Transfer Learning based Classification of Diabetic Retinopathy on the Kaggle EyePACS dataset

Maria Tariq, Vasile Palade, and YingLiang Ma

Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, The United Kingdom
tariqm16@uni.coventry.ac.uk, ab5839@coventry.ac.uk, ac7020@coventry.ac.uk

Abstract. Severe stages of diabetes can eventually lead to an eye condition called diabetic retinopathy. It is one of the leading causes of temporary visual disability and permanent blindness. There is no cure for this disease other than a proper treatment in the early stages. Five stages of diabetic retinopathy are discussed in this paper that need to be detected followed by a proper treatment. Transfer learning is used to detect the grades of diabetic retinopathy in eye fundus images, without training from scratch. The Kaggle EyePACS dataset is one of the largest datasets available publicly for experimentation. In our work, an extensive study on the Kaggle EyePACS dataset is carried out using pre-trained models ResNet50 and DenseNet121. The Aptos dataset is also used in comparison with this dataset to examine the performance of the pre-trained models. Different experiments are performed to analyze the images from the different classes in the Kaggle EyePACS dataset. This dataset has significant challenges including image noise, imbalanced classes, and fault annotations. Our work highlights potential problems within the dataset and the conflicts between the classes. A clustering technique is used to get informative images from the normal class to improve the model's accuracy to 70%.

Keywords: Kaggle EyePACS, Pre-trained models, Transfer learning, Deep Learning, Fine tuning

1 Introduction

Diabetic retinopathy (DR) is an eye complication that can be developed in diabetes, as high blood sugar levels in diabetes damage the eye's retina with time. There are two types of diabetes; Type 1, in which the body does not produce insulin, and Type 2, in which the body produces insulin but does not know how to use it [1]. DR is one of the primary causes of the rise in blindness globally. According to the [1], 422 million adults (aged 20 to 79 years) in 2014 suffered from Type 2 diabetes. Both Type 1 and Type 2 patients are at potential risk of having DR. The population increased to 463 million in 2019 and was predicted to increase to 700 million adults by 2045 [2]. In 2015, there were 2.6 million

people that were visually disabled because of DR, and it is expected to rise to 3.2 million by 2020 [3], making DR the leading cause of preventable blindness. The DR is reversible if proper treatment is carried out in the early stages, but there is no permanent cure for this ailment in the later stages [4].

DR can be categorized into five stages; normal, mild, moderate, severe or non-proliferative, and proliferative [5]. It progresses slowly through these stages without proper screening and treatment. During DR, different lesions start appearing gradually in the eye, like microaneurysms in mild DR [6], hemorrhages and exudates in the moderate DR, formation of new blood vessels in non-proliferative DR, and fragile blood vessels and scar tissues in proliferative DR [5]. These lesions slowly distort the retina and further harm the macula. Regular screening and proper treatment after diagnosis are required to prevent this eye-threatening disease [7]. Detection of small lesions is difficult in the initial stages, but it can be very helpful in reducing the risk of severity. The other thing is the correct diagnosis of all five stages of DR to get proper treatment [8]. Human experts and ophthalmologists are available to manually diagnose the signs of DR, which is time-consuming and qualitative. In recent years, much work has been done on the automated detection of DR with the development of relevant technologies [9].

Deep Learning (DL) is an essential tool for processing medical images for classification, object detection [10], and localization [11]. It uses Convolutional Neural Networks (CNNs) to extract features from the images automatically and then distinguishes between images of different classes [12]. In our work, in-depth research on the Kaggle EyePACS dataset is performed to analyze the behavior of the largest available DR dataset. The eye fundus images are first processed through computer vision using different techniques to improve the quality of images. Pre-trained models like ResNet50 and DenseNet121 are trained through transfer learning for multiclass classification to assist human experts in diagnosis. Aptos dataset is used in comparison with the EyePACS dataset to investigate the performance of the developed classification models. In this paper, all experiments are mainly carried out on the Kaggle EyePACS dataset, which has five classes of diabetic retinopathy, as shown in Fig. 1. During classification, many challenges of the EyePACS dataset, such as noise, incorrect labeling, and imbalanced classes, are highlighted. However, this paper focused on the behavior of this dataset, conflicted classes within the dataset, and the potential steps taken to train the model and increase its performance.

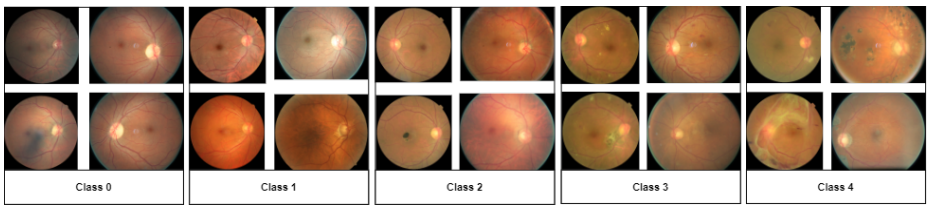

Fig. 1: Images of the five classes of the Kaggle EyePACS dataset.

2 Related Work

Convolutional neural networks get along well with images but need much time for training [13]. Meanwhile, transfer learning was introduced to achieve better accuracy in less time. It is used to train a previously trained model on an entirely different problem by transferring its learning. The model does not need to be trained from scratch; instead, it learns new data in less time and with reasonable accuracy. GoogLeNet and AlexNet have been used for transfer learning on the Messidor dataset [14]. They have done three experiments with two, three, and four classes to get a test accuracy of 74.5%, 68.8%, and 57.2%, respectively. They have also hypothesized that low accuracy in four classes is due to noise and incorrect labeling [14]. In [8], authors have used Inception-v3 for transfer learning. They have trained their model to do binary classification with a small dataset and managed to get an accuracy of 90.9% with 3.94% of the loss. Inception modules are considered to extract differently sized features of input images in one level of convolution [8]. So, Gulshan et al. have also used Inception-v3 to train their model on binary classification. The model is trained on 0 and 1 as one class and 2,3,4 as another class to suggest if the patient needs a referral or not [15].

While working with the Kaggle EyePACS dataset in [5], authors have used data preprocessing and some traditional data augmentation techniques. They have performed two binary classifications; one with healthy (0) and sick (1,2,3,4 classes), and the second with low (0,1) and high (2,3,4 classes). For first classification, they have 94.5% sensitivity and 90.2% specificity. For the second, they have got 98% sensitivity and 94% specificity. For five classes, they have obtained 0.85 of Quadratic Weighted Kappa and 0.74 of F1-score on their test set. In [16], authors have developed a CNN-based system of DR classification using AlexNet, VGG16, and InceptionNet-V3. They have used the Kaggle EyePACS dataset and mentioned the problems within this dataset. The images were handpicked by domain experts to avoid the false labelling of the dataset and achieved a 5-fold cross-validation with the average classification accuracy of 37.43%, 50.03% and 63.23% on AlexNet, VGG16, and InceptionNet-V3, respectively. In [17], authors have trained and tested their model on the Kaggle EyePACS dataset. They have achieved a relatively good accuracy of 70%, but on the skewed dataset with the majority of images in class 0. In [18], authors have done a predictive analysis on the Kaggle dataset using transfer learning techniques. It is relatively similar to our work, in which we will perform an intensive analysis of the eye fundus images from the Kaggle EyePACS dataset through different experiments using pre-trained models.

3 Pre-trained Models

Two pre-trained models were mainly used for the majority of experiments; ResNet50 and DenseNet121. ResNet50 was introduced with the increased network depth to train more and achieve a reasonable accuracy on the images. We

have achieved 92.1% top-5 accuracy and 3.57% top-5 error on ImageNet validation dataset. The architecture of the model is updated and combined with two dense layers for five class classification. DenseNet121 has more depth but slightly less accuracy than Inception-v3, which is 92.3%, and the top-5 error is 7.83% on ImageNet validation dataset. The DenseNet has dense connections between layers, fewer parameters, high accuracy, higher computational efficiency, and memory efficiency. This network advanced the previously developed network ResNet and improves its performance. Like the identity block of ResNet, this network uses a "dense block". The architecture of the DenseNet121 model is updated, where the base model is combined with the average global pooling layer and dense layer for five class classification in our DR detection problem.

4 Dataset

The Kaggle dataset EyePACS was sponsored by the California Healthcare Foundation in 2015, where they launched this competition with the support of a data science team to introduce artificial intelligence in the detection of Diabetic Retinopathy. The images were provided by EyePACS, which is a free platform for retinopathy screening. It consists of 88,696 images, which includes 35,126 images that are annotated for training. Labels are given on the scale of 0-4, which represent the grades of Diabetic Retinopathy. Label 0 shows normal class which includes 25810 images, Label 1 shows mild symptoms of DR which includes 2443 images, Label 2 is moderate DR which includes 5292 images, Label 3 shows symptoms of severe DR and has 873 images, and finally Label 4 shows proliferative DR with 708 images. These grades are given according to the standards of International Clinical Diabetic Retinopathy severity scale by a single specialist. The resolution of images is variable and approximately 3000 x 2000 pixels.

The other dataset we have used is Aptos 2019 (4th Asia Pacific Tele Ophthalmology Society Symposium). APTOS includes 5590 images, 3662 for training and 1928 for testing (Kassani et al., 2019). A clinician has rated each image with the same severity of diabetic retinopathy as in the EyePACS dataset. The number of images is 1805 in the normal class, 370 in the mild class, 999 in the moderate class, 193 in the severe and 295 in the proliferative class. The resolution of images is variable.

5 Methodology

In the proposed method, the dataset EyePACS is taken from the Kaggle public repository. This dataset contains images of different resolutions and grades in an excel file. A desktop PC with Nvidia Tesla K80 GPU was used to train the five classes of DR. TensorFlow was used as backend framework.

Data must be preprocessed to remove noise from the dataset and then fed into the pre-trained model for further training. Some preprocessing techniques were applied, which are discussed in this section. The Diabetic Retinopathy images

were cropped to the input size of the model, which varies from model to model. For ResNet50, we need 224 x 224 which is quite low, but for DenseNet121, we have changed the input layer of the model to accept the images of custom size 512 x 512.

5.1 Transfer Learning details

Following are the hyper-parameter details used in these transfer learning experiments.

Loss Function Several experiments have been conducted using two different loss functions. Categorical crossentropy loss is used for multiclass classification, but it did not perform well on our dataset due to the imbalanced nature of the dataset or small lesions in the images. The loss function is given below:

$$Loss = -\sum_{i=1}^{n} y_i \cdot \log(\hat{y_i})$$

This loss function shows the error between the actual and the predicted output. y_i is the probability for event i, which in total equals 1. n is the number of predictions in the output list.

Sparse Categorical Focal Loss is an extension to categorical crossentropy with the weighting factor $(1 - \hat{y_i})$. γ is the focusing factor used to adjust the rate smoothly. This focal loss works better if the dataset is imbalanced and if there are small lesions within the classes. In this work, focal loss is used with gamma equals to 2. The loss function is below:

$$Loss = \sum_{i=1}^{n} (1 - \hat{y_i})^{\gamma} \cdot y_i \cdot \log(\hat{y_i})$$

Early Stopping Early stopping is used to stop training automatically based on some metric. The metric is usually the validation accuracy or loss that needs to be achieved for the performance evaluation of the model. When this metric stops improving after some epochs, it waits until reaches the value of patience. Patience is the number of epochs without any improvement in the metric. After these epochs, it automatically terminates the training cycle. It increases the model's performance by avoiding overfitting and saving time. The metric used in this work is validation accuracy and the patience value is 70.

Optimizer and Learning rate An optimizer calculates the change after each training cycle and updates the model's weights. It minimizes the loss value to increase the accuracy. We have tested two optimizers, stochastic gradient descent (SGD) and adam optimizer. SGD is calculated by going through all the training examples. This optimizer did not work for our work; however, the Adam optimizer works well and converges faster for our problem. It has less computation

time and needs fewer parameters to tune. The learning rate is set to 0.001, which is considered the best to train the model.

Model Layers In the base model, the initial layers of the model have not been trained and frozen to fine-tune the model. Only the last few layers have been trained to extract informative features from the images. After the base model, the global average pooling layer is used to down-sample a patch's features by taking average values from the feature map. It also reduces the problem of overfitting by learning invariant features. We have used Softmax as an activation function [19], which is used to transform the output before calculating loss in the training cycle. Softmax is used with a dense layer of 5 neurons, and each neuron represents each class.

5.2 Training using Pre-trained models:

EyePACS dataset is the one with the most number of images, but it has a lot of noise, imbalanced classes, and false annotations. We will look into the problems of the Kaggle EyePACS dataset through the conducted experiments.

Experiment on conflict classes: This dataset has two major classes, Class 0 and Class 2, with 25810 and 5292 images, respectively. It was considered better to train the majority classes initially and analyze the results. We resized our input images to 224 x 224 for ResNet50 and randomly down-sampled Class 0 to 5292. The highest accuracy in the two classes was 51%, and the accuracy seemed to be stuck at 50% in the subsequent epochs, which can be seen in Table. 1.

Table 1: Conflicting Classes

	Classes	Model	Epochs	Accuracy	Class 0	Class 2
Exp 1	0 and 2	ResNet50	120	51%	0.61	0.33
	0 and 2	ResNet50	200	50%	0.32	0.61
Exp 2	0 and 1	ResNet50	260	50%	0.26	0.62
	0 and 1	ResNet50	280	52%	0.23	0.65

The same experiment was repeated on Class 0 and Class 1; Class 1 is the next majority class and has 2443 images, so Class 0 was randomly down-sampled to 2443 images. The model responded similarly to Class 0 and Class 1 as the accuracy stuck at 51%. We can say that Class 0 (normal) conflicts with class 1 and class 2. There can be two reasons for this conflict: a mixing between these classes with faults in the annotations, or the model is not good enough to learn

small lesions in the initial stages of DR. If we combine conflict classes 0, 1, and 2 as one Majority class and 3 and 4 as Minority class, then it achieves good accuracy, which can be seen in Experiment 3 of Table. 2.

Experiment on Three Classes: As illustrated in Table 2, it is noticeable that a good accuracy is achieved in Exp 4 and 5. One class is taken from initial grades

Table 2: Experiments on three classes

	Classes	Model	Resolution of images	Epochs	Accuracy
Exp 3	0, 1 and 2 (Majority), 3 and 4 (Minority)	DenseNet121	224 x 224	80	99.9%
Exp 4	0, 3, and 4	DenseNet121	224 x 224	160	66%
Exp 5	1, 2, and 3	DenseNet121	224 x 224	80	63.5%
Exp 6	1, 3 and 4	DenseNet121	224 x 224	80	69%

like 0, 1, and 2, and the other class from severe classes like 3 and 4. It might be due to visible lesions in the images. When the model is trained for minority classes in Exp 6, it can be seen that the DenseNet121 model differentiates well between classes 1, 3, and 4, minority classes. An accuracy of 69% is achieved in 80 epochs.

Experiment on Five classes: In Exp 7, DenseNet121 is trained to perform multiclass classification on five classes of DR. The images are resized to a higher resolution of 512 x 512. The accuracy achieved in five classes, with all the traditional image preprocessing techniques, is 48%. The F1-score of each class shows the conflicting nature between classes 0, 1, and 2. In order to defend the ability of the model to learn the lesions, the Aptos dataset was taken to perform multiclass classification on five classes. 80% percent of data was taken from each class for training, and 20% of data was taken for testing. Images were resized to 380 x 380. Our model successfully learned the classes in experiment 8 and achieved a test accuracy of 93% on five classes. The images have good quality, and it is easy to see the small lesions and difference between those classes. Eventually, we can hypothesize that our model is good enough to learn small lesions and differentiate well between five classes. However, this dataset is relatively small, so we cannot standardize this dataset to build a generalized model for DR classification.

In experiment 9, only 700 images were taken from each class to train a Support Vector Machine (SVM). SVM is a non-parametric algorithm implemented

Table 3: Experiments on five classes

	Model	Dataset	Classes	Accuracy	F1-score
Exp 7	DenseNet121	EyePACS dataset	5	48%	Class 0 (0.31) Class 1 (0.48) Class 2 (0.29) Class 3 (0.56) Class 4 (0.68)
Exp 8	DenseNet121	Aptos dataset	5	93%	Class 0: 1.00, Class 1: 0.94, Class 2: 0.84, Class 3: 0.95, Class 4: 0.92
Exp 9	Support vector machines	EyePACS dataset	5	52.57%	Class 0: 0.35, Class 1: 0.35, Class 2: 0.36, Class 3: 0.79, Class 4: 0.79

to give the upper estimation of the model's accuracy. The F1-score of the 0, 1, and 2 classes is low, confirming the conflict between these three classes, and our highest accuracy is 52.57%. The five-class classification accuracy is higher on SVM than on neural networks. The results of these experiments can be seen in Table 3.

6 Discussion

In this section, the challenges in the Kaggle EyePACS datasets are highlighted and discussed. It has a lot of noise and wrong labeling; however, it is the most used dataset due to its large size. Different image preprocessing techniques have been used to improve noise and increase the quality of images. Data augmentation is implemented during training time to balance the classes of this dataset. Although, the accuracy did not improve as expected. Two pre-trained models, ResNet50 and DenseNet121, were chosen because of their valuable contributions in the medical field to perform multiclass classification. During the training, it was noticed that the model successfully recognized mild classes (0, 1, and 2) from severe classes (3 and 4). However, it did not perform well in differentiating the mild classes (0, 1, and 2) because of the negligible difference between those images. Moreover, class 0 is the shared class that conflicts with both class 1 and class 2, which is why the accuracy got stuck at 50% for these classes. Class 0 is the normal grade class, which holds 70% of the images from the training dataset. So, it can be considered that class 0 has a higher chance of having junk data that requires to be separated.

We have also applied a k-means clustering on class 0 to distribute it into 3 clusters. The purpose of clustering is to separate the informative images from the junk images into one cluster. Each cluster is then investigated with the rest of the classes to see if there is any one cluster that improves the accuracy of the model. After the K-means clustering, the pre-trained model DenseNet121 is used to extract features from all the images of class 0 divided into three clusters. These three clusters are considered as class 0 and then trained one by one with other classes 1, 2, 3, and 4. 10-fold cross-validation is performed to estimate the model on small training data better.

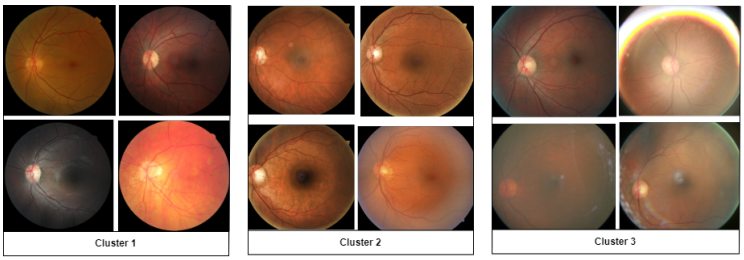

Fig. 2: Images from three different clusters.

The model's accuracy increases to 70% on five-class classification when trained on 180 epochs. Our model successfully detects the mild stages of DR, especially class 1 with the small lesion (microaneurysms) of diabetic retinopathy with an F1-score of 0.67. In addition, the detection for the severe stages of DR is also improved with a comparatively better F1-score. The accuracy on the other two clusters is relatively low, which is 42%.

In Fig. 2, we can see some random images from the three clusters 1, 2, and 3. Our model performed well on cluster 2 with 70% accuracy on five classes. It can be seen in Table 4 that the model did well in classifying the four severity classes (1, 2, 3, and 4).

7 Conclusion and Future Work

In this paper, we have done a detailed predictive analysis of the Kaggle Eye-PACS dataset. This dataset is important because it is the largest publicly available dataset with five classes. However, this dataset has many challenges like poor quality, imbalanced classes, and incorrect labeling. In our analysis, we have highlighted the drawbacks of this dataset through different experiments using transfer learning. ResNet50 and DenseNet121 were used as the deep learning models to perform five-class classification. The dataset has three conflict classes, considered to be incorrect-labeled or confused classes; normal, mild, and moderate classes with very few initial symptoms, which is why it is hard to distinguish

Table 4: K-means clustering on Class 0

	Cluster	Classes	Accuracy	Model	Epochs	F1-score
Exp 10	Cluster 1	5	42%	DenseNet121	180	Class 0: 0.01, Class 1: 0.46, Class 2: 0.22, Class 3: 0.44, Class 4: 0.62
	Cluster 2	5	70%	DenseNet121	180	Class 0: 0.13, Class 1: 0.67, Class 2: 0.73, Class 3: 0.85, Class 4: 0.88
	Cluster 3	5	42%	DenseNet121	180	Class 0: 0.07, Class 1: 0.48, Class 2: 0.31, Class 3: 0.41, Class 4: 0.57

between them. The Aptos dataset is also used to perform multiclass classification and compared to the EyePACS dataset. However, this dataset is small and insufficient to build a generalized model for DR classification. In future work, it is essential to generate new images for the stages of DR to make a new large dataset that will be good enough to be utilized in real life to help experts in diagnosing Diabetic Retinopathy.

References

1. Roglic, G., et al.: Who global report on diabetes: A summary. International Journal of Noncommunicable Diseases 1(1), 3 (2016)
2. Teo, Z.L., Tham, Y.C., Yu, M., Chee, M.L., Rim, T.H., Cheung, N., Bikbov, M.M., Wang, Y.X., Tang, Y., Lu, Y., et al.: Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. Ophthalmology 128(11), 1580–1591 (2021)
3. Cheloni, R., Gandolfi, S.A., Signorelli, C., Odone, A.: Global prevalence of diabetic retinopathy: protocol for a systematic review and meta-analysis. BMJ open 9(3), e022188 (2019)
4. Shibib, L., Al-Qaisi, M., Ahmed, A., Miras, A.D., Nott, D., Pelling, M., Greenwald, S.E., Guess, N.: Reversal and remission of t2dm–an update for practitioners. Vascular Health and Risk Management 18, 417 (2022)
5. Islam, S.M.S., Hasan, M.M., Abdullah, S.: Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images. arXiv preprint arXiv:1812.10595 (2018)
6. Gori, N., Kadakia, H., Kashid, V., Hatode, P.: Detection and analysis of microanuerysm in diabetic retinopathy using fundus image processing. vol 3, 907–911 (2017)

7. Cavan, D., Makaroff, L., da Rocha Fernandes, J., Sylvanowicz, M., Ackland, P., Conlon, J., Chaney, D., Malhi, A., Barratt, J.: The diabetic retinopathy barometer study: global perspectives on access to and experiences of diabetic retinopathy screening and treatment. Diabetes research and clinical practice 129, 16–24 (2017)
8. Hagos, M.T., Kant, S.: Transfer learning based detection of diabetic retinopathy from small dataset. arXiv preprint arXiv:1905.07203 (2019)
9. Gao, Z., Li, J., Guo, J., Chen, Y., Yi, Z., Zhong, J.: Diagnosis of diabetic retinopathy using deep neural networks. IEEE Access 7, 3360–3370 (2018)
10. Mathe, S., Pirinen, A., Sminchisescu, C.: Reinforcement learning for visual object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2894–2902 (2016)
11. Al, W.A., Yun, I.D.: Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images. IEEE transactions on medical imaging 39(4), 1245–1255 (2019)
12. Sungheetha, A., Sharma, R.: Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. Journal of Trends in Computer Science and Smart technology (TCSST) 3(02), 81–94 (2021)
13. Abràmoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M.: Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investigative ophthalmology & visual science 57(13), 5200–5206 (2016)
14. Lam, C., Yi, D., Guo, M., Lindsey, T.: Automated detection of diabetic retinopathy using deep learning. AMIA summits on translational science proceedings 2018, 147 (2018)
15. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 316(22), 2402–2410 (2016)
16. Wang, X., Lu, Y., Wang, Y., Chen, W.B.: Diabetic retinopathy stage classification using convolutional neural networks. In: 2018 IEEE International Conference on Information Reuse and Integration (IRI). pp. 465–471. IEEE (2018)
17. AATILA, M., LACHGAR, M., HRIMECH, H., KARTIT, A.: Diabetic retinopathy classification using resnet50 and vgg-16 pretrained networks. International Journal of Computer Engineering and Data Science (IJCEDS) 1(1), 1–7 (2021)
18. Salvi, R.S., Labhsetwar, S.R., Kolte, P.A., Venkatesh, V.S., Baretto, A.M.: Predictive analysis of diabetic retinopathy with transfer learning. In: 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE). pp. 1–6. IEEE (2021)
19. Sharma, S., Sharma, S., Athaiya, A.: Activation functions in neural networks. towards data science 6(12), 310–316 (2017)