applied sciences

MDPI

Article

Transfer Learning for Diabetic Retinopathy Detection: A Study of Dataset Combination and Model Performance

A. M. Mutawa *, Shahad Alnajdi and Sai Sruthi

Department of Computer Engineering, College of Engineering and Petroleum, Kuwait University,
P.O. Box 5969, Safat 13060, Kuwait

* Correspondence: dr.mutawa@ku.edu.kw

Abstract: Diabetes' serious complication, diabetic retinopathy (DR), which can potentially be life-threatening, might result in vision loss in certain situations. Although it has no symptoms in the early stages, this illness is regarded as one of the "silent diseases" that go unnoticed. The fact that various datasets have varied retinal features is one of the significant difficulties in this field of study. This information impacts the models created for this purpose. This study's method can efficiently learn and classify DR from three diverse datasets. Four models based on transfer learning Convolution Neural Network (CNN)—Visual Geometry Group (VGG) 16, Inception version 3 (InceptionV3), Dense Network (DenseNet) 121, and Mobile Network version 2 (MobileNetV2)—are employed in this work, with evaluation parameters, including loss, accuracy, recall, precision, and specificity. The models are also tested by combining the images from the three datasets. The DenseNet121 model performs better with 98.97% accuracy on the combined image set. The study concludes that combining multiple datasets improves performance compared to individual datasets. The obtained model can be utilized globally to accommodate more tests that clinics perform for diabetic patients to prevent DR. It helps health workers refer patients to ophthalmologists before DR becomes serious.

Keywords: convolutional neural network; deep learning; diabetic retinopathy; image classification; medical imaging; transfer learning

check for updates

Citation: Mutawa, A.M.; Alnajdi, S.; Sruthi, S. Transfer Learning for Diabetic Retinopathy Detection: A Study of Dataset Combination and Model Performance. *Appl. Sci.* **2023**, *13*, 5685. https://doi.org/10.3390/app13095685

Academic Editors: Sami Bourouis, Hammam Alshazly and Ali Javed

Received: 13 April 2023
Revised: 1 May 2023
Accepted: 3 May 2023
Published: 5 May 2023

1. Introduction

Diabetes is a severe (can be long-lasting) condition where the body cannot generate enough insulin or utilize the insulin it produces. The longer a patient has diabetes and has poor blood sugar management, the greater the chance of complications, some of which could be fatal. Numerous illnesses, including kidney damage, nerve damage, eye damage, cardiovascular disease, hearing loss, Alzheimer's, skin issues, infections, foot damage, and numerous other complications, can be brought on by diabetes [1].

Globally, 1.5 million deaths were recorded from diabetes in 2019, according to the World Health Organization (WHO). The organization approved five worldwide coverage and treatment goals for diabetes in May 2022, which are to be met by 2030 [2]. Diabetes affected 536.6 million people aged 20 to 79 worldwide in 2021, with a forecast rise to 783.2 million by 2045. Diabetes affected men and women at almost the same proportions, while people aged 75 to 79 had the highest prevalence. Estimates show that high-income countries had a greater frequency than low-income ones (11.1% vs. 5.5%) and in urban regions (12.1%) than in rural ones (8.3%) in 2021. Over a billion people globally, or more than 10.5% of all adults, now have diabetes, which affects a little over half a billion people [3].

To avoid complications from DR, the WHO and the American Academy of Ophthalmology advise patients with diabetes to have eye exams at least once a year. This early evaluation would prevent progression caused by any examination delay [4]. However, for two critical reasons, this advice is inappropriate in many countries. First, several patients, particularly in developing countries, cannot afford routine eye tests. Secondly, not enough medical professionals to meet the demand for ophthalmologists and not enough screening equipment due to the rise in diabetes patients [5]. Additionally, patients with early-stage DR do not show explicit symptoms or have vision problems.

Due to excessive blood sugar levels, DR is a secondary illness or consequence of diabetes that damages a diabetic person's eyes. Identification of diabetic patients with retinopathy helps to avoid vision loss. Retinopathy can be avoided with early identification and a healthy lifestyle. Through techniques like image processing and deep learning models, artificial intelligence detects DR [6–8]. Proliferative (PrDR) and non-proliferative (NPrDR) retinopathy are the two stages of DR grading. The retina is healthy without DR, and there are three levels of NPrDR: mild, moderate, and severe [6,9].

Transfer learning applies knowledge from one problem to a similar one. It allows a pre-trained model to apply information from a massive quantity of labeled data to a similar task and then fine-tune it to the target task using relatively small amounts of labeled data [10]. Collecting labeled data is both time-consuming and expensive in various practical applications. Transfer learning can solve this problem by starting with a pre-trained model that is already familiar with the core patterns and features of the data. It enables satisfactory results with a substantially smaller amount of labeled data than traditional methods require [11].

This work aims to detect DR from fundus images. It uses three publicly accessible datasets. Four pre-trained models based on CNN—MobileNetV2, VGG16, InceptionV3, and DenseNet121—have been used in this work, with several evaluation parameters, including loss, accuracy, recall, precision, and specificity. Each dataset is implemented with the same transfer learning methods, and the results are analyzed. Finally, the three datasets are combined to form a single data set and are tested on the models. The main objective of this work is as follows:

- To evaluate the performance of models built using MobileNetV2, DenseNet121, VGG16, and InceptionV3 on retinal fundus images.
- To train and test a model using three publicly available datasets and integrate it into a new dataset.

Hence the research question for this study is as follows: How do four pre-trained models (MobileNetV2, VGG16, InceptionV3, and DenseNet121) perform on the combined dataset compared to three independent datasets in terms of accuracy, recall, specificity, and precision for DR classification tasks?

The remaining study structure is classified as the review of the previous studies in Section 2. The materials and methods in Section 3 describe different datasets used, the class imbalance problem, data augmentation, transfer learning, and the model implementation procedure. Section 4 presents the experimental results by analyzing four models with three datasets and an integrated dataset, followed by Section 5′s study discussion. The work is concluded in Section 6 with future enhancements.

2. Related Works

By creating unique features and designing software that analyses retinal images, numerous techniques have been investigated to address the problem of DR grading. This kind of approach is complex and labor-intensive. The ophthalmologist must also use his knowledge of designing traits and labeling retinal pictures. On the other hand, this necessity is one of the significant issues that must be addressed [12]. Most of these traditional approaches are unsuccessful in figuring out the answers and understanding how things fit together. It retains many irrelevant picture-specific features, such as rotation, brightness, and size, which impact performance and accuracy rather than learning valuable data [13]. High-performance optic recognition and classification to identify and separate the afflicted retinal areas have been feasible with the advent of AI approaches such as machine learning (ML) and deep learning (DL) [14].

The basis of the transfer learning strategy theory is that the approach was previously trained on a sizable image database and could be adjusted to fit the necessary set of data [15].

This technique has become popular over the past few years because of the availability of large pre-trained models and the lack of labeled data for many real-world challenges [16]. The Quadratic Weighted Kappa was employed for analyzing the DR prediction in a study by Chilukoti [17]. They used EfficientNet b3 pre-trained model for the classification. In a paper published by Viji [18], different models like VGG, Xception, ResNet, Inception, and EfficientNet were employed and achieved an accuracy of 99.36% with 0.986 recall.

In [19], from the input dataset, features were extracted using a CNN model, and the classification was performed using a support vector machine model. Compared to classifying using the CNN model, this strategy offered faster execution. Medical imaging is crucial at all fundamental health difficulty levels and various medical diseases. DR grading from fundus photos has been the subject of much research using a variety of transfer learning approaches using publicly available datasets. Rahhal et al., [20] employed Inception [21], VGG16 [22], DenseNet [23], MobileNet [24], and ResNet models [25], and VGG16 led to a 100% accuracy rate. The learnable parameters are reduced when multiple layers are stacked without using spatial pooling for CNN models, which are simpler to train [26]. Kothare and Malpe [27] constructed a support vector machine and Naive Bayes model with a binary pattern approach. The algorithm was applied to choose pertinent features. As opposed to this, the models were used to categorize patients as having the condition. Concerning accuracy, execution time, and memory use, the support vector machine performs better than the naive Bayes.

Hussain et al. [28] sought to build an accessible predictor and classifier for hard exudates using artificial neural networks (ANN). Feature extraction and detection were performed using the Speed Up Robust Features technique. For classification, however, Feed-Forward Backpropagation ANN was employed. The work's major flaw is that it only used a dataset containing 48 photos. Ahmad et al., compare multiple designs, including Inception-ResNetV2, ResNet50, NASNet, InceptionV3, VGG16, DenseNet121, Xception, and VGG19 in the article [29]. The models carried out classification and localization tasks, which were trained on a proprietary dataset and tested on the Messidor-2 dataset. The CNN and Contrast Limiting Adaptive Histogram Equalization (CLAHE) approaches increased the area under the curve and the fundus images' quality (23). The CLAHE technique has been introduced, amplifying the retinal vessels to enhance the preprocessing stage.

Various ML classifiers have been merged to enhance DR detection and reduce the error rate [30]. Accurate diagnosis and DR measurement can be performed using fundus images. An ophthalmologist must have much training and dedication to analyze DR using fundus images manually. DR was effectively categorized using deep learning techniques like CNN and transfer learning using models like ResNet, VGG, and GoogleNet [20,31,32]. In a previous article by Fayyaz [33], the authors employed a variety of SVM kernels with Ant Colony System (ACS) feature selection method. It evaluates a technique for detecting DR using 250, 550, and 750 features. The cubic SVM classifier significantly outperformed all other kernels, with an accuracy of 92.6%, 91.8%, and 93%, respectively, for 250, 550, and 750 attributes. With the APTOS-2019 dataset, a different study by Nahiduzzaman [34] utilized an extreme learning machine (ELM) model and reached an accuracy of 97.27%.

A CNN architecture employed for image segmentation tasks is called U-Net. In the work by Jena [35], segmentation was performed utilizing an asymmetric deep learning architecture for DR screening using U-Net networks. CLAHE is used to analyze and improve the green channel images. For APTOS and MESSIDOR, the non-DR detection accuracy was 98.6% and 91.9%, while the PrDR detection accuracy was 96.8% and 95.76%.

Data augmentation is the most popular method for addressing imbalance issues with image categorization [36]. It is a collection of strategies for producing more data pieces from existing data to enhance the quantity of data artificially. The images can be improved in various ways, including cropping the image, inverting it horizontally or vertically, resizing it, and rotating it [37]. In a paper by Mungloo-Dilmohamud [38], using a data augmentation strategy enhances the standard of transfer learning in classifying DR images. Table 1 shows the summary table of the related works in DR.

Table 1. Summary of the related works in DR.

Reference	Datasets	Models	Metrics	Results
[39]	Kaggle data	CNN with grid search	Accuracy	89%
[34]	EyePACS, APTOS	Extreme learning machine	Accuracy	91.78% (EyePACS), 97.27% (APTOS)
[40]	APTOS19, DIARETDB1	Ensemble model-DenseNet101, ResNext	Accuracy	96.98% (binary), 86.08% (multi)
[15]	IDRiD	VGG16, InceptionV3	Accuracy	77.6% with VGG16
[35]	APTOS, MESSIDOR	CNN and SVM	Accuracy	97.25% (APTOS), 94.8% (MESSIDOR)
[41]	ODIR, IDRiD	Vision transformer	Accuracy	82.35%
[33]	APTOS	AlexNet, ResNet-101 with ACS	Accuracy	93%
[17]	EyePACS	Efficientnet-b3	Quadratic Weighted Kappa	0.87
[42]	EyePACS, APTOS	Source-Free Transfer Learning	Accuracy	91.2%
[43]	EyePACS (training), APTOS (testing)	ResNet, EfficientNet, and Swin Transformer	Accuracy	76.35%
[20]	DDR	DenseNet 121, InceptionV3, ResNet153, VGG16, MobileNet, and InceptionResNet	Accuracy	100%
[44]	Kaggle	multi-class feature extraction deep forest	Accuracy	74%
[45]	Own dataset	Inception-v4 with ensemble method	AUC	0.994
[46]	IDRiD	CNN	Accuracy	68%
[31]	DDR, EyePACS, and IDRiD	VGG16	Accuracy	85.4% (DDR), 77.9% (EyePACS), and 89.8% (IDRiD)
[47]	Messidor-2, EyePACS, and DIARETDB0	CNN + SVD + Inception-V3, GoogLeNet, AlexNet, ResNet	Accuracy	94.59% (Messidor-2), 97.92% (EyePACS), and 93.52% (DIARETDB0)
[48]	Messidor-2, EyePACS, and DIARETDB0	U-net + VGGNet	Accuracy	93.95% (Messidor-2), 96.6% (EyePACS), and 92.25% (DIARETDB0)
[49]	Kaggle	CNN	Accuracy	83.5%
[50]	Own dataset	DenseNet-121 + CNN	Accuracy	84.47%
[51]	IDRiD	VGG16	Accuracy	93.1%
[52]	APTOS	ResNet, Inception V3, InceptionResNet-V2, DenseNet-169, Xception, EfficientNet-B4	quadratic weighted kappa	0.824
[53]	APTOS	AlexNet, Res-Net18, SqueezeNet, GoogleNet, VGG16, and VGG19	Accuracy	97.9%
[54]	EyePACS	Inception-V3	Accuracy	90.9%

3. Materials and Methods

Even though ML first appeared in the healthcare industry several years ago, plenty of openings must be filled, and advancements should be considered. Most medical datasets about DR have a few problems. Many datasets only have a few images, which is not enough for the learning process to work effectively. Additionally, many datasets lack labels, forcing ophthalmologists to analyze and categorize the images according to their medical knowledge. Most of the studies have investigated the DR classification on a specific dataset. In this study, three different publicly accessible datasets are employed for the DR classification task, and an integrated dataset is generated by combining the three. The methodology of the study is explained in Figure 1. The study follows a binary classification task.

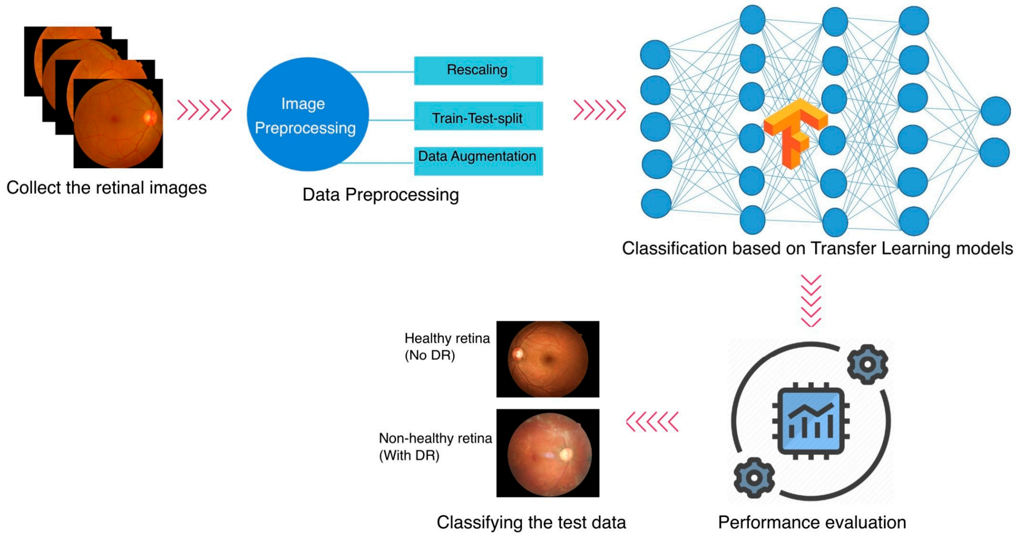
Collect the retinal images
Image Preprocessing
Rescaling
Train-Test-split
Data Augmentation
Data Preprocessing
Classification based on Transfer Learning models
Healthy retina (No DR)
Non-healthy retina (With DR)
Classifying the test data
Performance evaluation

Figure 1. The methodology for DR classification.

3.1. Datasets

Three datasets from India, the United States, and China are used in this study, namely Asia Pacific Tele-Ophthalmology Society (APTOS) [55], Eye Picture Archive Communication System (EyePACS) [56], and Ocular Disease Intelligent Recognition (ODIR) [57] datasets, respectively. For the APTOS and EyePACS datasets, only the normal and PrDr retinal databases are employed in this study. The PrDR is labeled class 1, and the normal retina class 0.

The APTOS DR severity scale is a reliable and valid tool for grading the severity of DR. It has been extensively applied in Asia-Pacific clinical practice and research. It has also been utilized in creating and assessing computer-aided diagnostic tools for DR that employ retinal pictures to identify and categorize the condition's severity automatically. The dataset contains 5990 retinal images, of which 3662 are in the training set and 1928 are in the testing stage. The test phase data has no specific labels provided; hence, the 3662-training data are downloaded for this study. There are 1805 class 0 images and 295 class 1 images within it. The statistics indicate that it is approximately six times as many images in class 0 as in class 1.

A Chinese dataset called ODIR was compiled by Shanggong Medical Technology Co., Ltd., Shanghai, China, from data on 5000 patients at various hospitals. Eight labels comprise this dataset, one normal and the other seven representing various retinal illnesses, including pathological myopia, hypertension, glaucoma, cataract, age-related macular degeneration, DR, and other abnormalities. There are 14,400 images in this dataset. DR is only considered for this study, which consists of 2873 normal retinal images and 1608 with DR.

A free retinopathy screening platform with headquarters in the United States offers the EyePACS dataset. EyePACS has been extensively utilized in numerous DR screening programs, making it possible to test sizable populations efficiently and affordably. There are 35,126 images total for both eyes, divided into five classes, where 73.4% of the images in this dataset are healthy and free of DR symptoms, which is a significant imbalance issue [58]. The work is based on a randomly selected portion of the images. The class distribution of this dataset is 2010 images for class 0 and 708 images for class 1.

3.2. Data Augmentation

If a binary classification project is used, the dataset is deemed unbalanced if the first class's data set is disproportionately more minor than the second [59,60]. The minority class is the one with the least information. In comparison, most of the class is referred to as having a lot of information. The performance of the prediction process is greatly hampered by this disproportion distribution, particularly for the minority class, where it reduces the model's capacity for learning and causes it to be biased. Furthermore, if the minor class makes up less than 5% of the data, there is a severe problem. The accuracy in this situation will be at least 95%, even if the model incorrectly detects the minor class data. In some applications, this circumstance can be acceptable. However, this problem can undermine the model's validity in real-world applications.

Data augmentation methods come in various forms [38]. Nevertheless, geometric transformation is the most well-liked strategy [38,61]. It has therefore been applied to the chosen datasets in this study to tackle the class imbalance challenge. Different transformations have been employed, such as height shift, width shift, horizontal flip, vertical flip, and rotation. Table 2 shows the distribution of data before and after applying data augmentation. The data are combined to make a new dataset. The total count for the integrated data is 6688 for class 0 and 7110 for class 1.

Table 2. The data distribution in each class for the three datasets before and after applying data augmentation.

Dataset	Class 0		Class 1	
	Actual	Data Augmented	Actual	Data Augmented
APTOS	1805	1805	295	1770
EyePACS	2010	2010	708	2124
ODIR	2873	2873	1608	3216

The train-test-split method is chosen with a ratio of 70:30 for model selection. The training set is employed to fit the model or discover the underlying patterns in the data. The testing set examines the model's performance, specifically how effectively it generalizes to novel, untested data. The total number of training images is 1263, 1407, and 2011 for class 0 and 1242, 1488, and 2250 for class 1 concerning APTOS, EyePACS, and ODIR, respectively. Similarly, 542, 603, and 862 for class 0 and 528, 636, and 966 for class 1 are the count for the test set from APTOS, EyePACS, and ODIR, respectively.

3.3. Transfer Learning Models

Transfer learning is frequently employed when training neural networks to produce the best results [42,62]. It can be explained as training a model from start to finish on a dataset with the right amount of data from a related field, then using the target dataset's insufficient data to fine-tune it. Instead of beginning from scratch, this approach initializes the model with a strong foundation [63,64].

One of CNN's main flaws is the requirement for a sizable number of training samples. Its depth and number of parameters decide the data volume needed for a network; the more layers and parameters a network has, the more data it needs [65,66]. Otherwise, there would be a performance-decreasing overfitting issue. However, a significant obstacle in creating DL models for the medical industry is data scarcity. Consequently, it would be advantageous to apply the transfer learning principle to medical projects [12,67].

The second iteration of Google's MobileNet network is known as MobileNetV2. The main goal in creating this network is to keep costs and complexity low, making it easy to use for detection and classification on mobile devices or other devices with constrained resources like memory and energy. Low energy consumption models can benefit medical devices and resource-constrained developing nations. It demonstrates potency in extracting features, segmentation, and object detection [68].

The Visual Geometry Group (VGG) at Oxford University created the model known as VGG16 in 2014. The primary framework VGGNet has given rise to several modifications, including VGG1, VGG13, VGG16, and VGG19. Among those models, the last two versions, VGG16 and VGG19, are considered more common. Medical applications frequently employ the VGG16 architecture to recognize and categorize diverse clinical images [31]. The model's learnable parameter layers are represented by the model's weighted layers, which total 16. Despite having a modest kernel size, it has a very long processing time [19].

Google created InceptionV3 in 2015 as a third iteration of the InceptionV1 network, also known as GoogLeNet, based on the CNN model [69]. The goal of this version is to use less computational power than earlier versions. It has been demonstrated that Inception models are more efficient than VGG regarding memory usage, calculations, and several noticeably reduced parameters [62].

Strong gradient flow, efficient computing, smooth error signal propagation, and reduced feature complexity are benefits of the DenseNet121 model. The vanishing gradient problem is the primary goal of this network's construction for the first time [19]. As the number of layers rises, this issue arises. The more information is lost or deleted, the deeper the network is. DenseNet121 resolves this issue by interconnecting the network's layers [52]. The overview of the transfer learning model is depicted in Figure 2.

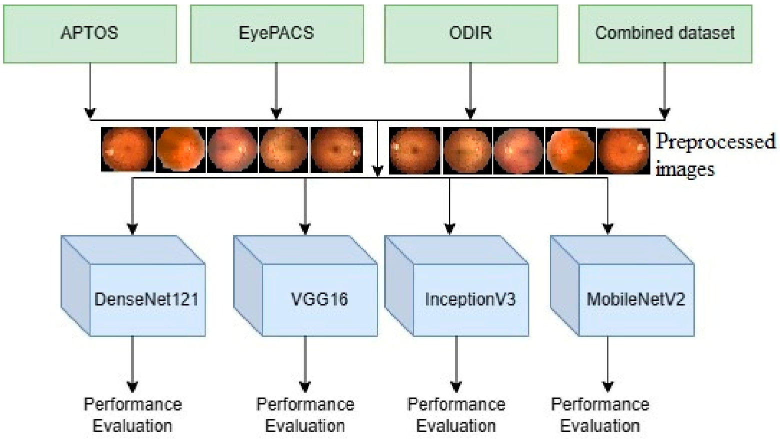

Figure 2. Overview of the transfer learning models.

3.4. Evaluation Metrics

Each data instance—in this case, an image—has four potential outcomes that are categorized, which are true positive (TrPos), true negative (TrNeg), false positive (FaPos), and false negative (FaNeg). A TrPos indicates that the model accurately predicted an optimistic class and a TrNeg suggests that the model accurately predicted a negative class. A FaPos denotes that the model produced an overly optimistic class forecast. Like the FaPos, a FaNeg implies that the model, in error, made a negative class prediction [70].

TrPos and TrNeg are the perfect examples in the medical industry since they involve a patient classification for the ill person and a patient classification for the healthy person. The FaNeg example, however, is the most severe instance where a sick person may be mistakenly labeled as healthy rather than unwell. This misclassification in some disorders can be fatal. In the DR, misdiagnosis can result in significant disease progressions that could eventually result in blindness.

The most widely used statistic for assessing a model's effectiveness at making predictions is its accuracy score (Equation (1)). A loss metric measures the model performance, the sum of mistakes on each occurrence in a training or testing dataset. The precision metric concerns how many positive cases are labeled as positive or how many patients with DR are indeed DR predictions. This score reflects how accurately the model identified the disease from the data. Precision (Equation (2)) is typically utilized when it is necessary to ensure that good forecasts come true.

$$\text{Accuracy} = (\text{TrPos} + \text{TrNeg})/(\text{TrPos} + \text{TrNeg} + \text{FaPos} + \text{FaNeg}), \tag{1}$$

$$\text{Precision} = \text{TrPos}/(\text{TrPos} + \text{FaPos}), \tag{2}$$

On the other hand, the specificity (Equation (3)) meter, or TrNeg rate, displays the number of negative classes accurately identified as such, or in our case, the number of healthy people correctly identified as such. A false alarm with a moderate cost would be to label a healthy person as ill. However, if the condition is serious, telling the person about it could set off a panic attack that could be harmful depending on his physical and mental health. Therefore, in some circumstances, having certainty about the diagnosis is advantageous before making it public, which is what the specificity metric offers.

$$\text{Specificity} = (\text{TrNeg})/(\text{TrNeg} + \text{FaPos}), \tag{3}$$

Among all the real positive cases, recall (also known as sensitivity or TrPos rate) (Equation (4)) determines the positive points that were correctly anticipated. The recall measure is significant and necessary in medical applications where the cost of the incorrect prediction, particularly the false negative instance, is quite expensive, especially when the disease has considerable side effects or is contagious and rapidly spreading.

$$\text{Recall or Sensitivity} = \text{TrPos}/(\text{TrPos} + \text{FaNeg}), \tag{4}$$

4. Results

The python libraries such as TensorFlow [71] and Scikit-learn [72] are used to implement the models in this study. NVidia Titan V GPU is used for DL purposes. VGG16, MobileNetV2, InceptionV3, and DenseNet121 are the four CNN models employed in this work with transfer learning. With a batch size of 32, the total epochs were fixed to 30. In medical image classification, recall measurement is usually regarded as a better performance indicator than accuracy, precision, or specificity because it more accurately depicts the clinical value of classification outcomes [73].

The top, dense, fully connected layers are removed in charge of classifying data to build the final classification layer according to the proposed method. Sigmoid activation is used for binary classification. To solve the overfitting problem, the EarlyStopping and ReduceLRonPlateau methods are utilized in this study as regularisation approaches. Depending on the supplied settings, the model will terminate training once it obtains the best outcome before it starts to overfit. EarlyStopping parameter examines the validation loss, and if it is stable or increasing for five epochs, then the model stops iterating and saves the best model. Similarly, if the loss value stays the same for two successive epochs, the ReduceLRonPlateau method modifies the learning rate by a factor of 0.1. Table 3 depicts the parameters employed in the study.

Table 3. Parameters utilized in this study.

Parameters	Value
Batch size	32
Epochs	30
learning rate	0.001
Optimizer	Adam
Activation	Sigmoid
Loss	binary_crossentropy

Table 4 displays the APTOS's performance with four models. The table shows that the MobileNetV2 and DenseNet121 models have identical recall, precision, and accuracy scores of 0.9850, indicating that they perform equally well on this task. The InceptionV3 model also has a better recall value of 0.9830 but a precision of 0.9683, suggesting a higher true positive rate. The VGG16 model has a much lower recall score of 0.3902 but a high precision score of 0.9952, indicating that it misses more true positive cases. The specificity value of all models is also high, meaning the model correctly identifies negative cases and has fewer false positives. Overall, the APTOS model performs very well, with all models achieving high accuracy scores of at least 0.6981. However, the DenseNet121 and MobileNetV2 models perform best on this task.

Table 4. APTOS model evaluation metrics results.

Model	Recall	Precision	Specificity	Accuracy	Train Loss	Valid Loss
MobileNetV2	0.9850	0.9850	0.9850	0.9850	0.0735	0.0949
DenseNet121	0.9850	0.9850	0.9850	0.9850	0.1927	0.1169
InceptionV3	0.9830	0.9683	0.9686	0.9757	0.0941	1.0060
VGG16	0.3902	0.9952	0.9982	0.6981	0.1231	1.6910

The following dataset, EyePACS, underwent the same process, and the results are illustrated in Table 5. The DenseNet121 model produced the best results with a minor loss and the highest accuracy. During 30 epochs with 64 batch sizes, DenseNet121 achieved scores of 0.1412 in training loss, 0.2933 in validation loss, and 0.8910 in accuracy, recall, precision, and specificity. The MobileNetV2 and InceptionV3 models also have relatively high recall, precision, and accuracy scores, with recall and precision scores of over 0.87 and accuracy scores of above 0.8773. The VGG16 model, on the other hand, has a lower recall score of 0.7154 but a high precision score of 0.9420, indicating that it accurately identifies true positive cases but misses many cases. The accuracy score for the VGG16 model is 0.8313, which is lower than the other models evaluated. The specificity value of all models is also high, meaning the model correctly identifies negative cases and has fewer false positives. Overall, the EyePACS model performs well, with all models achieving high accuracy scores of at least 0.8313. However, the DenseNet121 model is the best performer on this task.

Table 5. EyePACS model evaluation metrics results.

Model	Recall	Precision	Specificity	Accuracy	Train Loss	Valid Loss
MobileNetV2	0.8781	0.8781	0.8781	0.8781	1.6679	2.1374
DenseNet121	0.8910	0.8910	0.8910	0.8910	0.1412	0.2933
InceptionV3	0.8821	0.8793	0.8723	0.8773	0.2352	0.3494
VGG16	0.7154	0.9420	0.9536	0.8313	0.2617	0.5014

Looking at Table 6, the DenseNet121 model has the highest recall, precision, and accuracy scores, all of which are 0.7582. The MobileNetV2 model also has relatively high recall, precision, and accuracy scores, with 0.75. The VGG16 model has a recall score of 0.7743, which is higher than the other models, but a lower precision score of 0.7151, indicating that it identifies false positive cases. The accuracy score for the VGG16 model is 0.7177, the lowest among the models evaluated. The InceptionV3 model has the lowest recall score of 0.6739, indicating that it misses more true positive cases. The specificity values are reasonably high, with a score above 0.65 for all models. Overall, the ODIR model performs reasonably well, with all models achieving accuracy scores of at least 0.6723. However, the DenseNet121 and MobileNetV2 models perform best on this task.

Table 6. ODIR model evaluation metrics results.

Model	Recall	Precision	Specificity	Accuracy	Train Loss	Valid Loss
MobileNetV2	0.7500	0.7500	0.7500	0.7500	0.5072	0.5938
DenseNet121	0.7582	0.7582	0.7582	0.7582	0.4888	0.5601
InceptionV3	0.6739	0.6963	0.6705	0.6723	0.5453	0.6193
VGG16	0.7743	0.7151	0.6543	0.7177	0.5843	0.5480

Table 7 shows the evaluation metrics results for the model trained on a combined dataset, which includes data from the APTOS, ODIR, and EyePACS datasets. The DenseNet121 and VGG16 models have the highest recall, precision, and accuracy scores, all of which are 0.9897 and 0.9879, respectively. The MobileNetV2 model also has high recall, precision, and accuracy scores, with 0.9851. The InceptionV3 model has lower recall and precision scores than the other models, with scores of 0.9715 and 0.9727, respectively. However, its accuracy score of 0.9721 is still high. The specificity value of all models is also high, meaning the model correctly identifies negative cases and has fewer false positives. Overall, the model trained on the combined dataset performs very well, with all models achieving high accuracy scores of at least 0.9721.

Table 7. The model evaluation metrics results for the combined dataset.

Model	Recall	Precision	Specificity	Accuracy	Train Loss	Valid Loss
MobileNetV2	0.9851	0.9851	0.9926	0.9851	0.0109	0.0518
DenseNet121	0.9897	0.9897	0.9948	0.9897	0.0051	0.0368
InceptionV3	0.9715	0.9727	0.9864	0.9721	0.0523	0.0989
VGG16	0.9879	0.9879	0.9939	0.9879	0.0490	0.0465

5. Discussions

The data type plays a crucial role and significantly influences the outcomes. In this study, different datasets are combined regardless of the patient's origin and are trained on four different pre-trained models, namely VGG16, InceptionV3, MobilenetV2, and DenseNet121. The system had undergone proof-of-concept development on a relatively limited scale. It may be expanded (as a future enhancement) to incorporate more datasets and models from many new ethnic groups and nations.

Many researchers have used pretrained transfer learning models to predict and classify DR retinal images. Almost all works are trained on a single dataset. The VGG16, Densenet121, InceptionV3, and MobileNetV2 models are recognized for effective use as a CNN model for DR classification. Rahhal et al. [20] show an accuracy of 100% in classifying DR with the VGG16 model. With five classes in DR, Rocha et al. [31] attained an accuracy of 89.8% and 90.2% for six categories. Bilal et al. [47] trained InceptionV3 with three different datasets and achieved an accuracy of 97.92% with EyePACS, 94.59% in Messidor-2, and 93.52% with the DIARETDB0 dataset. Hagos and Kant [54] showed an accuracy of 90.9% with a binary classification of DR with the InceptionV3 model. Bagadi et al. [74] achieved an accuracy of 95% with the DenseNet121 model for the APTOS dataset in classifying the DR images. Sarki et al. investigated the subject of grading DR [75] using MobileNetV1 and MobileNetV2 as well as other transfer learning techniques. Using MobileNetV2, they could classify with 78.1% accuracy, and using MobileNetV1, only 58.3% accuracy. Sheikh S.O. [76] generated a model with the MobileNetV2 employing a dataset combined by EyePACS, APTOS, and Messidor2 and achieved an accuracy of 91.68%.

Figure 3 depicts the training and validation loss of this study. The APTOS model performs very well, with the DenseNet121 and MobileNetV2 models, with an accuracy of 98.5% and recall of 0.985. There is no massive difference in the loss value of these two models. For the EyePACS dataset, all the models perform better regarding loss value. The evaluation metrics show the DenseNet121 model with an accuracy and recall of 0.891.

Similar is the loss value results for the ODIR dataset. However, VGG16 offers a recall value of 0.7743 and the DenseNet121 with an accuracy of 75.82%.

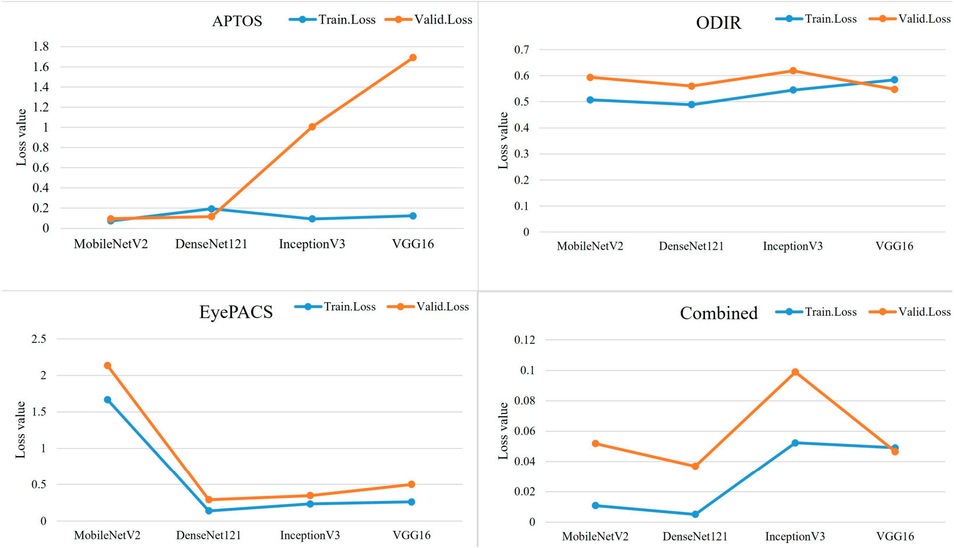

Figure 3. The training and validation loss for four datasets with four pre-trained models.

The combined dataset has all models with better training and validation loss. The DenseNet121 model has the highest recall, precision, and accuracy scores, all of which are 0.9897, which is high compared to Sheikh S.O. [76].

The paired *t*-test compares each model's performance on the combined dataset with their performance on three independent datasets. The alpha or the significant value is set to 0.05. There is a statistically significant difference between the evaluation metrics of the two datasets if the assessed *p*-value is less than 0.05. Hence, the null hypothesis is rejected. The significance is described in Table 8. The combined dataset's performance is compared with three individual datasets (APTOS, EyePACS, and ODIR) using four evaluation metrics (recall, precision, specificity, and accuracy) for DR classification tasks.

Table 8. The *p*-value of the combined dataset with an individual dataset based on a paired *t*-test. The significant value (alpha value) is set to 0.05.

Dataset	Recall	Precision	Specificity	Accuracy
APTOS-Combined	0.397	0.876	0.189	0.389
EyePACS-Combined	0.048	0.008	0.014	0.004
ODIR-combined	0.001	0.000	0.002	0.001

The results show no significant difference in performance for any matrices between the combined and APTOS datasets. However, there is a significant difference between the combined dataset and the EyePACS or ODIR datasets. Overall, these results suggest that the combined dataset performs differently than the EyePACS and ODIR datasets but is like the APTOS dataset. The study also concludes that combining multiple datasets improves performance compared to individual datasets alone.

Plotting the ROC (Receiver Operating Characteristic) Curve is a reliable method of assessing the classification accuracy of a classifier. By charting the True Positive Rate (TPR) and the False Positive Rate (FPR), we can see how the classifier behaves for each threshold. The better the model does in categorizing the data, the closer the ROC curve touches the upper left corner of the figure. To determine how much of the plot lies beneath the curve, we may compute the AUC (area under the curve). The model is better the closer AUC gets to 1 [77]. The AUC score and the ROC curve of the combined dataset evaluated on four models are illustrated in Figure 4. The black diagonal dashed line represent the 50%

area. From the plot, the combined model shows a better classification of DR and normal retinal images with the DenseNet121 model with an AUC score of 0.9947. The AUC score of MobileNetV2 is 0.9897, VGG16 is 0.9608, and InceptionV3 is 0.9781.

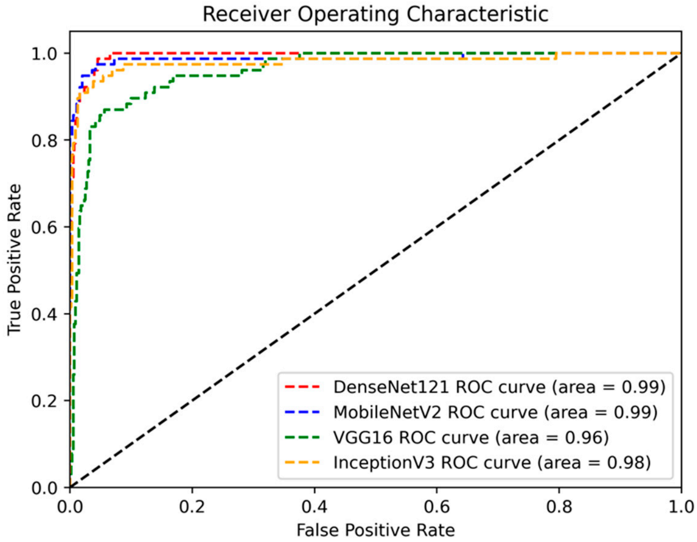

Figure 4. The ROC curve of the four different models evaluated on the combined dataset.

The AUC score of the other three datasets is depicted in Figure 5, the diagonal black dashed line represents the 50% area. The ODIR shows the least performance, and the APTOS has a similar performance to that of the combined (also concluded from the statistical analysis in Table 8). When using DenseNet121, the APTOS has a higher AUC score of 0.9902, and EyePACS has a higher AUC score of 0.9320. With the MobileNetV2 model, ODIR has a score of 0.7278 for AUC.

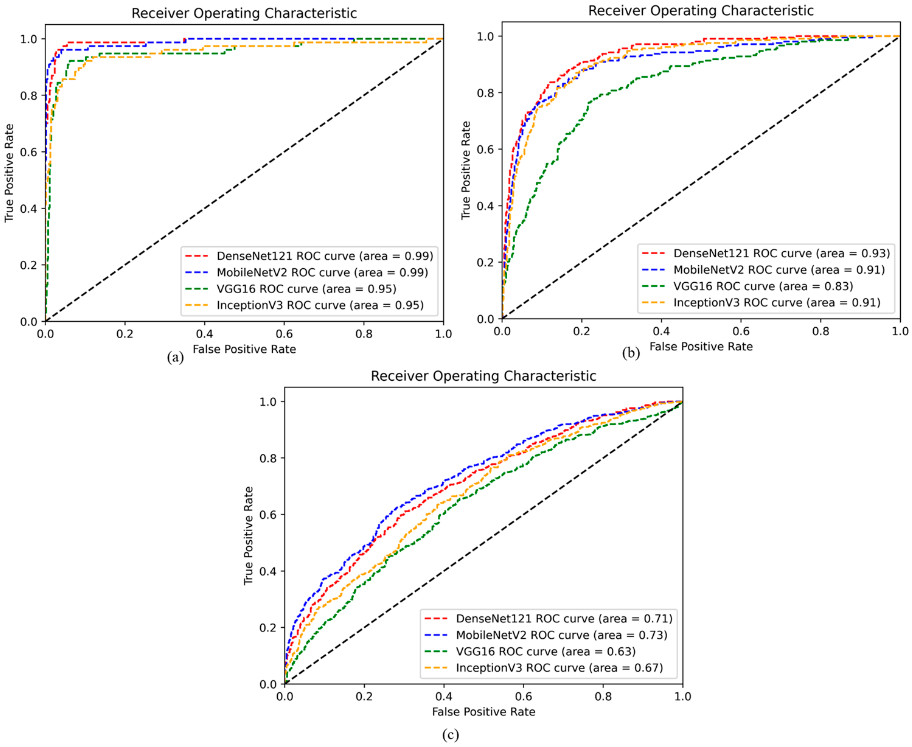

Figure 5. The ROC curve of the four different models evaluated on three datasets. (**a**) APTOS, (**b**) EyePACS, and (**c**) ODIR.

The comparison with previous works is depicted in Table 9. The transfer learning methods used in this study achieved better performance when compared to related works with binary and multi-classification.

Table 9. Comparison with other works related to deep learning models in terms of accuracy.

Reference	Dataset	Model	Label Count	Result
[39]	Kaggle	Grid search CNN	5	89%
[41]	ODIR, IDRiD	Vision transformer	6	82.35%
[40]	DIARETDB1	ResNext + DenseNet	2	96.98%
[33]	APTOS	AlexNet, ResNet-101 with ACS	2	93%
[15]	IDRiD	VGG16	5	77.6%
[17]	EyePACS	Efficientnet-b3	5	92%
[42]	Train EyePACS, Test APTOS	Source-Free Transfer Learning	5	91.2%
[9]	APTOS	CNN	2	95.3%
[20]	DDR	VGG16	5	100%
[78]	APTOS	Few-Shot Learning	2	99.73%
[79]	EyePACS	VGGNet	5	96.6%
[31]	DDR, EyePACS, and IDRiD	VGG16	5	85.4% (DDR) 77.9% (EyePACS) 89.8% (IDRiD)
[47]	Messidor-2, EyePACS, DIARETDB0	CNN + SVD + Inception-V3	5	94.59% (Messidor-2) 97.92% (EyePACS) 93.52% (DIARETDB0)
[50]	Own dataset	DenseNet-121 + CNN	2	84.47%
[80]	EyePACS	VGG16	2	75.7%
[81]	Own dataset	VGG16	3	87.28%
[76]	APTOS + EyePACS + Messidor	MobileNetV2	5	91.68%
[82]	EyePACS	VGGNet-s	5	95.68%
Our study	EyePACS + APTOS + ODIR	DenseNet121	2	98.97%

Evaluating models on multiple datasets can provide insight into how well they generalize to new and diverse data. Each dataset may have biases that might influence the overall findings. Consequently, merging the datasets that were publicly available from people of varying ethnicities is a potential solution to the issue of producing biased results. This information may be used to decide which model is appropriate for a particular task or to pinpoint areas that want development.

6. Conclusions

This study investigated the performance of four distinct transfer learning models with each of the four separate data sets. Using the datasets (APTOS, EyePACS, ODIR, and the combined) with preprocessing and data augmentation approaches, we got encouraging findings on our DR classification model in this study. Pretrained models for classifying DR classes include VGG16, InceptionV3, DenseNet121, and MobileNetV2. The APTOS data has the highest accuracy and recall (98.50%) with MobileNetV2 and DenseNet121 models. The test results of EyePACS showed that DenseNet121 has the highest accuracy and recall of 89.10%, while ODIR shows 75.82% accuracy with DenseNet121 and 0.7743 recall value with VGG16. In the combined dataset the DenseNet121 model outperforms other models with high accuracy, recall, specificity, and precision (98.97%). The statistical results suggest that the combined dataset performs differently than the EyePACS and ODIR datasets (p-value less than 0.05) but is like the APTOS dataset (p-value greater than 0.05). Using pre-trained models and transfer learning can increase the effectiveness and accuracy of DL models while requiring less training data and resources. It can be particularly beneficial in applications such as medical diagnosis. The established approach can be widely adopted to accommodate additional screenings for diabetic patients performed by clinics to identify DR and refer them to an ophthalmologist to begin the right course of treatment before it progresses into blindness.

Even though the transfer learning strategy described in this work for the categorization of DR has shown some encouraging findings, there are limitations to the work that needs to be taken into consideration. The study used only a small dataset size and evaluated balanced data. Also, it is not clinically validated. The study may be expanded (as a future enhancement) to incorporate more datasets and models from many new ethnic groups and nations. Also, the model can be deployed into a mobile application as a future enhancement to test with actual patient data.

Author Contributions: Conceptualization, A.M.M.; data curation, S.A.; methodology, A.M.M., S.A. and S.S.; software, S.A.; writing—original draft preparation, S.S.; writing—review and editing, A.M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Kuwait University Research Grant No. EO04/18.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: [55,57,58].

Acknowledgments: The authors would like to thank Kuwait University research sector for the continuous support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Simó, R.; Simó-Servat, O.; Bogdanov, P.; Hernández, C. Diabetic retinopathy: Role of neurodegeneration and therapeutic perspectives. *Asia-Pac. J. Ophthalmol.* **2022**, *11*, 160–167. [CrossRef] [PubMed]
2. World Health Organization. Diabetes. Available online: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 21 November 2022).
3. Sun, H.; Saeedi, P.; Karuranga, S.; Pinkepank, M.; Ogurtsova, K.; Duncan, B.B.; Stein, C.; Basit, A.; Chan, J.C.N.; Mbanya, J.C.; et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin. Pract.* **2022**, *183*, 109119. [CrossRef] [PubMed]
4. Rajini, H.N. A Novel Approachfor the Diagnosis of Diabetic Retinopathy Using Convolutional Neural Network. In Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 15–16 March 2019; pp. 1102–1107.
5. Pires, R.; Avila, S.; Wainer, J.; Valle, E.; Abramoff, M.D.; Rocha, A. A data-driven approach to referable diabetic retinopathy detection. *Artif. Intell. Med.* **2019**, *96*, 93–106. [CrossRef]
6. Sesikala, B.; Harikiran, J.; SaiChandana, B. A Study on Diabetic Retinopathy Detection, Segmentation and Classification Using Deep and Machine Learning Techniques. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1419–1424.
7. Sri, K.S.; Priya, G.K.; Kumar, B.P.; Sravya, S.D.; Priya, M.B. Diabetic Retinopathy Classification Using Deep Learning Technique. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1492–1496.
8. Vipparthi, V.; Rao, D.R.; Mullu, S.; Patlolla, V. Diabetic Retinopathy Classification Using Deep Learning Techniques. In Proceedings of the 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 17–19 August 2022; pp. 840–846.
9. Thomas, N.M.; Jerome, S.A. Grading and Classification of Retinal Images for Detecting Diabetic Retinopathy Using Convolutional Neural Network. In *Advances in Electrical and Computer Technologies*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 607–614.
10. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
11. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1345–1459. [CrossRef]
12. Asiri, N.; Hussain, M.; Al Adel, F.; Alzaidi, N. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artif. Intell. Med.* **2019**, *99*, 101701. [CrossRef]
13. Qummar, S.; Khan, F.G.; Shah, S.; Khan, A.; Shamshirband, S.; Rehman, Z.U.; Khan, I.A.; Jadoon, W. A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection. *IEEE Access* **2019**, *7*, 150530–150539. [CrossRef]
14. Sebastian, A.; Elharrouss, O.; Al-Maadeed, S.; Almaadeed, N. A Survey on Deep-Learning-Based Diabetic Retinopathy Classification. *Diagnostics* **2023**, *13*, 345. [CrossRef]
15. Jiwani, N.; Gupta, K.; Sharif, M.H.U.; Datta, R.; Habib, F.; Afreen, N. Application of Transfer Learning Approach for Diabetic Retinopathy Classification. In Proceedings of the 2023 International Conference on Power Electronics and Energy (ICPEE), Bhubaneswar, India, 3–5 January 2023; pp. 1–4.
16. Torrey, L.; Shavlik, J. Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI Global: Hershey, PA, USA, 2010; pp. 242–264.

17. Chilukoti, S.V.; Shan, L.; Maida, A.S.; Hei, X. A Reliable Diabetic Retinopathy Grading via Transfer Learning with Quadratic Weighted Kappa Metric. *Res. Sq.* **2023**. [CrossRef]
18. Vij, R.; Arora, S. A novel deep transfer learning based computerized diagnostic Systems for Multi-class imbalanced diabetic retinopathy severity classification. *Multimed. Tools Appl.* 2023, *in press*. [CrossRef]
19. Qomariah, D.U.N.; Tjandrasa, H.; Fatichah, C. Classification of Diabetic Retinopathy and Normal Retinal Images Using CNN and SVM. In Proceedings of the 2019 12th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 18 July 2019; pp. 152–157.
20. Rahhal, D.; Alhamouri, R.; Albataineh, I.; Duwairi, R. Detection and Classification of Diabetic Retinopathy Using Artificial Intelligence Algorithms. In Proceedings of the 2022 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 21–23 June 2022; pp. 15–21.
21. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
22. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014. [CrossRef]
23. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
24. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Doshi, D.; Shenoy, A.; Sidhpura, D.; Gharpure, P. Diabetic Retinopathy Detection Using Deep Convolutional Neural Networks. In Proceedings of the Computing, Analytics and Security Trends (CAST), International Conference on Computing, Analytics and Security Trends (CAST), Pune, India, 11 July 2016; pp. 261–266.
27. Kothare, K.S.; Malpe, K. Design and Implementation of Inspection Model for knowledge Patterns Classification in Diabetic Retinal Images. In Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 27–29 March 2019; pp. 1220–1223.
28. Hussain, M.A.; Islam, S.O.B.; Tiwana, M.; Qureshi, W. Detection and Classification of Hard Exudates with Fundus Images Complements and Neural Networks. In Proceedings of the 2019 5th International Conference on Control, Automation and Robotics (ICCAR), Beijing, China, 19–22 April 2019; pp. 206–211.
29. Ahmad, M.; Kasukurthi, N.; Pande, H. Deep Learning for Weak Supervision of Diabetic Retinopathy Abnormalities. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 573–577.
30. Bilal, A.; Sun, G.; Li, Y.; Mazhar, S.; Khan, A.Q. Diabetic Retinopathy Detection and Classification Using Mixed Models for a Disease Grading Database. *IEEE Access* **2021**, *9*, 23544–23553. [CrossRef]
31. da Rocha, D.A.; Ferreira, F.M.F.; Peixoto, Z.M.A. Diabetic retinopathy classification using VGG16 neural network. *Res. Biomed. Eng.* **2022**, *38*, 761–772. [CrossRef]
32. Nadeem, M.W.; Goh, H.G.; Hussain, M.; Liew, S.-Y.; Andonovic, I.; Khan, M.A. Deep learning for diabetic retinopathy analysis: A review, research challenges, and future directions. *Sensors* **2022**, *22*, 6780. [CrossRef]
33. Fayyaz, A.M.; Sharif, M.I.; Azam, S.; Karim, A.; El-Den, J. Analysis of Diabetic Retinopathy (DR) Based on the Deep Learning. *Information* **2023**, *14*, 30. [CrossRef]
34. Nahiduzzaman, M.; Islam, M.R.; Goni, M.O.F.; Anower, M.S.; Ahsan, M.; Haider, J.; Kowalski, M. Diabetic Retinopathy Identification Using Parallel Convolutional Neural Network Based Feature Extractor and ELM Classifier. *Expert Syst. Appl.* **2023**, *217*, 119557. [CrossRef]
35. Jena, P.K.; Khuntia, B.; Palai, C.; Nayak, M.; Mishra, T.K.; Mohanty, S.N. A Novel Approach for Diabetic Retinopathy Screening Using Asymmetric Deep Learning Features. *Big Data Cogn. Comput.* **2023**, *7*, 25. [CrossRef]
36. Van Dyk, D.A.; Meng, X.-L. The art of data augmentation. *J. Comput. Graph. Stat.* **2001**, *10*, 1–50. [CrossRef]
37. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
38. Mungloo-Dilmohamud, Z.; Heenaye-Mamode Khan, M.; Jhumka, K.; Beedassy, B.N.; Mungloo, N.Z.; Peña-Reyes, C. Balancing Data through Data Augmentation Improves the Generality of Transfer Learning for Diabetic Retinopathy Classification. *Appl. Sci.* **2022**, *12*, 5363. [CrossRef]
39. Rajamani, S.; Sasikala, S. Artificial Intelligence Approach for Diabetic Retinopathy Severity Detection. *Informatica* **2023**, *46*. [CrossRef]
40. Mondal, S.S.; Mandal, N.; Singh, K.K.; Singh, A.; Izonin, I. EDLDR: An Ensemble Deep Learning Technique for Detection and Classification of Diabetic Retinopathy. *Diagnostics* **2023**, *13*, 124. [CrossRef] [PubMed]
41. Gu, Z.; Li, Y.; Wang, Z.; Kan, J.; Shu, J.; Wang, Q. Classification of Diabetic Retinopathy Severity in Fundus Images Using the Vision Transformer and Residual Attention. *Comput. Intell. Neurosci.* **2023**, *2023*, 1305583. [CrossRef] [PubMed]
42. Zhang, C.; Lei, T.; Chen, P. Diabetic retinopathy grading by a source-free transfer learning approach. *Biomed. Signal Process. Control* **2022**, *73*, 103423. [CrossRef]
43. Xu, S.; Huang, Z.; Zhang, Y. Diabetic Retinopathy Progression Recognition Using Deep Learning Method. Available online: http://cs231n.stanford.edu/reports/2022/pdfs/20.pdf (accessed on 21 November 2022).

44. Qin, X.; Chen, D.; Zhan, Y.; Yin, D. Classification of diabetic retinopathy based on improved deep forest model. *Biomed. Signal Process. Control* **2022**, *79*, 104020. [CrossRef]
45. Li, F.; Wang, Y.; Xu, T.; Dong, L.; Yan, L.; Jiang, M.; Zhang, X.; Jiang, H.; Wu, Z.; Zou, H. Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs. *Eye* **2022**, *36*, 1433–1441. [CrossRef]
46. Jiwani, N.; Gupta, K.; Afreen, N. A Convolutional Neural Network Approach for Diabetic Retinopathy Classification. In Proceedings of the 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), Indore, India, 23–24 April 2022; pp. 357–361.
47. Bilal, A.; Zhu, L.; Deng, A.; Lu, H.; Wu, N. AI-Based Automatic Detection and Classification of Diabetic Retinopathy Using U-Net and Deep Learning. *Symmetry* **2022**, *14*, 1427. [CrossRef]
48. Bilal, A.; Sun, G.; Mazhar, S.; Imran, A.; Latif, J. A Transfer Learning and U-Net-based automatic detection of diabetic retinopathy from fundus images. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2022**, *10*, 663–674. [CrossRef]
49. Saranya, P.; Umamaheswari, K.M.; Sivaram, M.; Jain, C.; Bagchi, D. Classification of Different Stages of Diabetic Retinopathy Using Convolutional Neural Networks. In Proceedings of the 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 19–21 January 2021; pp. 59–64.
50. Islam, M.T.; Al-Absi, H.R.H.; Ruagh, E.A.; Alam, T. DiaNet: A Deep Learning Based Architecture to Diagnose Diabetes Using Retinal Images Only. *IEEE Access* **2021**, *9*, 15686–15695. [CrossRef]
51. Goel, S.; Gupta, S.; Panwar, A.; Kumar, S.; Verma, M.; Bourouis, S.; Ullah, M.A. Deep Learning Approach for Stages of Severity Classification in Diabetic Retinopathy Using Color Fundus Retinal Images. *Math. Probl. Eng.* **2021**, *2021*, 1–8. [CrossRef]
52. Al-Smadi, M.; Hammad, M.; Baker, Q.B.; Sa'ad, A. A transfer learning with deep neural network approach for diabetic retinopathy classification. *Int. J. Electr. Comput. Eng.* **2021**, *11*, 3492. [CrossRef]
53. Khalifa, N.E.M.; Loey, M.; Taha, M.H.N.; Mohamed, H.N.E.T. Deep Transfer Learning Models for Medical Diabetic Retinopathy Detection. *Acta Inform. Med.* **2019**, *27*, 327–332. [CrossRef] [PubMed]
54. Hagos, M.T.; Kant, S. Transfer learning based detection of diabetic retinopathy from small dataset. *arXiv* **2019**. [CrossRef]
55. APTOS 2019 Blindness Detection. Available online: https://www.kaggle.com/competitions/aptos2019-blindness-detection/overview (accessed on 5 September 2022).
56. Cuadros, J.; Bresnick, G. EyePACS: An adaptable telemedicine system for diabetic retinopathy screening. *J. Diabetes Sci. Technol.* **2009**, *3*, 509–516. [CrossRef]
57. Ocular Disease Intelligent Recognition ODIR-5K. Available online: https://odir2019.grand-challenge.org/ (accessed on 5 September 2022).
58. Diabetic Retinopathy Detection. Available online: https://www.kaggle.com/c/diabetic-retinopathy-detection (accessed on 5 September 2022).
59. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
60. Lango, M.; Stefanowski, J. What makes multi-class imbalanced problems difficult? An experimental study. *Expert Syst. Appl.* **2022**, *199*, 116962. [CrossRef]
61. Agarwal, S.; Bhat, A. A survey on recent developments in diabetic retinopathy detection through integration of deep learning. *Multimed. Tools Appl.* **2022**, *82*, 17321–17351. [CrossRef]
62. Zeng, X.; Chen, H.; Luo, Y.; Ye, W. Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network. *IEEE Access* **2019**, *7*, 30744–30753. [CrossRef]
63. Gao, Z.; Li, J.; Guo, J.; Chen, Y.; Yi, Z.; Zhong, J. Diagnosis of diabetic retinopathy using deep neural networks. *IEEE Access* **2018**, *7*, 3360–3370. [CrossRef]
64. Costa, P.; Araújo, T.; Aresta, G.; Galdran, A.; Mendonça, A.M.; Smailagic, A.; Campilho, A. EyeWes: Weakly Supervised Pre-Trained Convolutional Neural Networks for Diabetic Retinopathy Detection. In Proceedings of the 2019 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 27–31 May 2019; pp. 1–6.
65. Elsharkawy, M.; Elrazzaz, M.; Sharafeldeen, A.; Alhalabi, M.; Khalifa, F.; Soliman, A.; Elnakib, A.; Mahmoud, A.; Ghazal, M.; El-Daydamony, E.; et al. The Role of Different Retinal Imaging Modalities in Predicting Progression of Diabetic Retinopathy: A Survey. *Sensors* **2022**, *22*, 3490. [CrossRef] [PubMed]
66. Saeed, F.; Hussain, M.; Aboalsamh, H.A. Automatic Diabetic Retinopathy Diagnosis Using Adaptive Fine-Tuned Convolutional Neural Network. *IEEE Access* **2021**, *9*, 41344–41359. [CrossRef]
67. Kandel, I.; Castelli, M. Transfer learning with convolutional neural networks for diabetic retinopathy image classification. A review. *Appl. Sci.* **2020**, *10*, 2021. [CrossRef]
68. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
69. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
70. Erickson, B.J.; Kitamura, F. Magician's corner: 9. Performance metrics for machine learning models. *Radiol. Artif. Intell.* **2021**, *3*, e200126. [CrossRef]

71. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**. [CrossRef]
72. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
73. Anwar, S.M.; Majid, M.; Qayyum, A.; Awais, M.; Alnowami, M.; Khan, M.K. Medical image analysis using convolutional neural networks: A review. *J. Med. Syst.* **2018**, *42*, 226. [CrossRef]
74. Bagadi, L.; Pavankumar, E.; Likitha, A.; Niranjan, K.; Nani, B. Comparative Analysis of Different Models for Diabetic Retinopathy Classification. In *Innovations in Electronics and Communication Engineering*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 253–258.
75. Sarki, R.; Michalska, S.; Ahmed, K.; Wang, H.; Zhang, Y. Convolutional neural networks for mild diabetic retinopathy detection: An experimental study. *bioRxiv* **2019**. [CrossRef]
76. Sheikh, S.O. Diabetic Reinopathy Classification Using Deep Learning. Available online: https://qspace.qu.edu.qa/bitstream/handle/10576/15230/Sarah%20Obaid%20Sheikh%20_OGS%20Approved%20Thesis.pdf?sequence=1&isAllowed=y (accessed on 19 December 2022).
77. Fan, J.; Upadhye, S.; Worster, A. Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Med.* **2006**, *8*, 19–20. [CrossRef]
78. Murugappan, M.; Prakash, N.; Jeya, R.; Mohanarathinam, A.; Hemalakshmi, G. A Novel Attention Based Few-shot Classification Framework for Diabetic Retinopathy Detection and Grading. *Measurement* **2022**, *200*, 111485. [CrossRef]
79. Jabbar, M.K.; Yan, J.; Xu, H.; Ur Rehman, Z.; Jabbar, A. Transfer Learning-Based Model for Diabetic Retinopathy Diagnosis Using Retinal Images. *Brain Sci.* **2022**, *12*, 535. [CrossRef] [PubMed]
80. Khaled, O.; El-Sahhar, M.; El-Dine, M.A.; Talaat, Y.; Hassan, Y.M.I.; Hamdy, A. Cascaded Architecture for Classifying the Preliminary Stages of Diabetic Retinopathy. In Proceedings of the 9th International Conference on Software and Information Engineering, Cairo, Egypt, 11–13 November 2021; pp. 108–112.
81. Le, D.; Alam, M.; Yao, C.K.; Lim, J.I.; Hsieh, Y.-T.; Chan, R.V.P.; Toslak, D.; Yao, X. Transfer Learning for Automated OCTA Detection of Diabetic Retinopathy. *Transl. Vis. Sci. Technol.* **2020**, *9*, 35. [CrossRef] [PubMed]
82. Wan, S.; Liang, Y.; Zhang, Y. Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Comput. Electr. Eng.* **2018**, *72*, 274–282. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.