisions made by clinicians, and what the risk model predicts.

Because DR directly affects the blood vessels, it is the main cause of preventable blindness, with a prevalence of 22% in people with diabetes [6]. However, early detection and treatment can prevent vision loss due to diabetes. Thus, there are currently treatments that can considerably reduce the risks of blindness through early diagnosis. For example, retinography is a non-invasive and accessible method that generates high-quality images of the inner and back surface of the eyes through a color camera showing the blood vessels and revealing possible lesions due to DR, which can be seen in Figure 1a, showing a healthy patient.

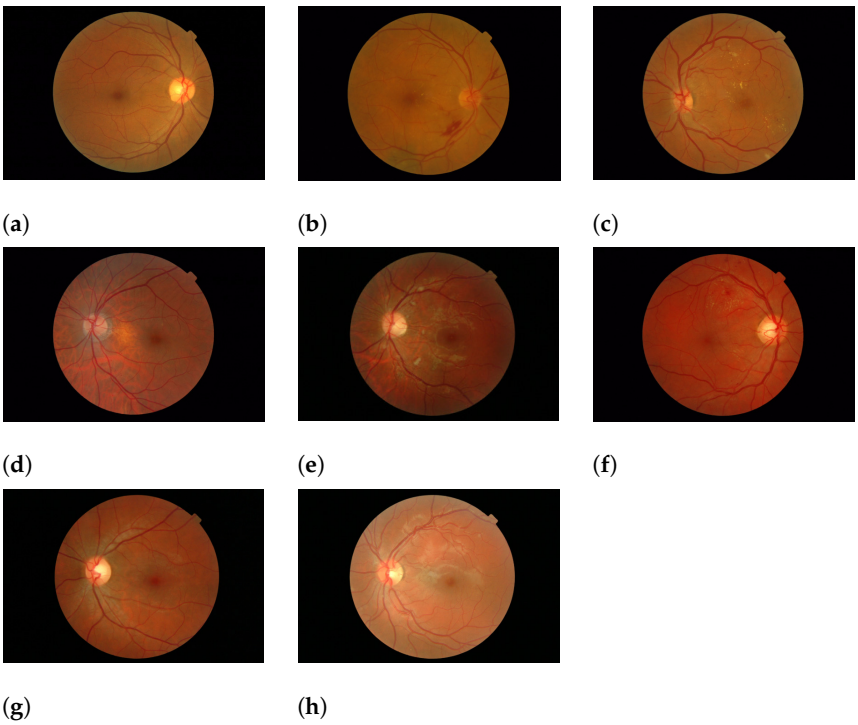


Figure 1. DR lesions examples. (**a**) Healtly patient. (**b**) Hemorrhages. (**c**) Exudates. (**d**) Microaneurysms. (**e**) Cotton wool spots. (**f**) Retinal venous beading. (**g**) Patient with glaucoma. (**h**) Patient with macular degeneration.

Diabetic retinopathy is characterized by microangiopathy in the venous capillaries, such as microaneurysms, microvascular occlusions, arteriovenous short circuits, neovascularization, exudates, and hemorrhages [3]. In addition, retinal venous beading and cotton wool spots can be found. Neovascularization consists of the generation of new blood vessels due to a prolonged absence of blood in determined regions of the eye, which generally have irregular, weak walls that are prone to rupture. These new blood vessels are identified as prominences in the contour of the optic nerve. Retinal hemorrhages (Figure 1b) appear when the blood vessels weaken and begin to bleed, producing irregular reddish spots, which cause either partial or total blindness due to the light occlusion of the photoreceptors. Exudates (Figure 1c) are presented when substances rich in proteins spread into the eye from the vessels, which form brilliant deposits of different sizes, covering small spots to large areas. Microaneurysms (Figure 1d) appear as small reddish spots in different areas, which could easily be mistaken for small hemorrhages, but are induced by dilations of the blood vessels and not by blood draining into the vitreous body. Therefore, in diagnosis, these structures are used to evaluate the DR grade [3]. Cotton wool spots (Figure 1e) are presented when there are tight retinal arteries due to chronic hypertension and atherosclerosis [7]. Venous beading (Figure 1f) is characterized by the beaded appearance of the retinal venules due to long hypoxia cycles and a changing lumen's dilation and constriction.

The damage that DR causes to small retinal vessels also increases the risk of glaucoma (GLA), cataracts, and other eye problems. GLA (Figure 1g) is an eye condition that damages the optic nerve. This damage is caused by higher-than-normal pressure in the eye. Open-angle glaucoma is characterized by irregular blind spots in peripheral or central vision, often occurring in both eyes or tunnel vision in very advanced stages. The above figure shows a fundus image of a patient with glaucoma.

Furthermore, age-related macular degeneration (AMD) is an important cause of irreversible blindness worldwide. It is a multi-factorial disease with cuases such as age, race, gender, diet, cardiovascular illnesses, and genetic risk factors. In early stages, AMD presents the appearance of extracellular deposits in the Bruch's membrane. In late stages, AMD is characterized by the appearance of geographic atrophy (GA), where there is photoreceptor loss, retinal thinning, and a progressive, irreversible loss of central vision. In many cases, GA could evolve into choroidal neovascularization, a formation of new blood vessels in the Bruch's membrane, which causes hemorrhages that lead to further loss of vision [8]. Figure 1h shows a patient with macular degeneration.

Due to DR being characterized by several retinal lesions, it is necessary to develop new multi-class image classification algorithms and address the imbalance between the number of images per class found in most datasets. In addition, these algorithms must present an accurate classification to support medical diagnosis. Therefore, in this study, we present a knowledge distillation (KD) model using a new combination of loss functions. First, we used a pre-trained teacher model to distill the target dataset. Then, a student model was trained on the distilled dataset to match the performance of the teacher model. For evaluation, the proposed method was tested on two datasets: one imbalanced dataset containing 1000 images across 39 classes and a second balanced dataset with 3592 images into eight categories. Experimental results show that the DR lesions classification system overcomes the accuracy of some state-of-the-art methods. The contributions of this work are summarized as follows:

- Knowledge Distillation Model and Combination of Loss Functions:
  - We present a knowledge distillation model and develop a new combination of the Kullback–Leibler divergence and categorical cross-entropy loss functions to address the problem of sample number imbalance in the dataset.
  - To the best of our knowledge, this combination of loss functions is the first to be used.
- Teacher–Student Training Framework:

  - We utilize a pre-trained teacher model to distill the dataset, providing a refined set of data for training.
  - We train a student model on the distilled dataset to achieve performance comparable to the teacher model, ensuring effective knowledge transfer and maintaining a compact and efficient final model.
- Dataset and Methodology:
  - We perform an analysis of different CNN architectures by testing and evaluating them to determine the best network that would generate the best classification results.
- Experimental Setup and Results:
  - Experimental results using a public and in-house dataset demonstrates that the proposed knowledge distillation model significantly reduces overfitting against baseline models. In addition, the distilled model shows a robust generalization and overcomes some state-of-the-art methods.

The rest of the paper is divided as follows: Section 2 presents a review of some recent works on DR detection, DR grading, and DR lesion classification. Section 3 describes the fundus image datasets used, the CNN architectures, and the loss functions utilized to classify DR lesions. Section 4 details the proposed knowledge distillation-based classification model for DR lesions. Next, Section 5 presents the experimental setup, evaluation metrics, DR lesions classification performance analysis, and comparison against other works in the literature, followed by a discussion of the results. Finally, the conclusions are presented in Section 6, followed by future work.

2. Related Work

Various methods have been proposed to detect DR in fundus images. For example, extracting blood vessels to use their information for DR diagnosis. Thus, Sundaram and et al. [9] presented a hybrid segmentation method of blood vessels by fusing morphological operators, vessel enhancement at multiple scales, and the bottom hat algorithm. In comparison, Colomer et al. [10] proposed a blood vessel segmentation approach using sparse representations and dictionary-based learning. The proposed method was evaluated using RGB and intensity fundus images via two public databases.

Other works use a machine learning approach to generate DR grading related to the severity of the illness. For example, in [11], Sharif and Shah applied feature extraction to classify the grading of DR. Contrast-limited adaptive histogram equalization (CLAHE), Gaussian curve fitting, and independent component analysis (ICA) were used as a preprocessing stage, and a multi-class Gaussian Bayes classifier and multi-support vector machine (SVM) were applied to classify the grade of DR. Kaur and Mittal [12] reported a lesion segmentation method by a means and iterative clustering approach, taking into account the heterogeneity and bright and weak edges, which are used to grade nonproliferative diabetic retinopathy.

On the other hand, other works directly classified retina lesions to identify DR through generic classifiers. For example, in [13], Estudillo-Ayala et al. presented a method using multi-directional fractional-order Gaussian filters, the differential evolution algorithm, and the Kittler thresholding method to detect microaneurysms and hemorrhages in fundus images. Then, the extracted structures were classified by means of an SVM. In another work, Wang et al. [14] proposed a semi-supervised method using a series of both healthy and ill fundus images without annotation. The proposed method separated background from blood vessels and background noise applying image processing steps. Furthermore, background noise was modulated as a stochastic variable through a mixture of Gaussian filters for normal and abnormal images, respectively. Thus, both the image background and the noise of the background noise identification were joined in a whole model to detect the lesions, for example, haemorrhages, exudates, and cotton wool spots. In [15], Biswas et al. developed an intelligent system for the detection of DR using an SVM to detect foveal avascular zones and microaneurysms. In [16], an automatic detection method of exudates

using fundus images was presented. This method involved the following steps: shifting color correction, optic disc removal, and exudates detection. Afrin and Shill [17] presented a DR grading system based on retinal lesion detection (e.g., microaneurysms, blood vessels, and exudates). Similarly, in [18], Adal et al. reported an automatic detection method of lesions based on longitudinal retinal changes caused by small red lesions in normalized images. This method reduces illumination variations and improves the contrast of small features in the retina. Sidibé et al. [19] used a sparse coding approach for retinal lesions classification, specifically for detecting fundus images with exudates or drusen and images without lesions. They applied a linear SVM over the sparse coded features, improving the bag-of-visual-word approach. In [20], Ghasemi Falavarjan et al. reported an advance in analyzing ultra-wide-field retinal images, enabling accurate measurements of peripheral retinal lesions. Finally, some less recent techniques are those of [21,22]. The first work presented a method to detect exudates by applying adaptive histogram equalization and thresholding for area calculation. Then, micro-aneurysms were detected using top-hat and bottom-hat transforms. Micro-aneurysms were finally classified using Otsu thresholding and the Hough transform. The second method used the m-Mediods-based modeling approach together with the Gaussian mixture model.

Although the above methods obtained promising results, they are not suitable for classifying multiple DR lesions. Thus, CNNs have been widely used for multiclass problems, improving the ability of DR lesion diagnosis. Ashraf et al. [23] focused on identifying red lesions of small size that are less discriminative for early DR detection. A modified ResNet50 model was used as a CNN as an alternative to transfer learning issues such as over-fitting, domain adaptation, and performance degradation. In another work, Alsubai et al. [24] proposed a quantum-based deep CNN approach because of its parallel computing capacity for DR classification in fundus images. Priya et al. [25] and Bilal et al. [26] presented reviews of current methods of detecting non-proliferative diabetic retinopathy, exudates, hemorrhages, and microaneurysms using deep, intelligent systems. Abdelmaksoud et al. [27] presented a comprehensive computer-aided diagnosis using deep learning for DR classification. First, a prepossessing stage removes noise, performs quality enhancements, and resizes fundus images for standardization. Next, the system classifies healthy and DR patients. Finally, blood vessels, exudates, hemorrhages, and microaneurysms are extracted. Hassan et al. [28] evaluated semantic segmentation, scene analysis, and hybrid DL systems for retinal lesion detection using optical coherence tomography (OCT) images. Some lesions extracted were intra-retinal fluid, drusen, hard exudates, chorioretinal anomalies, and subretinal fluid. In [29], Playout et al. presented a convolutional multi-task architecture with reinforcing learning for red and bright lesions' simultaneous segmentation.

In contrast to traditional deep learning methods, which are resource-intensive, containing large amounts of parameters, and are not suitable for direct clinical applications, this work proposes the use of the knowledge distillation (KD) algorithm, which is considered a dataset reduction method that focuses on the creation of more compact and representative sets of the original dataset since it takes into account the most important characteristics of images and saves their information. This method not only reduces computational complexity and storage requirements but can also improve the efficiency and generalization of models trained on previously manipulated sets. Thus, Luo et al. [30] proposed a knowledge distillation approach to a large teacher network, which later guides a student network using labels of the images for DR grading. In addition, the authors incorporated class activation mapping (CAM), including an attention module and a mimicking stage. Abbasi et al. [31] presented a knowledge distillation approach to transfer knowledge of a complex model to a simple model with fewer parameters. The proposed method uses unlabeled data to transfer DR knowledge and incorporate it into a low-resource embedded system. A VGG network was used to train the teacher model, and then a simple model (student model) was trained based on the teacher's knowledge. The proposed approach was used for binary DR classification. In another work, Gao et al. [32] reported a collaborative learning-based knowledge distillation framework, which was used to enhance

fundus images, increasing the detecting retinopathy accuracy and reducing the running time of the model. The collaborative strategy allowed us to incorporate several student models of different scales and architectures to extract relevant diagnosis information. Next, a transfer learning approach was applied. In [33], Islam et al. proposed a knowledge distillation approach where two models were fused (ResNet152V2 and Swin transformer) as teacher models. Then, the knowledge learned from the teacher model was used to train the lightweight student model (Xception) for the diabetic retinopathy classification task. In addition, a pre-processing stage, formed by several methods such as image denoising, ROI selection, thresholding, bounding box selection, cropping, unsharp masking filter, and gamma transformation, were applied to the APTOS and IDRiD datasets for both binary and multi-class classification. Ju et al. [34] proposed a DL strategy to learn from long-tailed fundus datasets, where a hybrid knowledge distillation method was used for DR lesion classification. In [35], Wang et al. presented a knowledge distillation approach focusing on the transfer of lesion knowledge applied to DR lesion segmentation. Finally, Salguero et al. [36] analyzed something similar to what we are proposing in this paper, in which they processed the information with different unsupervised classification methods and relied on natural language processing with the latent semantic analysis method. Their dataset belonged to several images of ovarian cancer pathology. When comparing the results obtained with 40% of the data and with the entire dataset, they saw that the score obtained was 87%.

3. Materials and Methods

In this section, the fundus image datasets used in this work are introduced. Two datasets are presented: the JSIEC1K dataset, containing 1000 images across 39 categories, and an in-house dataset with 3592 images balanced across eight categories. The characteristics, sources, and composition of these datasets are detailed.

Additionally, the fundamentals of convolutional neural networks are explained, including their architecture, key components such as convolutional layers and pooling layers, and the process of training these networks. Following this, two specific CNN architectures are discussed: Inception-v3 and YOLOv8. The Inception-v3 model, with its innovative use of inception modules and the YOLOv8 architecture, known for its real-time object detection capabilities, are described in detail. Finally, the theory behind the loss functions used in the study, specifically the Kullback–Leibler divergence and categorical cross-entropy loss, is presented to provide a comprehensive understanding of the proposed method's theoretical foundations.

3.1. Fundus Image Datasets

For the development of this project, we explored multiple fundus image datasets, where we found that the images tend to be highly consistent across different sources. Thus, the specialized equipment used for capturing these images generally produces very similar outputs, regardless of the specific device (see [37]). However, the queried datasets lacked sufficient images or labels to train deep-learning models. Therefore, in this work, we selected two datasets due to their characteristics described below. The first one includes 1000 images belonging to 39 different categories, but presents an imbalance between them and shows a non-uniform distribution. The second one contains 3592 into eight different categories but maintains a balance between them.

3.1.1. JSIEC1K Dataset

For the initial tests, the JSIEC1K dataset [38] was used, obtained from Kaggle public datasets (https://www.kaggle.com/datasets/linchundan/fundusimage1000/, accessed on 13 February 2024). It is an imbalanced dataset, which is a collection of fundus images consisting of a total of 1000 images divided into 39 different categories. It is important to mention that the distribution of images is not uniform across all categories, as there are some with a significantly higher number of samples compared to others. This imbalance in

the dataset, coupled with the large number of classes being trained with so few samples, may affect the test results. The images come from the Joint Shantou International Eye Center (JSIEC) located in Guangdong Province, China. They are a small portion of a larger set consisting of 209,494 fundus images. The main objective of this dataset is to be used for training and testing deep learning platforms, which are a type of artificial intelligence. Table 1 shows the classes and number of images per class of the JSIEC1K dataset.

Table 1. Classes and the number of images in each class of the JSIEC1K dataset.

Class	#	Class	#	Class	#
Normal	38	ERM	26	Vitreous particles	14
Tessellated fundus	13	MH	23	Fundus neoplasm	8
Large optic cup	50	Pathological myopia	54	Massive hard exudates	13
DR1	18	Possible glaucoma	13	Yellow-white spots-flecks	29
DR2	49	Optic atrophy	12	Cotton-wool spots	10
DR3	39	Severe hypertensive retinopathy	15	Vessel tortuosity	14
BRVO	44	Disc swelling and elevation	13	Chorioretinal atrophy	15
CRVO	22	Dragged Disc	10	Preretinal hemorrhage	10
RAO	16	Congenital disc abnormality	10	Fibrosis	10
Rhegmatogenous RD	57	Retinitis pigmentosa	22	Laser Spots	20
CSCR	14	Bietti crystalline dystrophy	8	Silicon oil in eye	19
VKH disease	14	Peripheral retinal degeneration	14	Blur fundus without PDR	114
Maculopathy	74	Myelinated nerve fiber	11	Blur fundus with suspected PDR	45

3.1.2. In-House Dataset

On the other hand, an in-house and balanced dataset was also used to perform the tests. It consists of images from the "Messidor-2" dataset [39,40], which was downloaded from Kaggle public datasets (https://www.kaggle.com/datasets/geracollante/messidor2, accessed on 13 February 2024), containing 1748 fundus images of 1440 × 960 and 2240 × 1488 pixels. Out of the 1748 images, 1705 images were grouped into eight groups: macular degeneration (55), exudates (149), hemorrhages (82), cotton wool spots (56), microaneurysms (156), glaucoma (69), venous beading (59), and healthy patients (449). Later, a data augmentation strategy using the Python package Augmentor was performed in order to generate a balance set with respect to the number of healthy patient images, providing 3592 images (449 for each group).

In contrast to the JSIEC1K dataset, this dataset shows less diversity but a more balanced approach. Table 2 shows the classes and number of images per class of the in-house dataset.

Table 2. Classes and the number of images in each class of the in-house dataset.

Class	#	Class	#	Class	#
Beading	449	Macular Degeneration	449	Exudate	449
Glaucoma	449	Hemorrhages	449	Cotton-wool spots	449
Microaneurysms	449	Healthy	449		

3.2. Convolutional Neural Networks

A CNN is composed of so-called convolutional layers, which are in charge of extraction of local features from a set of images by applying different convolution kernels to filters. As a result of these convolutions, features maps are obtained. Thus, a convolution layer containing k kernels could detect K different features (feature maps) after being trained. In Figure 2, a convolutional layer and its corresponding feature maps are presented using a gray-scale image and a bank of K filters.

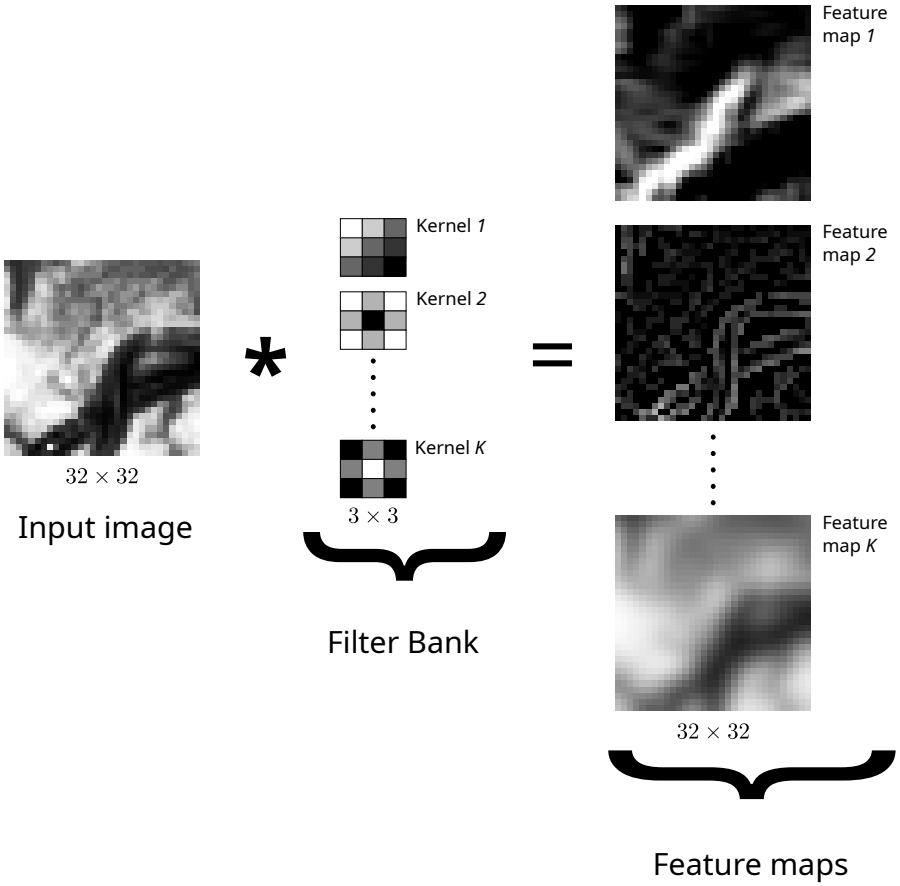


Figure 2. Example of a convolutional layer. A filter bank of k filters generates the corresponding k feature maps.

In a nutshell, each feature map F_k with $k = 1, 2, \ldots, K$ could be mathematically described as is shown in Equation (1):

$$F_k = \sum_n W_k[n] * I[n], \tag{1}$$

where $I[n]$ is a multi-spectral image, n-th is the color band number, $W_k[n]$ is the sub-kernel to the band n, and $*$ represents the 2D convolution operator.

Furthermore, a pooling layer, or downsampling layer, is commonly incorporated at the output of each convolutional layer. Its dimension reduction decreases the number of neurons in the convolutional network and reduces overfitting and computational complexity.

Mathematically, the pooling layer slides a two-dimensional filter over each feature map. Thus, for a feature map of dimension $m \times n$, the output of the pooling layer will be $(m - f)/s + 1 \times (n - f)/s + 1$, where f is the size of the filter and s is stride length.

The pooling layers could be classified into three categories: a max-pooling layer, a min-pooling layer, and an average-pooling layer, where the max-pooling layer is the most commonly used. The max-pooling layer is used to reduce the dimension of the feature maps and to improve robustness to changes in the position of the feature in the image or local translation invariance. It is reached by selecting the maximum response in each patch of each feature map.

A simple CNN architecture used for classification tasks contains several pairs of convolutional-pooling layers. Furthermore, another kind of layer, the fully connected layer, is used for classification tasks. In this layer, each of the outputs in the previous layer is connected with a weight to each neuron in the layer. Finally, the output classification labels in the CNN are generated by an activation function. Figure 3 shows a typical CNN architecture.

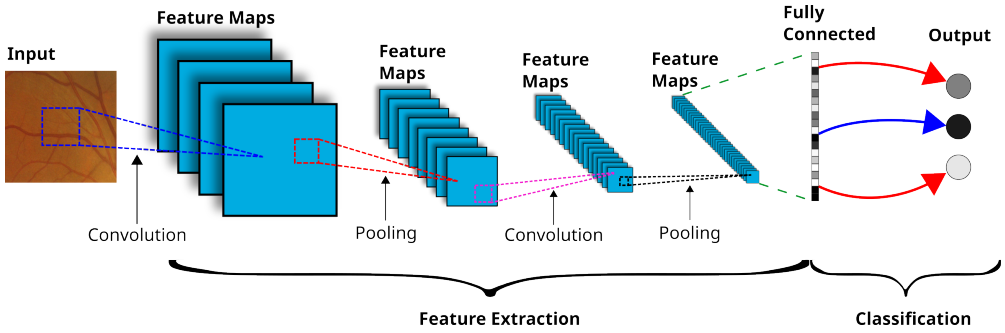


Figure 3. Typical CNN architecture.

Training a CNN consists of adjusting the unknown parameters of the network, i.e., the weights and biases of all connections. In this regard, the backpropagation algorithm is the most frequently used process to iteratively adjust these weights through the minimization of a loss function, for example, the difference between the estimated classes (outputs) and the references (labels).

On the other hand, there are pre-trained CNN models that could be used in cases where a sufficiently large dataset is not available. This strategy is known as transfer learning, and it allows us to fine-tune the last layers, e.g., last fully-connected layer, instead of training the whole model from scratch.

In transfer learning, the weights of a pre-trained model are transferred to the current model to be trained, in contrast to randomly initializing its weights [41]. Therefore, for many image classification problems, such as medical image classification, this approach is enough. In this regard, VGG16, Inception-v3, and ResNet50 are three of the models with the highest score in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC2014) [42].

3.2.1. Inception-v3

Inception-v1 is a 22-layer network that was started as a module for Googlenet, a CNN developed by researchers at Google, that used deeper nets than needed for image classification. Inception-v3 CNN [43], developed by the same group as the original Inception, is composed of 48 convolutional layers and "Inception modules", which are used to reduce the parameter number and to maintain the efficiency of the network. Furthermore, inception modules allow us to handle information at different scales efficiently, improving the efficiency and performance of image recognition tasks.

In Inception-v3, the authors showed how large convolution kernels can be expressed more efficiently by a series of smaller convolutions, surpassing its ancestor GoogLeNet on the ImageNet benchmark. Thus, Inception-v3 implements three types of Inception modules (Table 3); these modules allow us to replace large convolutions by applying convolutions of different sizes.

Table 3. Inception module types, convolutions applied, equivalent convolution, and the output feature maps in each module.

Inception Module	Convolutions Applied	Equivalent Convolution	Output Feature Maps
A	$(3 \times 3) \times 2$	(5×5)	288
B	$(3 \times 1) \times 1$ and $(1 \times 3) \times 1$	(3×3)	768
C	$(1 \times 7) \times 1$ and $(7 \times 1) \times 1$	(7×7)	2048

Inception-v3 introduces several improvements over its predecessors, such as the factorization of large convolutions into smaller convolutions and the use of asymmetric convolutions, reducing both the parameter number and the complexity of the model. Figure 4 shows the Inception-v3 architecture.

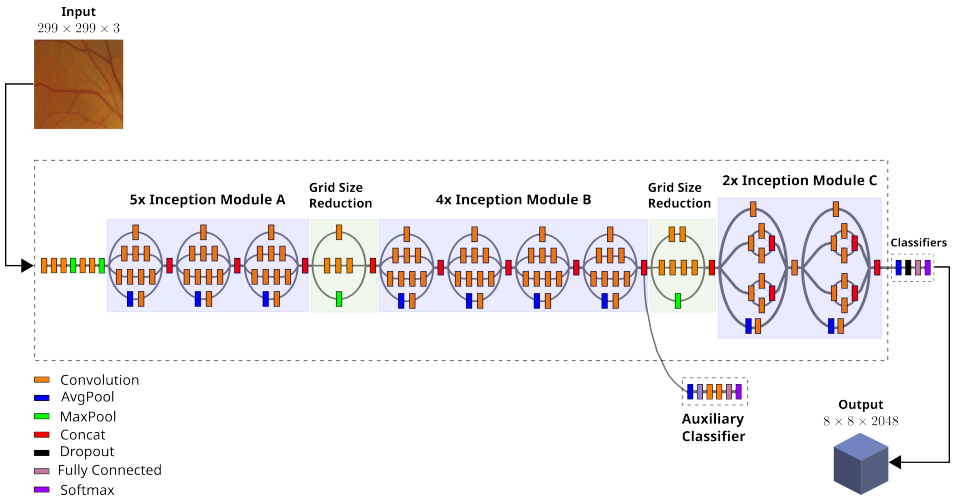


Figure 4. Inception-v3 architecture.

3.2.2. YOLOv8

Another popular CNN architecture is YOLO (you only look once). YOLO is a family of object detection models running in real-time that has been key in computer vision evolution. It was introduced in a paper titled "You Only Look Once: Unified, Real-Time Object Detection", published in 2016 by Joseph Redmon et al. with the YOLOv2 version [44]. In that work, a novel approach to object detection in images was presented, which stood out for its ability to perform real-time detection with a single pass of the convolutional neural network.

In 2022, YOLOv8 was released by Ultralytics, the original YOLO team [45]. YOLOv8 is building upon its past versions and brings in novel features and enhancements in performance and flexibility. It was designed with relevant characteristics such as accuracy, speed, and ease of use for a wide range of tasks, e.g., object detection and image segmentation and classification, even with large amounts of data. YOLOv8 is composed of two parts: the so-called backbone and the head. Figure 5 shows the YOLO architecture.

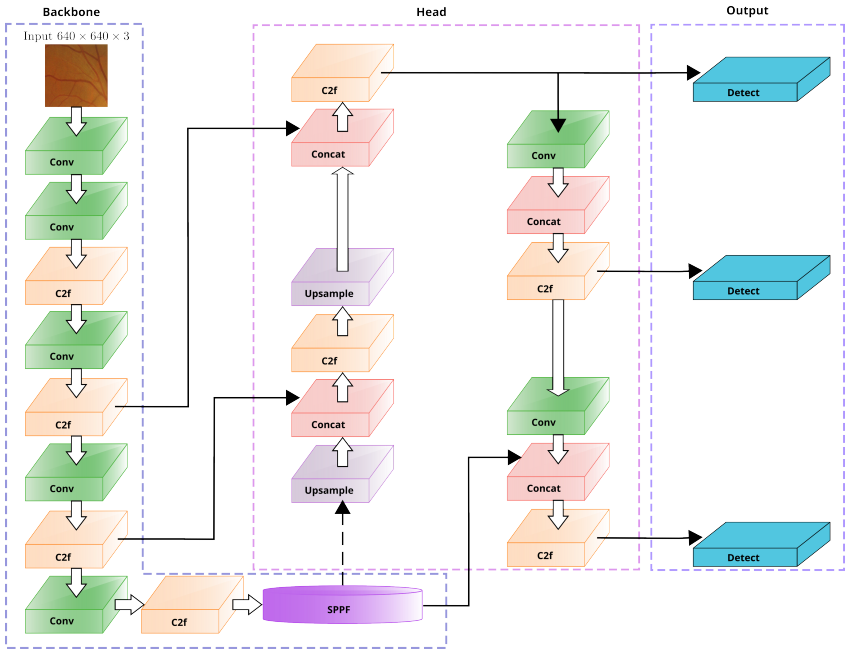


Figure 5. YOLOv8 network architecture. Conv stands for convolution, C2f stands for convolution to feature, Concat means concatenation, SPPF corresponds to spatial pyramid pooling fast, Upsample returns an interpolated version, and Detect performs the object detection task.

A modified CSPDarknet53 architecture was used as the backbone of YOLOv8. It is composed of 53 convolutional layers and cross-stage partial connections, which improves the information flow over layers. On the other hand, the head section is composed of several convolutional layers followed by fully connected layers. These layers predict the bounding boxes, the probabilities of the objects identified, and the objectness punctuation.

As the main feature, a self-attention mechanism is used in the head of YOLOv8. This mechanism focuses on different sections of the image and adjusts the relevance of the features according to their importance on the task performed.

On the other hand, YOLOv8 can detect objects of different sizes in the image using a multi-scale approach. This is possible using a pyramid network composed of multiple layers that detect objects at different scales (see Figure 5).

3.3. Kullback–Leibler Divergence and Categorical Cross-Entropy Loss

We used a combination of two different loss functions, Kullback–Leibler divergence [46] and the categorical cross-entropy loss function [47].

On the one hand, KL divergence allows us to measure the deviation between two probability distributions through their relative entropy. KL divergence is shown in Equation (2).

$$D_{KL}(p(x) \parallel q(x)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}, \tag{2}$$

where $p(x)$ and $q(x)$ are two probability distributions of a random variable x into the discrete sample space X.

It is important to say that KL divergence is not a distance measure. This is because it is not symmetric, i.e., $D_{KL}(p(x) \parallel q(x)) \neq D_{KL}(q(x) \parallel p(x))$.

On the other hand, categorical cross-entropy is a loss function widely used in multi-class classification problems. It is the union of the softmax activation function plus classic cross-entropy. Therefore, it is commonly called softmax loss.

The softmax activation function is given by Equation (3):

$$f(Z)_i = \frac{e^{z_i}}{\sum_j^C e^{z_j}}, \tag{3}$$

where Z is an input vector of C real values, and z_i and $f(Z)_i$ are output vectors, whose elements range between 0 and 1 and sum up to 1. Thus, $f(Z)_i$ consists of C probabilities, which are proportional to the exponential of their real input values.

Finally, the categorical cross-entropy definition is shown in Equation (4):

$$CCE = -\sum_{i}^{C} t_i \log(f(z)_i), \tag{4}$$

where t_i is the target vector (ground truth labels).

4. Knowledge Distillation-Based Classification Model

In this work, a knowledge distillation approach to classify DR lesions in fundus images is proposed. The main aim of this work is to develop a simple deep learning model, capable of being used in smaller devices that do not have the computational capacity to process all fundus images, but only a reduced set of images that only take into account the most important characteristics of the dataset.

The proposed framework of lesion classification using a knowledge distillation method consists of a two-stage model, as shown in Figure 6. It is composed of two main phases: a teacher and a student model. In the first phase, the teacher model extracts the relevant information related to DR lesions by applying a deep CNN approach. In the second phase, by a transfer learning approach, the extracted relevant DR diagnosis information is used to train the student model.

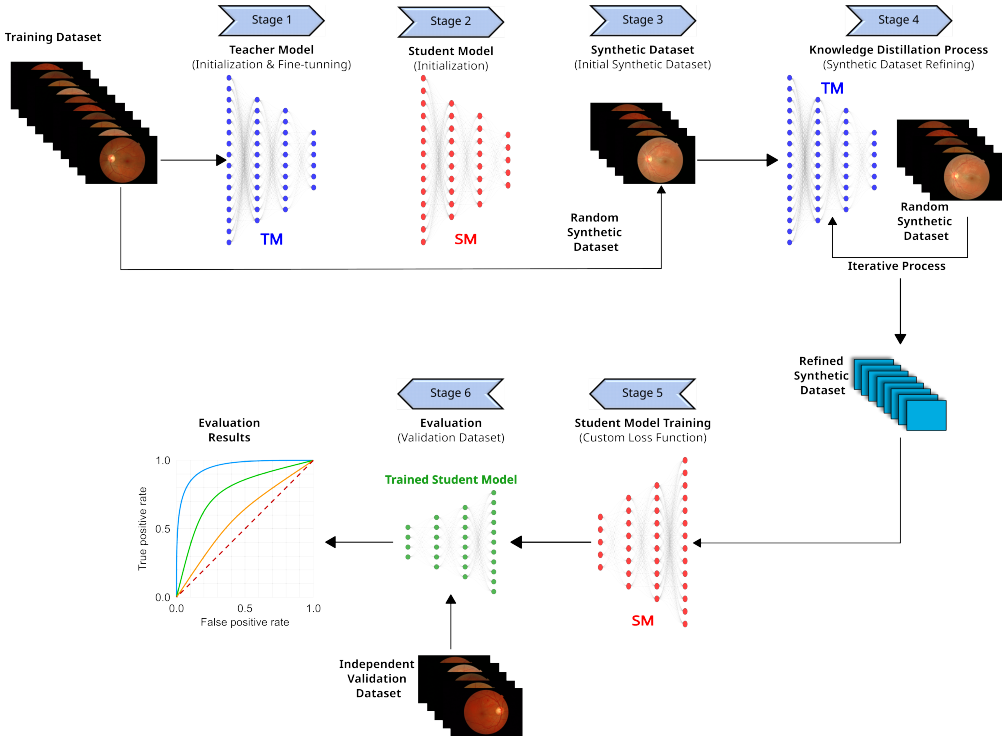


Figure 6. Knowledge distillation-based DR lesion classification framework.

We used Inception-v3 in its Keras implementation as a base and loaded it without its top layer, allowing it to adapt to our data with a different number of target classes. In this case, the model is prepared to work with images of size 224 × 224 and three color channels (RGB). Two tests were conducted to determine whether or not to use pre-trained weights: one with pre-trained weights from ImageNet benchmark and another with no pre-trained weights for training from scratch.

Therefore, it was observed that using the pre-trained weights from ImageNet benchmark, the model's accuracy improved between 1.1 and 2.3%, depending on whether data augmentation was not used or used (respectively) to counteract the unbalanced dataset.

On top of this base Inception-v3 model, custom layers were added to adapt the model to a specific task. First, the output of Inception-v3 was flattened to convert it into a vector. Then, a dense (fully connected) layer with 1024 neurons and ReLU activation was added, followed by a dropout layer with a rate of 0.2 to reduce overfitting. Finally, an output dense layer was added with as many neurons as classes to be predicted (in this case, 39) using a softmax activation function for multiclass classification.

The model was compiled with the RMSprop optimizer, a learning rate of 0.0001, and categorical cross-entropy loss was used, which is suitable for multiclass classification tasks. Tests were also conducted with an ADAM optimizer, and based on the results, it was decided to use ADAM in future iterations of the model, as it provided better results.

The knowledge distillation method employed in this study consists of several key stages (Figure 6), each contributing to the overall goal of creating a compact, efficient representation of the original dataset. These stages encompass the entire process from initial model preparation to final evaluation. We will now describe each of these stages in detail.

Stage 1: Teacher Model Preparation. This stage involves initializing an Inception-v3 model that has been pre-trained on ImageNet. The use of a pre-trained model leverages transfer learning, providing a strong foundation of general visual features. This model is then fine-tuned on the fundus image dataset to be distilled, allowing it to specialize in retinal image analysis while retaining its broader visual understanding. The resulting fine-tuned model becomes the teacher model, embodying the comprehensive knowledge that we aim to distill into a more compact form.

Stage 2: Student Model Initialization. In this stage, another Inception-v3 model architecture is initialized. Like the teacher model, it starts with weights from ImageNet pre-training. However, unlike the teacher model, this student model will not be trained on the full original dataset. Instead, it is prepared to learn from the distilled dataset that will be created in subsequent stages. This approach aims to create a model that can achieve performance comparable to that of the teacher model while using significantly less training data.

Stage 3: Synthetic Dataset Initialization. This stage marks the beginning of the knowledge distillation process. A small subset of samples is randomly selected from the original training dataset. These samples form the initial synthetic dataset, which will serve as the starting point for the distillation process. The goal is to refine this small set of samples to encapsulate the essential information from the entire original dataset.

Stage 4: Knowledge Distillation Process. This is the core of the knowledge distillation process. It involves an iterative procedure to optimize the synthetic dataset. In each iteration, the student model, using the custom loss function, is trained on the current version of the synthetic dataset for one epoch. The student's performance is then evaluated on the original training dataset to assess how well it has captured the knowledge. Gradients of the loss with respect to the synthetic dataset are computed using automatic differentiation. These gradients guide the update of the synthetic dataset through gradient descent, aiming to minimize the loss on the real dataset. This procedure is carried out to a specified number of iterations, progressively refining the synthetic dataset to better represent the knowledge contained in the original dataset.

Stage 5: Student Model Training. Once the knowledge distillation process is complete, the student model is trained on the final distilled synthetic dataset. This training uses a custom loss function that adds Kullback–Leibler divergence and categorical cross-entropy. This combined loss function helps the student model not only learn the correct classifications but also mimic the probability distributions of the teacher model's outputs, potentially capturing some of the nuanced knowledge of the teacher.

Stage 6: Evaluation. In the final stage, the trained student model is evaluated on a separate validation dataset depending on the one it was trained. Its performance is compared to that of the teacher model. This comparison allows us to assess how well the knowledge distillation process has worked. Ideally, the student model should achieve performance comparable to the teacher model despite being trained on a much smaller, synthetic dataset.

The overall method applies knowledge distillation to compress a large training dataset into a smaller synthetic dataset while aiming to maintain model performance. It leverages transfer learning through ImageNet pre-training and uses meta-learning-based performance matching in the distillation process. The goal is to create a compact, efficient training dataset that encapsulates the essential information from the original dataset, allowing for faster and more resource-efficient model training without significant loss in performance.

The teacher model approach in this knowledge distillation process demonstrates several significant strengths across its various stages. In the preparation phase, the model leverages the powerful Inception-v3 architecture, pre-trained on ImageNet, providing a robust foundation of general image features. This transfer learning approach allows the model to start with a rich set of visual representations, which is particularly beneficial for medical imaging tasks where large-scale, domain-specific datasets may be limited. The strategy of freezing pre-trained layers while adding custom top layers strikes a balance between preserving general image understanding and adapting to the specific nuances of retinal image classification.

The student model initialization phase shows strength in its flexibility, setting all layers to be trainable. This approach allows the student model to fully adapt to the task at hand, potentially refining the transferred knowledge from ImageNet to be more specific to retinal imagery. The use of the same Inception-v3 architecture for both teacher and student models

facilitates direct knowledge transfer, ensuring that the student can effectively learn from the teacher's representations.

The knowledge distillation process itself demonstrates a sophisticated approach to learning, employing a custom loss function that combines Kullback–Leibler divergence and categorical cross-entropy. This dual-objective function balances the goals of mimicking the teacher's soft predictions and maintaining high task-specific performance. By using the teacher's predictions as soft targets, the student model can capture nuanced information beyond just the hard class labels, leading to improved generalization and performance.

The method employed in this study leverages a custom loss function that combines Kullback–Leibler divergence and categorical cross-entropy, offering several advantages. This dual-component loss function facilitates effective knowledge distillation by encouraging the student model to mimic the probability distributions of the teacher model's predictions, capturing nuanced information beyond mere class labels. The categorical cross-entropy component ensures the student model maintains high task performance by focusing on correct classifications. By using the teacher's predictions as soft targets through KL divergence, the student model can learn from the teacher's uncertainties, potentially leading to better generalization. This combination also provides a regularizing effect, which is crucial when training on a small synthetic dataset, helping to prevent overfitting. The dual objective allows the student to balance between mimicking the teacher and performing well on the specific classification task.

5. Results and Discussion

5.1. Experimental Setup

The proposed method was implemented on a workstation custom computer running the Ubuntu® 22.04.4 LTS x86_64 (Canonical Ltd., London, UK) operating system. This system was equipped with a 12th Gen Intel® i7-12700 (Intel Corp., Santa Clara, CA, USA), an NVIDIA GeForce RTX® 3090 24 GB graphics card (NVIDIA Corp., Santa Clara, CA, USA), and 64 GB of RAM.

5.2. Evaluation Metrics

Two metrics were used to evaluate the classification performance of the proposed method: the accuracy (*ACC*) and the *F*1-score.

On the one hand, *ACC* measures the relation between correct samples concerning the number total. It is formally defined as follows:

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

where TP are the true-positive samples correctly classified, TN are the true-negative samples correctly classified, FP are the false-positive samples or positive samples incorrectly classified, and FN are the false-negative samples or negative samples incorrectly classified.

On the other hand, in the case of imbalanced datasets, the *F*1-score has been used to provide robust results. This measure considers both the recall and precision ability of the model. The *F*1-score is calculated as follows:

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{6}$$

Both *ACC* and the *F*1-score are measured between 0 and 1, where values close to 1 are better.

5.3. Classification Performance Analysis

In this section, we present an evaluation of the classification performance of the proposed knowledge distillation model. First, we report the results of the baseline models, the Inception-v3 and YOLOv8 models. Next, we present the knowledge distillation results. Finally, we compare the proposed approach to some state-of-the-art works.

5.3.1. Baseline Models

The Inception-v3 baseline model serves as a crucial benchmark for evaluating the efficacy of our proposed knowledge distillation technique in DR lesion classification using fundus images. This baseline, leveraging transfer learning from ImageNet pre-training, provides a robust point of comparison across multiple dimensions. By juxtaposing the performance metrics of this baseline against our distilled model, we can quantitatively assess the impact of the distillation process on classification accuracy, training efficiency, and generalization capability. The baseline model, trained on both datasets, allows us to measure the degree of data compression achieved through distillation while maintaining or potentially improving performance.

In addition to the Inception-v3 baseline, we also implemented a YOLOv8 model as a secondary baseline. This implementation, consisting of a few lines of code, serves as the quickest and most straightforward way to establish a functional model. The YOLOv8 baseline represents a "quick and dirty" approach, with no custom parameters modified, providing a reference point for minimal effort implementation. By including this rapid deployment baseline, we aim to contextualize our results not only against a sophisticated model like Inception-v3 but also against a solution that prioritizes speed and ease of implementation. This dual baseline approach allows us to comprehensively evaluate our distillation technique across a spectrum of model complexities and implementation efforts, offering insights into its versatility and effectiveness in various practical scenarios.

Thus, the Inception-v3 baseline model, trained on both the JSIEC1K and in-house datasets, provides a crucial benchmark for evaluating our proposed knowledge distillation technique. For the in-house dataset, the model achieved a training accuracy of 11.44% and a validation accuracy of 12.50%, with corresponding loss values of 2.0796 and 2.0794, respectively.

This training process took 52 min and 58.40 s, a relatively short time that nonetheless yielded poor performance, indicating significant challenges with this dataset. In contrast, the JSIEC1K dataset showed more promising results, with the model reaching a training accuracy of 66.75% and a validation accuracy of 35.94%. The loss values for this dataset were 1.0304 for training and 4.7077 for validation, suggesting potential overfitting. The training time for JSIEC1K was 34 min and 8.08 s, slightly less than that required for the in-house dataset. These baseline results highlight several key issues: the stark performance difference between datasets, the challenge of generalization as evidenced by the gap between training and validation accuracies, and the non-trivial time investment required even for suboptimal results. This sets a clear context for our knowledge distillation approach, which aims to address these limitations by creating a more efficient and generalizable method for retinal image classification. Figures 7 and 8 show the training and validation curves for accuracy and loss using the Inception-v3 baseline models on the JSIEC1K and in-house datasets, respectively.

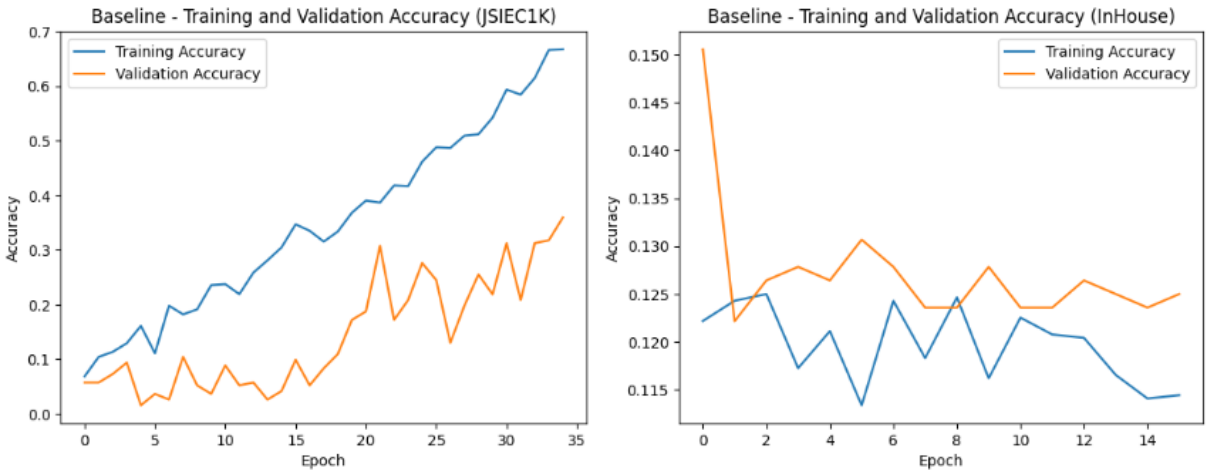


Figure 7. Training and validation accuracy curves for the Inception-v3 baseline models on JSIEC1K (**left**) and in-house (**right**) datasets over training epochs.

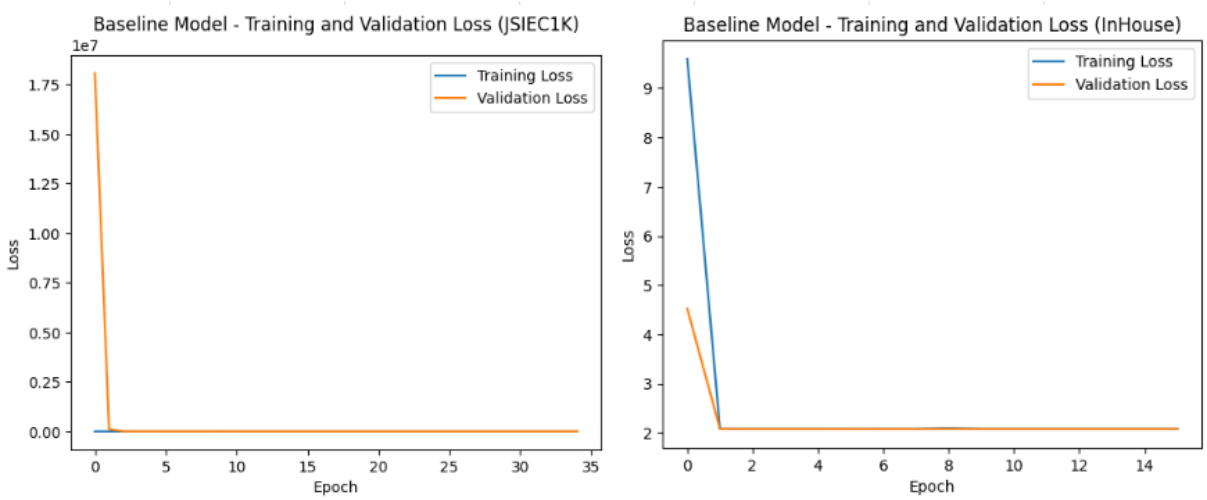


Figure 8. Training and validation loss curves for the Inception-v3 baseline models on JSIEC1K (**left**) and in-house (**right**) datasets over training epochs.

The YOLOv8 approach, implemented as a quick and straightforward baseline, demonstrates intriguing results when applied to both the in-house and JSIEC1K datasets. For the in-house dataset, YOLOv8 achieved a training accuracy of 67.60% with a loss of 0.0648, while the validation loss was 1.5974. This training process took 81 min and 46.16 s, showing a significant improvement in accuracy compared to the Inception-v3 baseline, albeit with a longer training time. The JSIEC1K dataset saw even better results, with YOLOv8 reaching a training accuracy of 90.9% and a loss of 0.05347, while the validation loss was 2.8493. This training was completed in 43 min and 51 s, faster than the in-house training dataset despite the higher accuracy achieved. Figure 9 shows the training and validation loss curves using the YOLOv8 baseline models on the JSIEC1K and in-house datasets.

Table 4 shows a summary of the results using the baseline models Inception-v3 and YOLO-v8 on the JSIEC1K and in-house datasets.

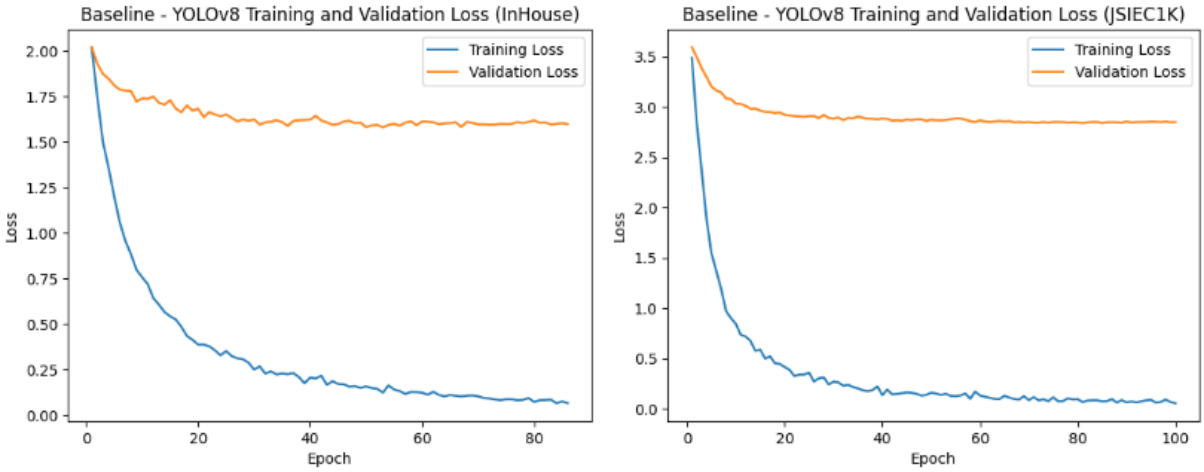


Figure 9. Training and validation loss curves for YOLOv8 baseline models on in-house (**left**) and JSIEC1K (**right**) datasets over training epochs.

Table 4. Results of the baseline models Inception-v3 and YOLO-v8 on the JSIEC1K and in-house datasets.

Model	Dataset	Training Time	Training Acc (%)	Validation Acc (%)	Training Loss	Validation Loss
Inception-v3	In-House	52 m 58.40 s	11.44	12.50	2.0796	2.0794
	JSIEC1K	34 m 8.08 s	66.75	35.94	1.0304	4.7077
YOLOv8	In-House	81 m 46.16 s	67.60	58.47	0.0648	1.5874
	JSIEC1K	43 m 51.00 s	90.90	81.52	0.5347	2.8493

From Table 4, the results reported highlight some key points, which are listed below:

- The YOLOv8 model demonstrates superior training accuracy on both datasets compared to the Inception-v3 baseline, particularly on the JSIEC1K dataset where it achieved outstanding 90.90% accuracy.
- The training times, while longer than the Inception-v3 baseline, are still relatively short considering the accuracy gains. However, high validation losses, especially for the JSIEC1K dataset, suggest potential overfitting issues.
- The absence of validation accuracy figures limits our ability to fully assess the model's generalization capabilities. Nevertheless, these results underscore the potential of the YOLOv8 architecture for rapid prototyping and baseline establishment in retinal image classification tasks, providing a strong point of comparison for our knowledge distillation technique.

5.3.2. Knowledge Distillation Results

The application of knowledge distillation to the Inception-v3 model trained on the in-house dataset has yielded exceptional results, marking a significant advancement in retinal image classification performance. Figure 10 shows some classification results using the knowledge distillation proposal trained on the in-house dataset. The estimate class is shown at the top left of each image. The distilled model achieved a remarkable training accuracy of 99.01% and a validation accuracy of 97.30%, with corresponding loss values of 0.0912 for training and 0.3284 for validation. These impressive outcomes were obtained after a training duration of 45 min qne 22.4 s, demonstrating the effectiveness of the distillation process in capturing and condensing essential information from the original dataset. The distilled model's performance represents a substantial improvement over both the Inception-v3 and YOLOv8 baselines previously tested on the in-house dataset. The high validation accuracy of 97.30% is particularly noteworthy, as it indicates excellent generalization capability, a crucial factor for the practical application of machine learning models in medical imaging. This strong generalization performance suggests that the

distilled model has successfully captured the key features and patterns within the retinal images, allowing it to make accurate predictions on unseen data. Table 5 compares the knowledge distillation approach with the Inception-v3 model against the baseline model YOLOv8, both trained on the in-house dataset.

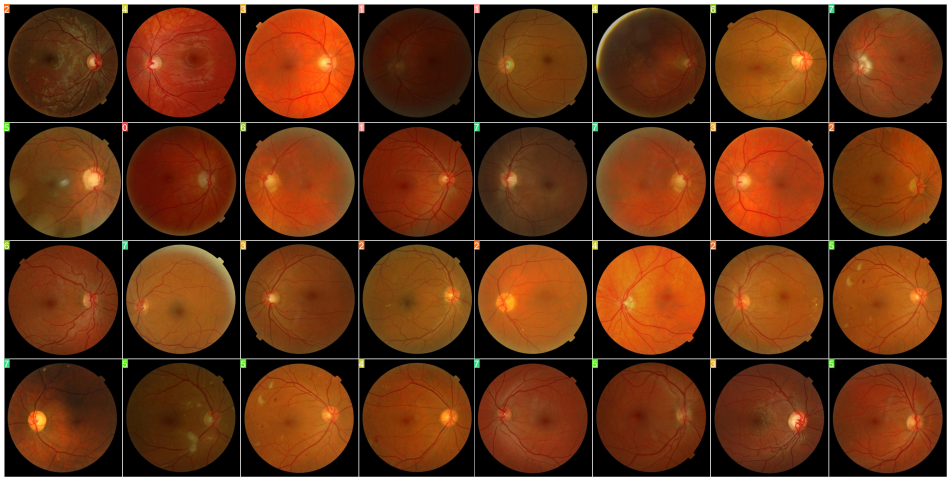


Figure 10. Fundus images examples and the classes assigned by the knowledge distillation proposal trained on the in-house dataset. The estimate class is shown at the top left of each image using the numbers 0: 'Beading', 1: 'Macular Degeneration', 2: 'Exudade', 3: 'Glaucoma ', 4: 'Hemorrhages', 5: 'Cotton-wool spots', 6: 'Microaneurysms', and 7: 'Healthy'.

Table 5. Comparison of the knowledge distillation approach against the baseline model YOLOv8 on the in-house dataset.

Model	Training Time	Training Acc (%)	Validation Acc (%)	Training Loss	Validation Loss
YOLOv8 (baseline)	81 m 46.16 s	67.60	58.47	0.0648	1.5874
KD	45 m 22.40 s	99.01	97.30	0.0912	0.3284

One of the most significant advantages of the distillation approach is the reduction in overfitting, a common challenge in deep learning models. The narrow gap between the training accuracy (99.01%) and validation accuracy (97.30%) indicates that the distillation process has effectively mitigated the overfitting issues observed in baseline models. This improved balance between fitting the training data and generalizing to new data is critical for creating reliable and robust models for medical applications. The loss values achieved by the distilled model are also considerably lower than those of the baseline models, with a training loss of 0.0912 and a validation loss of 0.3284. These reduced loss values signify a more stable and well-fitted model, suggesting that the distillation process has enabled the model to learn a more accurate representation of the underlying data distribution. The low loss values, combined with high accuracies, indicate that the model is both confident and correct in its predictions, a desirable characteristic for any classification system, but particularly crucial in medical diagnostics.

While the training time for the distilled model (approximately 45 min) is shorter than that of the baseline Inception-v3 model, the substantial performance gains justify this additional computational pre-investment. In the context of developing robust medical imaging tools, the improved accuracy and generalization capabilities outweigh the increased training time, especially considering that training is a one-time process and the resulting model can be deployed efficiently.

On the other hand, the training and validation loss curves, shown in Figure 11, provide further insight into the learning process of the knowledge-distilled model. Both curves show rapid convergence, with sharp decreases in loss during the early epochs, followed by stabilization. This pattern is indicative of efficient learning and good model stability.

The quick convergence suggests that the distillation process has effectively captured the most relevant features of the dataset, allowing the model to learn rapidly and reach high performance levels early in the training process.

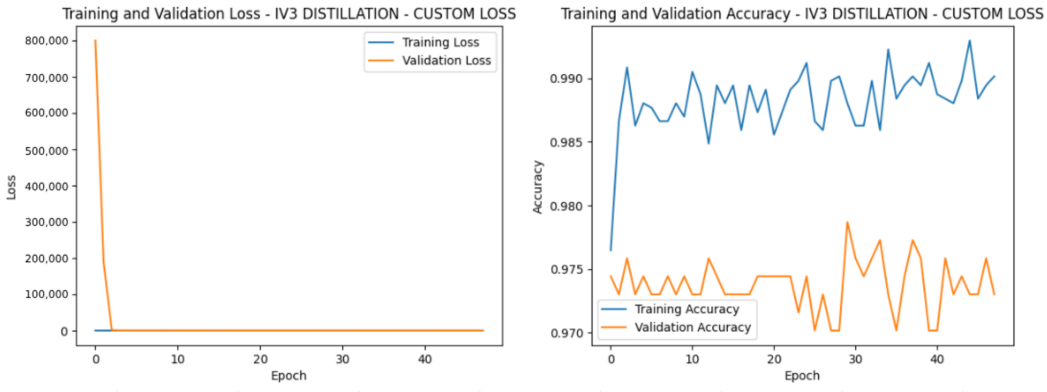


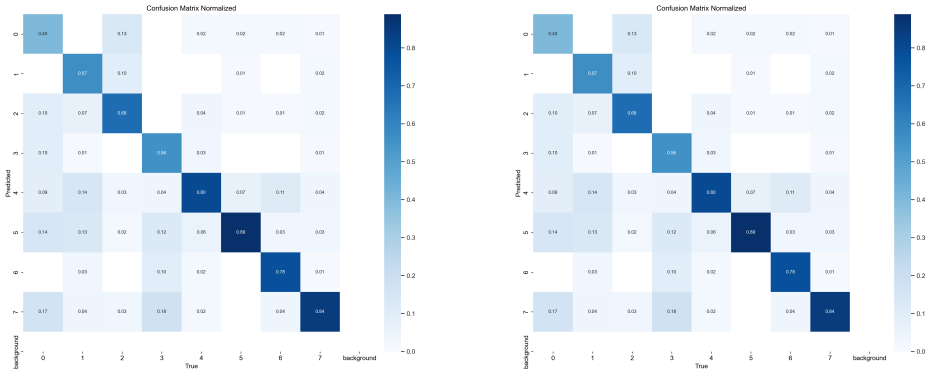
Figure 11. Training and validation loss (**left**) and accuracy (**right**) curves for the knowledge-distilled model with a custom loss function over training epochs.

Similarly, the accuracy curves reveal consistently high performance throughout the training phase. Both training and validation accuracies maintain levels above 97.0%, demonstrating the model's ability to learn effectively from the distilled dataset and generalize well to unseen data. This consistency in high accuracy across training and validation sets additionally confirms the strength of the distilled model. These findings strongly support the knowledge distillation model's effectiveness in DR lesion classification using fundus images. The distilled model not only outperforms the baselines in terms of raw accuracy but also demonstrates superior generalization capabilities. This suggests that the distillation process has successfully captured and condensed the essential information from the original dataset, enabling the model to learn more effectively and efficiently.

In addition, Figure 11 illustrates the training and validation performance of the knowledge-distilled model with a custom loss function on retinal image classification. The loss curves (left) demonstrate rapid convergence, with a sharp initial decrease in validation loss from approximately 800,000 to near 0 within the first few epochs, followed by consistent low loss values for both training and validation sets. The accuracy curves (right) show high performance throughout training, with training accuracy consistently above 98.50% and often exceeding 99.0%, while validation accuracy fluctuates between 97.0% and 98.0%. This aligns with the reported training accuracy of 99.01% and validation accuracy of 97.30%. The narrow gap between training and validation accuracies (less than 2%) indicates excellent generalization and reduced overfitting compared to baseline models. The stability of both loss and accuracy curves after initial rapid improvement suggests a well-fitted model with efficient learning characteristics. These results corroborate the exceptional performance of the distilled model, demonstrating significant improvements over Inception-v3 and YOLOv8 baselines on the in-house dataset and highlighting the effectiveness of knowledge distillation in capturing essential features for robust retinal image classification.

Finally, Figure 12 shows the confusion matrix and its normalized version of the knowledge distillation proposal trained on the in-house dataset.

The proposed knowledge distillation approach for DR lesion classification represents a significant advancement in the field of automated medical image analysis. By achieving high accuracy, excellent generalization, and efficient learning, this method shows promise for improving early diagnosis of diabetic retinopathy, potentially leading to better patient outcomes and more efficient healthcare delivery.



(**a**) (**b**)

Figure 12. Confusion matrix without (**a**) and with (**b**) normalization of the knowledge distillation proposal trained on the in-house dataset.

Despite its strong performance, the model's reliance on a fixed architecture (Inception-v3) may limit its adaptability to varying retinal image characteristics or different medical imaging tasks. The knowledge distillation process, although powerful, is constrained by the quality and biases of the teacher model, potentially propagating any inherent errors or limitations. The fixed hyperparameters, such as learning rate and loss function weighting, may not be optimal for all datasets or throughout the entire training process for different medical data sources. Additionally, the model's performance on extremely rare conditions or edge cases in retinal imagery might be limited due to the potential scarcity of such examples in the training data. Lastly, while the model shows high accuracy, its interpretability remains a challenge, which is crucial in medical applications where understanding the reasoning behind predictions is often as important as the predictions themselves.

In applying the pre-trained Inception-v3 architecture as a teacher model for dataset distillation, several challenges were encountered. The primary challenge involved adapting the model, originally trained on general image data, to the specific nuances of retinal imagery. This required careful fine-tuning to ensure effective capture of unique features in fundus images without overfitting to the training data. Another significant challenge was balancing the trade-off between model complexity and computational efficiency, particularly given the high-resolution nature of medical images.

These challenges were addressed through a systematic approach of experimentation and optimization. The model's hyperparameters were carefully adjusted: the top layers were fine-tuned while keeping pre-trained layers frozen, and custom data augmentation techniques specific to retinal images were implemented. To enhance the distillation process, a custom loss function combining Kullback–Leibler divergence with categorical cross-entropy was developed, allowing the student model to learn both from the teacher's soft predictions and the true labels.

It is noteworthy that the selection of Inception-v3 was not arbitrary. Multiple architectures were explored to determine the one that yielded the best results, including VGG16, ResNet, RetinaNet, and MobileNet. After extensive experimentation and comparison, the Inception-v3 architecture was selected due to its superior reliability and better generalization capabilities for the specific task of retinal image classification. This comprehensive approach to architecture selection ensured that the chosen model was well-suited to the unique challenges of the dataset and the requirements of the distillation process.

5.4. Comparison Study

In this section, we compare the proposed knowledge distillation classification of DR lesions against some state-of-the-art works. First, we compare our knowledge distillation model with a similar work that uses different distilled dataset strategies applied to ovarian cancer diagnoses. Second, we compare our method with five works on DR classification in fundus images using KD.

In the research by Salguero et al. [36] they had a problem with training being too expensive, so they opted to reduce the size of the training dataset by selecting images that best expressed the features. Their dataset consisted of 103 ovarian cancer diagnoses, of which 23 were detected at stage 3 or 4. The latter were chosen given the large difference shown in a tissue at that stage and a healthy one. They randomly took 13 cases with cancer and 13 with normal tissue. They describe that the distillation process selects a small patch showing greater variance in the matrix of eigenvalues S obtained from singular value decomposition (SVD). For this, they used different classifiers such as an SVM or multilayer perceptron (MLP). One of the several experiments they carried out was the comparison of performances of several sets in a MLP neural network. They used five different sets: the distilled dataset, the entire dataset, two randomly selected datasets, and a set obtained using principal component analysis (PCA).

Table 6 compares the $F1$-score of the proposed KD classification DR lesion approach and the results of ovarian cancer detection reported in [36] using different data subsets.

Table 6. $F1$-scores comparison of the proposed KD method against the approach of [36] using different data subsets.

Set Name	Their F-Score	Our F-Score
Distilled set (44%)	0.8784	0.9091
Complete set (100%)	0.7890	0.8875
Random set (44%)	0.7678	0.9076
Random set (5%)	0.7404	0.8932
PCA (σ = 95%)	0.6920	—
PCA (σ = 99%)	0.8146	—

In contrast to our results, they obtained performance with training models both with 40% of the total sample size and with the entire image set as about 0.87 as $F1$-score. On the other hand, using the data subset obtained by data distillation, the $F1$-score was superior. They concluded that using this methodology, the training time was reduced by 50%. Furthermore, the results obtained with the proposed KD method overcame the results of [36] using the distilled, complete, and random sets. The evaluation with the PCA set was not carried out in our proposal.

On the other hand, the proposed DR lesions classification approach was compared with five state-of-the-art works on DR classification in fundus images using a knowledge distillation strategy. The compared works are described as follows.

Luo et al. [30] proposed a self-knowledge distillation approach (SKD) for DR grading. In addition, a CAM-attention module (CAM-AM), focusing on regions with DR lesions, and a mimicking module (MM), which holds its original hierarchy in the distillation process, was proposed. Regarding the loss function, L_2 and KL divergence were used for the teacher and the student model, respectively. Messidor (1200 images) and IDRID (516 images) fundus datasets were used to test their proposal. In this work, no pre-processing stage was applied.

In [31], Abbasi et al. used knowledge distillation as the transfer knowledge strategy. First, in the pre-processing stage, the low-contrast images were eliminated, and then a histogram equalization was applied. Second, unlabeled data were used to transfer DR knowledge to a simple model capable of being implemented into a low-resource embedded system. Both the teacher and student model used a VGG network with KL divergence as a loss function. The proposal was evaluated in the Messidor fundus dataset for binary DR classification.

Gao et al. [32] presented a collaborative learning-based knowledge distillation strategy for DR detection. It included several student models of different scales and architectures to extract relevant diagnosis information. This allowed for enhancing fundus images while increasing the DR detection accuracy and reducing the computation time. On the one hand, ResNet-18 architecture was used as a teacher model, whereas the BEit transformer and ConvNeXt-CNN architectures were used as student models. On both architectures, KL divergence as a loss function was used. On the other hand, the authors applied image

cropping, histogram equalization, and contrast stretching in an in-house (1521 images) fundus dataset.

In [34], Ju et al. used a long-tailed fundus dataset, i.e., where a few classes represent the majority of the images. The authors proposed DR lesion classification using a knowledge distillation approach. First, the labeled dataset was used for hierarchy-aware pre-training and multi-label marginalization classification (MLMC). Second, an instance-wise class-balanced sampling (ICS) approach was applied. Third, a hybrid knowledge distillation (HKD) method was used to train a less biased representation. Finally, two in-house datasets of one million images with 50 classes were used to conduct the experiments. A ResNet-50 architecture was used for both the teacher and student model, and KL divergence as a loss function. In this proposal, no pre-processing stage was applied.

Wang et al. [35] presented a lesion-aware knowledge distillation (LKD) approach to DR lesion segmentation in three classes. The proposed LKD strategy uses a lesion embedding queue in the global training, and it was tested on the IDRiD (81 images) and DDR (757 images) fundus datasets. The MCA-UNet and UNet networks were used as teacher and student models in the KD method, respectively. The CLAHE algorithm was applied to the entire dataset as a pre-processing stage.

In Table 7, we present a comparison of the proposed KL-based DR lesion classification method against five state-of-the-art methods [30–32,34,35]. Thus, we can see that the proposed knowledge distillation model overcomes the proposal of [31] regarding the training time and the methods of [32,35] in the number of parameters of the model. Furthermore, our proposal overcomes all methods in accuracy. In addition, our proposal does not require pre-processing of the datasets, allowing it to be applied directly to the fundus image datasets.

Table 7. Comparison of the proposed method against five state-of-the-art methods.

Work/ Year	Method	Classification	Dataset/ Images-Classes	Pre-Processing	Teacher Model	Student Model	Loss Function	Time	Parameters (M)	Acc (%)
[30]/ 2020	SKD + CAM-AM + MM	DR grading	Messidor/1.2K, IDRiD/516	None	—	—	L_2 (Teacher) KL (Student)	—	—	92.9 (Messidor) 67.96 (IDRiD)
[31]/ 2021	KD	DR detection	Messidor/1.2K-2C	Images removal and histogram equalization	VGG	VGG	KL	5.12 h	—	82.32
[32]/ 2023	KD	DR detection	In-house/1.521K-2C	Image cropping, histogram equalization, and contrast stretching	ResNet-18	BEiT	KL	—	81.18	98.77
[34]/ 2024	MLMC + ICS + HKD	DR lesion classification	In-house/1M-50C	None	ResNet-50	ResNet-50	KL	—	—	85.79
[35]/ 2024	LKD	DR lesion segmentation	IDRiD/81-3C, DDR/757-3C	CLAHE	MCA-UNet	UNet	KL	—	7845	—
Ours/ 2024	KD	DR lesion classification	JSIEC1K/1K-39C, In-house/1.705K-8C	None	Inception-v3	Inception-v3	KL + CCE	45 m 22.40 s	71.38	99.01 (Messidor/In-House)

6. Conclusions

The knowledge distillation approach employed in this study has demonstrated remarkable effectiveness in enhancing the performance of retinal image classification for diabetic retinopathy detection. One of the most significant achievements of the distilled model is its superior generalization capability. The narrow gap between training (99.01%) and validation (97.30%) accuracies suggests that the model has effectively mitigated overfitting issues observed in baseline models.

This substantial improvement enables more effective and efficient learning, which is crucial in the context of medical image analysis. The high accuracy and robust generalization of the distilled model suggest strong potential for clinical application in DR screening.

The distilled model shows a significant reduction in overfitting compared to the baseline models, as evidenced by the close alignment of training and validation performance metrics. Such a characteristic is particularly important in medical applications where generalizability is crucial for accurate diagnosis across diverse patient populations. Finally,

the proposed method was compared against some state-of-the-art works, overcoming these and maintaining a compact and efficient final model.

Future research directions could explore the application of this knowledge distillation technique to other medical imaging tasks or larger, more diverse datasets. This could include adapting the method for the detection and classification of other retinal diseases, such as age-related macular degeneration or glaucoma. In addition, extending its application to different medical imaging modalities, e.g., fluorescein angiography or optical coherence tomography, could be explored. Additionally, investigating the model's interpretability and its performance in real-world clinical settings would be valuable next steps. This could involve collaborating with ophthalmologists to assess the model's performance in comparison to human experts and to understand how the model's predictions align with clinical decision-making processes. Finally, an integration of this model into existing clinical workflows could be analyzed. This could be reached by involving several steps: (1) developing a user-friendly interface for clinicians, (2) ensuring seamless connectivity with current imaging systems, and (3) implementing robust data handling and privacy measures.

Author Contributions: Conceptualization, E.M.-A., A.L.-F., S.J.-H. and D.R.; Data curation, E.M.-A., A.L.-F., S.J.-H. and J.B.; Formal analysis, E.M.-A., A.L.-F., S.J.-H. and D.R.; Funding acquisition, E.M.-A.; Investigation, E.M.-A., A.L.-F., S.J.-H. and D.R.; Methodology, E.M.-A., A.L.-F., S.J.-H. and D.R.; Project administration, E.M.-A.; Resources, E.M.-A., A.L.-F., S.J.-H. and J.B.; Software, A.L.-F. and S.J.-H.; Supervision, E.M.-A. and D.R.; Validation, A.L.-F., S.J.-H. and J.B.; Visualization, A.L.-F., S.J.-H. and J.B.; Writing—original draft, E.M.-A., A.L.-F. and S.J.-H.; Writing—review and editing, E.M.-A., A.L.-F., S.J.-H., D.R. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Universidad Panamericana under the Program "Fomento a la Investigación UP 2023" grant UP-CI-2023-MX-10-ING.

Data Availability Statement: The image datasets used in this work are publicly available from Kaggle: JSIEC1K [38]: https://www.kaggle.com/datasets/linchundan/fundusimage1000/ (accessed on 13 February 2024), MESSIDOR-2 [39,40]; https://www.kaggle.com/datasets/geracollante/messidor2/ (accessed on 13 February 2024).

Acknowledgments: Ernesto Moya-Albor, Alberto Lopez-Figueroa, Sebastian Jacome-Herrera, and Jorge Brieva thank the Universidad Panamericana for all their support in this work. Diego Renza thanks the Universidad Militar Nueva Granada for the support in this work.

Conflicts of Interest: The authors declare there are no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
AMD	Age-related Macular Degeneration
CAM	Class Activation Mapping
CAM-AM	CAM-Attention Module
CCE	Categorical Cross-Entropy
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
DL	Deep Learning
DR	Diabetic Retinopathy
GA	Geographic Atrophy
GLA	Glaucoma
HKD	Hybrid Knowledge Distillation
ICA	Independent Component Analysis
IDF	International Diabetes Federation
ISC	Istance-wise Class-balanced Sampling
JSIEC	Joint Shantou International Eye Center
KD	Knowledge Distillation
KL	Kullback–Leibler
LKD	Lesion-aware Knowledge Distillation
MLMC	Multi-label Marginalization Classifier

MLP	Multilayer Perceptron
MM	Mimicking Module
OCT	Optical Coherence Tomography
PCA	Principal Component Analysis
SKD	Self-knowledge Distillation
SVD	Singular Value Decomposition
SVM	Support Vector Machine
YOLO	You Only Look Once

References

1. International Diabetes Federation. *IDF Diabetes Atlas*, 10th ed.; Technical Report; International Diabetes Federation: Brussels, Belgium, 2021. Available online: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed on 27 February 2024).
2. Wong, T.Y.; Tan, T.E. The Diabetic Retinopathy "Pandemic" and Evolving Global Strategies: The 2023 Friedenwald Lecture. *Investig. Ophthalmol. Vis. Sci.* **2023**, *64*, 47. [CrossRef]
3. Prado-Serrano, A.; Guido-Jiménez, M.; Camas-Benítez, J. Prevalence of diabetic retinopathy in Mexican population [Prevalencia de retinopatía diabética en población mexicana]. *Rev. Mex. Oftalmol.* **2009**, *83*, 261–266.
4. Porta, M.; Klein, R.; Klein, B.; Kohner, E. Saving Sight: A History of Diabetic Eye Disease. *Front. Diabetes* **2020**, *29*, 221–241. [CrossRef]
5. Ometto, G.; Erlandsen, M.; Hunter, A.; Bek, T. The role of retinopathy distribution and other lesion types for the definition of examination intervals during screening for diabetic retinopathy. *Acta Ophthalmol.* **2017**, *95*, 400–404. [CrossRef]
6. Tomić, M.; Vrabec, R.; Ljubić, S.; Prkačin, I.; Bulum, T. Patients with Type 2 Diabetes, Higher Blood Pressure, and Infrequent Fundus Examinations Have a Higher Risk of Sight-Threatening Retinopathy. *J. Clin. Med.* **2024**, *13*, 2496. [CrossRef] [PubMed]
7. Rousso, L.; Sowka, J. Recognizing abnormal vasculature: A guide to following and educating patients who face this class of sight-threatening diagnoses. *Rev. Optom.* **2017**, *154*, 82–87. Available online: https://www.reviewofoptometry.com/article/recognizing-abnormal-vasculature (accessed on 27 February 2024).
8. Peralta-Ildefonso, M.J.; Moya-Albor, E.; Brieva, J.; Lira-Romero, E.; Perez-Ortiz, A.C.; Coral-Vazquez, R.; Estrada-Mena, F.J. Nuclear density analysis in microscopic images for the characterization of retinal geographic atrophy. In Proceedings of the 15th International Symposium on Medical Information Processing and Analysis, Bellingham, WA, USA, 6–8 November 2020; Volume 11330. [CrossRef]
9. Sundaram, R.; KS, R.; Jayaraman, P.; B, V. Extraction of Blood Vessels in Fundus Images of Retina through Hybrid Segmentation Approach. *Mathematics* **2019**, *7*, 169. [CrossRef]
10. Colomer, A.; Naranjo, V.; Engan, K.; Skretting, K. Assessment of sparse-based inpainting for retinal vessel removal. *Signal Process. Image Commun.* **2017**, *59*, 73–82. [CrossRef]
11. Sharif, M.; Shah, J. Automatic screening of retinal lesions for grading diabetic retinopathy. *Int. Arab. J. Inf. Technol.* **2019**, *16*, 766–774.
12. Kaur, J.; Mittal, D. Estimation of severity level of non-proliferative diabetic retinopathy for clinical aid. *Biocybern. Biomed. Eng.* **2018**, *38*, 708–732. [CrossRef]
13. Estudillo-Ayala, M.d.J.; Aguirre-Ramos, H.; Avina-Cervantes, J.G.; Cruz-Duarte, J.M.; Cruz-Aceves, I.; Ruiz-Pinales, J. Algorithmic Analysis of Vesselness and Blobness for Detecting Retinopathies Based on Fractional Gaussian Filters. *Mathematics* **2020**, *8*, 744. [CrossRef]
14. Wang, R.; Chen, B.; Meng, D.; Wang, L. Weakly Supervised Lesion Detection From Fundus Images. *IEEE Trans. Med. Imaging* **2019**, *38*, 1501–1512. [CrossRef] [PubMed]
15. Biswas, S.; Upadhya, R.; Das, N.; Das, D.; Chakraborty, M.; Purkayastha, B. An Intelligent System for Diagnosis of Diabetic Retinopathy. *Adv. Intell. Syst. Comput.* **2020**, *1139*, 97–110. [CrossRef]
16. Elbalaoui, A.; Fakir, M. Exudates detection in fundus images using mean-shift segmentation and adaptive thresholding. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2019**, *7*, 145–153. [CrossRef]
17. Afrin, R.; Shill, P. *Automatic Lesions Detection and Classification of Diabetic Retinopathy Using Fuzzy Logic*; Institute of Electrical and Electronics Engineers Inc.: New York City, NY, USA, 2019; pp. 527–532. [CrossRef]
18. Adal, K.; Van Etten, P.; Martinez, J.; Rouwen, K.; Vermeer, K.; Van Vliet, L. An automated system for the detection and classification of retinal changes due to red lesions in longitudinal fundus images. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 1382–1390. [CrossRef] [PubMed]
19. Sidibé, D.; Sadek, I.; Mériaudeau, F. Discrimination of retinal images containing bright lesions using sparse coded features and SVM. *Comput. Biol. Med.* **2015**, *62*, 175–184. [CrossRef]
20. Ghasemi Falavarjani, K.; Tsui, I.; Sadda, S. Ultra-wide-field imaging in diabetic retinopathy. *Vis. Res.* **2017**, *139*, 187–190. [CrossRef]
21. Wu, H.; Zhang, X.; Geng, X.; Dong, J.; Zhou, G. Computer aided quantification for retinal lesions in patients with moderate and severe non-proliferative diabetic retinopathy: A retrospective cohort study. *BMC Ophthalmol.* **2014**, *14*, 126. [CrossRef]
22. Usman Akram, M.; Khalid, S.; Tariq, A.; Khan, S.; Azam, F. Detection and classification of retinal lesions for grading of diabetic retinopathy. *Comput. Biol. Med.* **2014**, *45*, 161–171. [CrossRef]

23. Ashraf, M.N.; Hussain, M.; Habib, Z. Deep Red Lesion Classification for Early Screening of Diabetic Retinopathy. *Mathematics* **2022**, *10*, 686. [CrossRef]
24. Alsubai, S.; Alqahtani, A.; Binbusayyis, A.; Sha, M.; Gumaei, A.; Wang, S. Quantum Computing Meets Deep Learning: A Promising Approach for Diabetic Retinopathy Classification. *Mathematics* **2023**, *11*, 2008. [CrossRef]
25. Priya, H.; Anitha, J.; Popescu, D.; Asokan, A.; Jude Hemanth, D.; Son, L. Detection and grading of diabetic retinopathy in retinal images using deep intelligent systems: A comprehensive review. *Comput. Mater. Contin.* **2021**, *66*, 2771–2786. [CrossRef]
26. Bilal, A.; Sun, G.; Mazhar, S. Survey on recent developments in automatic detection of diabetic retinopathy [Enquête sur les récents développements en matière de détection automatique de la rétinopathie diabétique]. *J. Fr. D'Ophtalmol.* **2021**, *44*, 420–440. [CrossRef]
27. Abdelmaksoud, E.; El-Sappagh, S.; Barakat, S.; Abuhmed, T.; Elmogy, M. Automatic Diabetic Retinopathy Grading System Based on Detecting Multiple Retinal Lesions. *IEEE Access* **2021**, *9*, 15939–15960. [CrossRef]
28. Hassan, T.; Akram, M.; Werghi, N. *Exploiting the Transferability of Deep Learning Systems across Multi-Modal Retinal Scans for Extracting Retinopathy Lesions*; Institute of Electrical and Electronics Engineers Inc.: New York City, NY, USA, 2020; pp. 577–581. [CrossRef]
29. Playout, C.; Duval, R.; Cheriet, F. A Novel Weakly Supervised Multitask Architecture for Retinal Lesions Segmentation on Fundus Images. *IEEE Trans. Med. Imaging* **2019**, *38*, 2434–2444. [CrossRef]
30. Luo, L.; Xue, D.; Feng, X. Automatic diabetic retinopathy grading via self-knowledge distillation. *Electronics* **2020**, *9*, 1337. [CrossRef]
31. Abbasi, S.; Hajabdollahi, M.; Khadivi, P.; Karimi, N.; Roshandel, R.; Shirani, S.; Samavi, S. Classification of diabetic retinopathy using unlabeled data and knowledge distillation. *Artif. Intell. Med.* **2021**, *121*, 102176. [CrossRef]
32. Gao, Y.; Ma, C.; Guo, L.; Zhang, X.; Ji, X. CLRD: Collaborative Learning for Retinopathy Detection Using Fundus Images. *Bioengineering* **2023**, *10*, 978. [CrossRef] [PubMed]
33. Islam, N.; Jony, M.M.H.; Hasan, E.; Sutradhar, S.; Rahman, A.; Islam, M.M. Toward Lightweight Diabetic Retinopathy Classification: A Knowledge Distillation Approach for Resource-Constrained Settings. *Appl. Sci.* **2023**, *13*, 12397. [CrossRef]
34. Ju, L.; Yu, Z.; Wang, L.; Zhao, X.; Wang, X.; Bonnington, P.; Ge, Z. Hierarchical Knowledge Guided Learning for Real-World Retinal Disease Recognition. *IEEE Trans. Med. Imaging* **2024**, *43*, 335–350. [CrossRef]
35. Wang, Y.; Hou, Q.; Cao, P.; Yang, J.; Zaiane, O.R. Lesion-aware knowledge distillation for diabetic retinopathy lesion segmentation. *Appl. Intell.* **2024**, *54*, 1937–1956. [CrossRef]
36. Salguero, J.; Prasanna, P.; Corredor, G.; Cruz-Roa, A.; Becerra, D.; Romero, E. Data distillation in computational pathology by choosing few representants of the original variance: A use case in ovarian cancer. *Expert Syst. Appl.* **2024**, *245*, 123028. [CrossRef]
37. Khan, S.M.; Liu, X.; Nath, S.; Korot, E.; Faes, L.; Wagner, S.K.; Keane, P.A.; Sebire, N.J.; Burton, M.J.; Denniston, A.K. A global review of publicly available datasets for ophthalmological imaging: Barriers to access, usability, and generalisability. *Lancet Digit. Health* **2021**, *3*, e51–e66. [CrossRef]
38. Cen, L.P.; Ji, J.; Lin, J.W.; Ju, S.T.; Lin, H.J.; Li, T.P.; Wang, Y.; Yang, J.F.; Liu, Y.F.; Tan, S.; et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **2021**, *12*, 4828. [CrossRef] [PubMed]
39. Decencière, E.; Zhang, X.; Cazuguel, G.; Laÿ, B.; Cochener, B.; Trone, C.; Gain, P.; Ordóñez Varela, J.R.; Massin, P.; Erginay, A.; et al. Feedback on a publicly distributed image database: The Messidor database. *Image Anal. Stereol.* **2014**, *33*, 231–234. [CrossRef]
40. Abràmoff, M.D.; Folk, J.C.; Han, D.P.; Walker, J.D.; Williams, D.F.; Russell, S.R.; Massin, P.; Cochener, B.; Gain, P.; Tang, L.; et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol.* **2013**, *131*, 351–357. [CrossRef] [PubMed]
41. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27 (NIPS'14)*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 3320–3328.
42. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [CrossRef]
43. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
44. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242. Available online: http://arxiv.org/abs/1612.08242 (accessed on 4 March 2024).
45. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 27 February 2024).
46. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
47. Zhang, Z.; Sabuncu, M.R. Generalized cross entropy loss for training deep neural networks with noisy labels. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 8778–8788.