International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 4, August 2021, pp. 3492~3501
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i4.pp3492-3501

A transfer learning with deep neural network approach for diabetic retinopathy classification

Mohammed Al-Smadi, Mahmoud Hammad, Qanita Bani Baker, Sa'ad A. Al-Zboon
College of Computer and Information Technology, Jordan University of Science and Technology, Irbid, Jordan

Article Info

Article history:

Received Jul 20, 2020
Revised Dec 9, 2020
Accepted Jan 13, 2021

Keywords:

Deep learning
Diabetic retinopathy
Image classification
Medical image processing
Transfer learning

ABSTRACT

Diabetic retinopathy is an eye disease caused by high blood sugar and pressure which damages the blood vessels in the eye. Diabetic retinopathy is the root cause of more than 1% of the blindness worldwide. Early detection of this disease is crucial as it prevents it from progressing to a more severe level. However, the current machine learning-based approaches for detecting the severity level of diabetic retinopathy are either, i) rely on manually extracting features which makes an approach unpractical, or ii) trained on small dataset thus cannot be generalized. In this study, we propose a transfer learning-based approach for detecting the severity level of the diabetic retinopathy with high accuracy. Our model is a deep learning model based on global average pooling (GAP) technique with various pre-trained convolutional neural net- work (CNN) models. The experimental results of our approach, in which our best model achieved 82.4% quadratic weighted kappa (QWK), corroborate the ability of our model to detect the severity level of diabetic retinopathy efficiently.

Corresponding Author:

Mohammed Al-Smadi
College of Computer and Information
Technology Jordan University of Science and
Technology Irbid, Jordan 22110
Email: masmadi@just.edu.jo

1. INTRODUCTION

Diabetic retinopathy is an eye disease caused by high blood sugar and pressure which damages the blood vessels in the back of the eye. Based on [1], around 40.3% of United States adults 40 years and older suffer from retinopathy with 8.2% have vision-threatening retinopathy. Diabetic retinopathy is the root cause of more than 1% of the blindness worldwide. People with Diabetic retinopathy are at the great risk of developing other eye diseases such as glaucoma and Cataracts. This disease is progressive meaning that it advances from one stage to a more serious stage, if it has not treated well. Early detection with effective treatment of the diabetic retinopathy can reduce vision loss by 90% [2].

To overcome the aforementioned problem and detect diabetic retinopathy early and efficiently, we have developed a deep learning model that is capable, with high accuracy, to detect if an eye suffers from diabetic retinopathy or not. If the eye suffers from diabetic retinopathy, our model detects the severity level of the disease and hence preventing the disease from progressing. Machine learning and mainly deep learning have improved drastically during the past decade [3]. Deep learning algorithms advanced many research fields such as speech recognition, decision making, and image processing.

Convolutional neural network (CNN) is a deep neural network model that is widely used in computer vision and image classifications. CNN consists of three main components: i) single or multiple convolutional blocks which is a central component of CNN, ii) sampling layers (pooling layers) such as max-

sampling and mean-sampling, and iii) number of fully connected layers. Image classification can be defined as the process of labeling images with a category from a predefined set of categories. The process of image classification consists of many phases starting from collecting a dataset of images, labeling them, pre-processing the images, image segmentation, features extraction, and finally, object classification using a deep learning model [4]. Many researchers built various deep learning architectures based on CNN such as LeNet-5 [5], AlexNet [6], ZFNet [7], VGGNet [8], GoogleNet [9], ResNet [10], Inception V2 [11], Inception V3 [12], InceptionResNet (Incption V4) [13], DenseNet [14], GapNet [15], SNet [16], Xception [17], EfficientNet [18]. These deep learning architectures are used for building various deep learning models.

In this research, we have designed and developed 6 different transfer learning techniques to detect the severity level of the diabetic retinopathy to stop blindness before it is too late. The pre-trained models are: i) ResNet [10], ii) Inception V3 [12], iii) InceptionResNet (Inception V4) [13], iv) DenseNet [14], (v) Xception [17], and vi) EfficientNet [18]. The proposed techniques were trained and evaluated on real-world medical images [19]. Each image in the training dataset is manually labeled with its severity level by a clinician.

To summarize, this paper makes the following contributions:

- Innovative transfer learning model: we have leveraged various state-of-the-art CNN architectures to build various transfer learning-based models. The CNN architectures have been used as pre-trained models to our models. CNN-based models have been used successfully in image classification tasks.
- Theory: in this paper, we show that leveraging transfer learning improves the performance of deep learning models and increases its detection accuracy.
- Experiments: we have conducted several experiments on a large medical image dataset. Our experimental results show the high ability of our model for detecting the severity level of diabetic retinopathy disease. In addition, we compared the performance of six different transfer learning-based models.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related work in the medical image processing filed. Section 3 describes our method to design and develop a deep learning model to detect the severity of a diabetic retinopathy eye. Section 4 presents our experimental results and findings and we discuss them in section 5. Finally, the paper concludes with avenues of future work on section 6.

2. RELATED WORK

Medical image classification is one of the most important images processing tasks. Its main goal is to classify medical images into different categories to help physicians and clinicians to diagnose patients faster [20]. Physicians rely on their practical experience as well as manually spotting of various features in an image to determine its medical condition. Such a process is error prone and tedious task. Therefore, the medical image classification emerged to help physicians classify medical images faster and more conveniently. To keep the paper concise and readable, we will compare and contrast between our research work in light of the related research effort in detecting diabetic retinopathy using machine and deep learning methods.

Diabetic retinopathy is one of the eye's diseases that is the root cause of blindness around the world. Detecting the severity level of diabetic retinopathy eye early is crucial for preventing possible advancement of this disease. Due to the importance of this problem, many researchers have developed various machine learning techniques for detecting diabetic retinopathy including [21-31].

Our research shares with the previous research effort the idea of detecting diabetic retinopathy but it is significantly different. For example, our dataset contains 3,562 original images whereas many previous work trained their model on a small dataset with less than 500 images such as [21-27]. Other research work trained their models on a bigger dataset such as [28] and [29] with 1,200 images and [30-32] with around 35,000 images. Nevertheless, our research work outperformed these research efforts in many ways. For example, [28] and [29] used traditional machine learning such as SVM and AdaBoost. In [30] and [31] used a single CNN model. Similarly, [32] used mainly two different models and their best obtained model achieved a kappa score of 0.72. However, in this research, we have utilized 7 different state-of-the-art deep learning models. Finally, we have developed a transfer deep learning model and our results outperformed the previous research efforts.

3. RESEARCH METHOD

3.1. The dataset

To train and evaluate our deep learning model, we have utilized the dataset available for the APTOS 2019 Blindness Detection Kaggle competition [19]. The dataset is a real-world dataset obtained from multiple clinics in India using different cameras over a period of time. The images are labeled by experts. However, they might contain some noise in both the images and the labels.

The clinics labeled the images in the dataset with the severity level of the diabetic retinopathy starting from normal eye image to proliferative diabetic retinopathy eye image. Table 1 shows the labels of the images in the dataset along with the number of images that belong to each label. The table shows that the dataset has two problems. First, the dataset is unbalanced. The images that belong to the "NO DR" class are more than half of the dataset. Therefore, if we train a classifier on this dataset, the classifier will be bias toward this class. Second, the dataset is relatively small for deep learning tasks. To solve the first problem, imbalanced data, we leveraged a data-oversampling technique. To overcome the second problem, small dataset, we used a data augmentation technique. Next section describes these two techniques in more detail. Figure 1 shows an illustrative example of each diabetic retinopathy severity level. The size of the images has been reshaped to fit the page.

Table 1. Dataset information

Severity level	Label	# Images
No diabetic retinopathy	NO DR	1805
Mild diabetic retinopathy	Mild	370
Moderate diabetic retinopathy	Moderate	999
Severe diabetic retinopathy	Severe	193
Proliferative diabetic retinopathy	Proliferative DR	195
Severity level	Label	# Images
Total images		3562

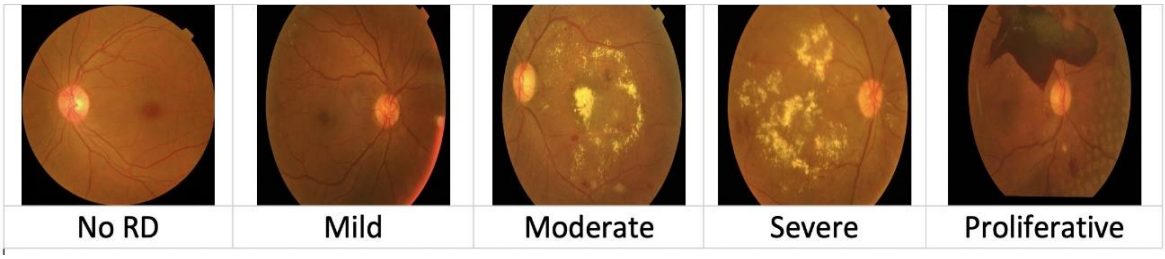

Figure 1. Images from the dataset showing eye scans with different severity levels of diabetic retinopathy

3.2. Data preprocessing

This section describes the data pre-processing techniques we leveraged to normalize the images as well as to enlarge the dataset to make it ready for deep learning tasks.

3.2.1. Image normalization

The images in our dataset are colored images with red, green, and blue channels (RGB) and their size varies from one image to another. Therefore, to standardize the size of the images, we reshaped the images to 512 x 512 pixels. We choose this size since its more efficient to run on our computers and to have enough features for the model to learn about the images.

Image normalization is a crucial step in deep learning that allows the gradient decent algorithm to converge faster and hence improving the performance of the deep learning model. There are several methods to normalize images, after converting them to integer vectors, such as: i) dividing each pixel in an image by the mean of that image vector, ii) subtract the mean per channel calculated over all images in the dataset, or iii) in picture images datasets, dividing each pixel by 255 is a simple and efficient technique. The third approach has been used to normalize the images in our dataset.

3.2.2. Data over-sampling

Over-sampling is a group of techniques to solve the imbalanced data problem. Such a technique tries to make the dataset balanced with equal number of instances in each class. The over-sampling technique that we used in this research is based on implementing a simple duplicate of random records from the minority classes. Figure 2 compares between the dataset before leveraging the over-sampling technique, Figure 2(a), and after leveraging the data over-sampling technique, Figure 2(b). As shown in the figure, after the data over- sampling, the resulted dataset is balanced. The total number of images in the dataset after the over-sampling is 7,935 images.

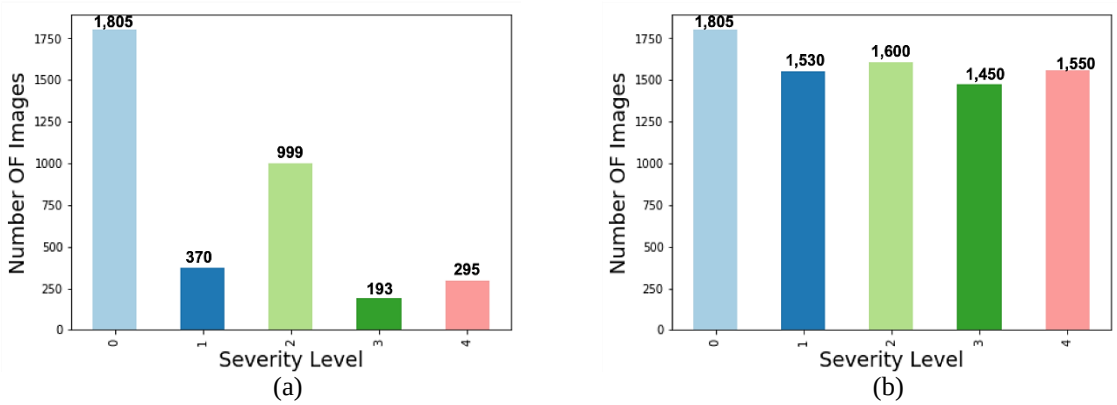

Figure 2. Severity distribution of images before and after applying the oversampling technique, (a) Original dataset, and (b) Dataset after over-sampling

3.2.3. Data augmentation

Training deep learning models such as DenseNet, ResNet, or EfficientNet, require large dataset to produce stable models. Small datasets produce models that overfit the training dataset and hence their results cannot be generalized. To avoid such a problem, we have performed a data-augmentation technique to enlarge the dataset. Data augmentation is a process of generating (manufacturing) data from the existing data to increase the diversity and the number of the instances in the dataset while maintaining the same label of the original image. Data augmentation techniques perform various operations on images including image scaling, geometric transformation, adding noise to images, changing the lighting conditions of the images, images flipping.

As depicted in Figure 3, we performed various data augmentation operations on the dataset including flip the image horizontally, flip an image vertically, scale the size of an image, rotate the image, shearing an image, and elastic and perspective transformation which tries to project an object of an image in a different point of view. To augment our images, we have utilized the ImgAug [33], a Python library for image augmentation. For each input image, we generate 64 different images other than the original one. After the data augmentation step, we ended up with 515,775 different labeled images.

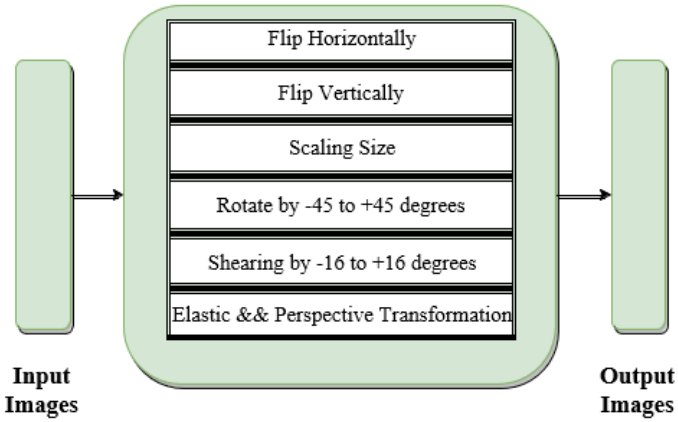

Figure 3. Image augmentation

3.2.4. Performing pseudo-label

The work of [34], the pseudo-label was implemented in this research to enhance the model performance. pseudo-label is a simple and efficient semi-supervised learning technique to improve the performance of deep neural network models. The model that uses pseudo-label is trained using supervised learning mechanism with labeled and unlabeled (test) data at the same time. For unlabeled data, the model is trained using the labeled data. Then, the trained model is used to predict the test (unlabeled) data. Finally, we

re-train the same model in a supervised mechanism using the labeled and predicted data and make a new prediction of the test data. Such a simple approach improved the performance of a state-of-the-art neural network model [34].

3.3. Transfer learning technique

Instead of training a deep learning model, mainly CNNs, from scratch, many researchers, especially in the medical images processing, leverage transfer learning technique to generate efficient models [35]. In this research, we developed a transfer-learning-based model after fine-tuning a pre-trained CNN model, trained on different images, and included the pre-trained model as an input to our model. The leveraged CNN models are pre-trained on the ImageNet dataset. The ImageNet dataset contains massive number of images and the CNN models are available to public. Such an idea greatly improved the performance of our model. Nevertheless, an ensemble model out of the best performing models is implemented.

3.4. Pre-trained models

In order to detect the severity level of diabetic retinopathy images and to compare between various pre-trained models, we have leveraged 6 state-of-the-art CNN models. The pre-trained models are: i) ResNet [10], ii) Inception V3 [12], iii) InceptionResNet (Inception V4) [13], iv) DenseNet [14], v) Xception [17], and vi) EfficientNet [18].

3.5. Global average pooling (GAP)-based classifier

Figure 4 overviews our classifier for detecting the severity level of diabetic retinopathy. As shown in the figure, we developed various deep learning models that leverage the transfer-learning technique to increase the performance of our model. Each model uses one of the CNN models, discussed in section 3.4., as a pre-trained model.

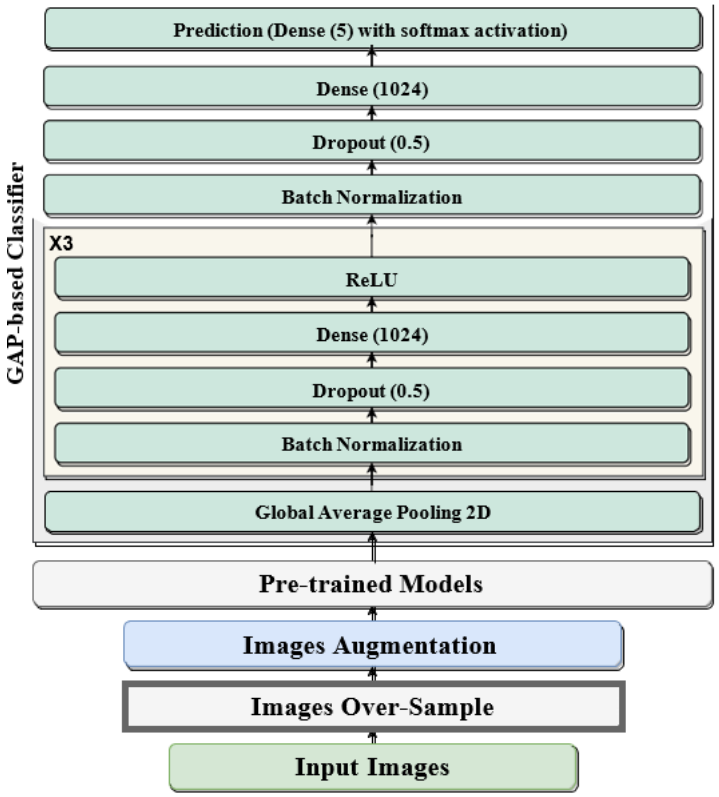

Figure 4. Our proposed classifier, a GAP-based deep learning neural network that leverages transfer learning

Our deep learning model leverages the GAP technique. GAP reduces the computational power required for the neural network to work and hence increasing its performance. GAP also is used as a regularization layer that reduces the overfitting problem and increases the generalizability of the results of the

classifier [36]. Nevertheless, GAP has shown an ability to act as an attention layer by discriminating regions of interest of the image and retain them to the final layer of the model [37]. Therefore, the GAP layer is placed after the pre-trained output layer to transfer attentive knowledge to the second part of the model. Batch normalization [11] and Dropout [38] regulation techniques are then used to reduce the overfitting problem and increase the learning capabilities of the classifier. The next layer in our deep neural network is the Dense layer, a fully connected layer with 1,024 neurons. Next, the output of the Dense layer fed to a rectified linear unit (ReLU) activation function. Those last five steps are repeated 3 times in our classifier, denoted as *X3* in Figure 4, before their output goes to the next level. The deeper the network the more vanishing the gradients will be. Therefore, a ReLU layer is added to the end of each of the three blocks to minimize the effect of vanishing gradient problem, where the gradients layer after layer are getting more and more smaller and are not back-propagated to the network layers, preventing the network from learning low level details of the images [10, 14, 39]. After repeating the previous steps three times, our classifier performs batch normalization and dropout regulation techniques, then another Dense layer with a ReLU activation function. Finally, the results of the previous layer are fed to a SoftMax function for final classification.

4. EXPERIMENTATION AND EVALUATION

This section discusses the experimentation setup and the evaluation procedure for the proposed models. The rest of this section is organized as follows: the experimentation setup and the parameters that were used to train the proposed techniques are discussed in subsection 4.1., the evaluation measure used to evaluate the proposed techniques is discussed in subsection 4.2. Finally, the models' evaluation results are presented in subsection 4.3.

4.1. Experimentation setup

The proposed techniques in this research were trained using the provided dataset (see section 3.1). All the transfer learning-based models were first trained without pseudo-label learning. Second, trained models were used to predict the label of the testing images for pseudo-label learning. Predicted labels and their associated images are then added to training dataset and used to train the proposed models.

Table 2 presents the hyper-parameters used in training the proposed models, where, *BS* stands for batch size and *LR* stands for learning rate. All the models were trained with a learning rate of 1e-4 for a maximum number of 75 epochs. All the models were trained with image size of 512x512 pixels except for the EfficientNet-B4 was trained using 380x380 pixels. All the experiments were conducted using the Kaggle kernel of the challenge. Small batch sizes were used to train the models due to the limited resources provided by the kernel and the high complexity of the used models.

Table 2. Models Hyper-parameters

Model	# Epoch	BS	LR	Input Image Size	# Channels
ResNet-50	75	6	1e-4	512*512	3
Inception-V3	75	8	1e-4	512*512	3
InceptionResNet-V2	75	6	1e-4	512*512	3
DenseNet-169	75	8	1e-4	512*512	3
Xception	75	8	1e-4	512*512	3
EfficientNet-B4	75	6	1e-4	380*380	3

4.2. Evaluation measure

The quadratic weighted kappa (QWK) [40] is used to evaluate the performance of the models. The QWK evaluates the level of agreement between the image target label and the predicted severity level. The Quadratic weighted kappa is computed using (1).

$$K = 1 - \frac{\sum_{i,j} w_{i,j} * O_{i,j}}{\sum_{i,j} w_{i,j} * E_{i,j}} \quad (1)$$

where, i represents the target label, j represents the predicted label, $O_{i,j}$ is an N*N matrix represents the received target and predicted labels, $E_{i,j}$ is an N*N matrix represents the expected target and predicted label, and $w_{i,j}$ is an N*N matrix represents a weight calculated based on the difference between the target and predicted label. $w_{i,j}$ is computed using (2):

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (2)$$

where, the *N* represents the number of testing samples, *i* represents the target label, and *j* represents the predicted label.

4.3. Results

Table 3 shows the results achieved by our transfer learning models. The best result was scored by the Inception-V3+GAP-based classifier with QWK=82.0%. The next best score was achieved by the DenseNet-169+GAP-based classifier with QWK=81.8%, whereas the third level was scored by the Xception+GAP-based classifier with QWK=80.9%. On the other hand, the ResNet-50+GAP-based classifier scored the worst results among the trained models with a QWK=77.6%.

As this research is based on a Kaggle challenge, and in order to achieve high results and rank, an ensemble based on simple average of the predictions of the top three performing models (i.e., DenseNet-169, Inception-V3, and Xception) was computed. The ensemble model outperforms the best performing model (i.e., Inception-V3) with 0.4% with a QWK of 82.4%. Although we finished the challenge with a rank of 71 (team name: Data_Science@JUST) the first team ("[ods.ai] topcoders") achieved a score of QWK=85.6% with only 3.2% of advancement over our achieved results in https://www.kaggle.com/c/aptos2019-blindness-detection/leaderboard. It is worth noting that 2,931 teams have participated in this Kaggle challenge.

Table 3. Results achieved by our proposed models

Model	Results (QWK)
ResNet-50+GAP-based classifier	77.6 %
InceptionResNet-V2+GAP-based classifier	79.6 %
EfficientNet-B4+GAP-based classifier	80.0 %
Xception+GAP-based classifier	80.9 %
DenseNet-169+GAP-based classifier	81.8 %
Inception-V3+GAP-based classifier	82.0 %
Ensemble (DenseNet-169, Inception-V3, Xception)	82.4 %

5. DISCUSSION

One of the risks for training transfer learning with deep neural networks is overfitting. Therefore, the callback function of "EarlyStopping" in https://keras.io/callbacks/ from Keras callbacks is used to stop the model training when the computed validation loss value is no more improving. As depicted in Figure 5, there is no gap between the computed train loss and the validation loss over model training epochs. This indicates that the model was trained without overfitting. Although the maximum number of epochs enabled for training was set to 75, the model stopped training after epoch 40 to prevent overfitting as shown in Figure 5. Both loss values declined together during model training due to the model layers responsible for regularization (i.e., batch normalization and dropout of Figure 4). As discussed earlier, batch normalization [11] and dropout [38] are regulation techniques used to reduce the overfitting problem and increase the learning capabilities of the classifier.

In order to show the strength of the proposed model architecture and the importance of the proposed image preprocessing techniques (see section 3), an ablation analysis was conducted on the best performing model (Inception-V3+GAP-based classifier). As presented in Table 4, relying on training the Inception-V3 model alone without transfer learning, the model scored the lowest results of QWK=63.8% with -13.7% lower than the achieved result by the transfer learning with the Inception-V3 model (i.e., Inception-V3+GAP-based classifier with QWK=82.0%). This finding shows the significance of the proposed transfer learning architecture. Moreover, this finding goes in line with findings reported in literature for the value of using transfer learning with deep neural networks in general [41] and for the medical image classification in particular [42].

The second highest impact was the ablation of pseudo-label technique, the model scored QWK 70.3% without pseudo-label with a score decline by -11.7%. This finding shows how efficient was the pseudo-label technique in improving the performance of the proposed model. The same finding was reported in the technique published in [34]. The ablation of the image augmentation scored a decline in the results with -3.4%. Previous research has demonstrated the effectiveness of using image data augmentation in enhancing models classification performance [43]. Although it was expected that image data augmentation would have a higher impact on enhancing the proposed model results, the data augmentation techniques were mainly traditional and simple (i.e. rotating, flipping, and cropping of input images).

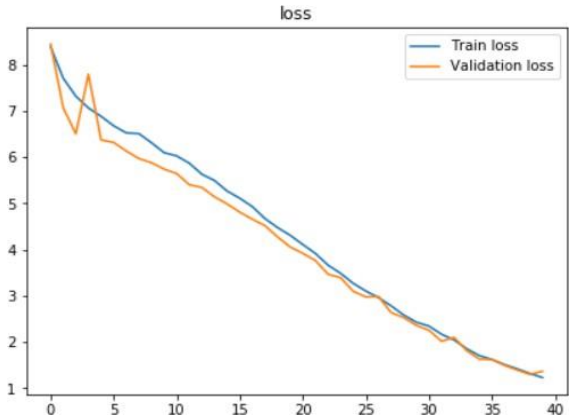

Figure 5. The loss values for the training and validation datasets during each epoch of training the model

Finally, the ablation of the oversampling technique had the lowest impact with a decline in the model score of -0.7%. As discussed earlier, a random oversampling technique was implemented to solve the imbalanced problem of the original dataset. However, the ablation analysis shows that ablating this technique has a very small affect on the model achieved results with only -0.7%. Although the random oversampling technique can be useful to avoid the negative affect of imbalanced data on achieved model results, it causes the model to overfit during training [44]. However, as we used the callback function of "EarlyStopping" from Keras callbacks to prevent model overfitting, the effect of the random oversampling was very low on the model achieved classification results.

Table 4. Ablation analysis results for the Inception-V3 + GAP-based model

Ablated Features	Result	Difference
Without Over-Sampling	81.3 %	-0.7 %
Without Augmentation	78.6 %	-3.4 %
Without Pseudo-Label	70.3 %	-11.7 %
Without Transfer Learning	68.3 %	-13.7 %
Inception-V3 + GAP-based classifier - Full	82.0 %	—

6. CONCLUSION

Diabetic retinopathy is a progressive eye disease caused by a high blood sugar or pressure. This disease, if not detected and treated well early, can cause vision loss. To that end, we have proposed a transfer learning approach for accurately detecting the severity level of diabetic retinopathy. Our model is a deep learning model based on global average pooling (GAP) technique with various pre-trained CNN models. We have utilized 6 state-of-the-art CNN models as pre-trained models to our GAP-base model and compared between them. Our best model, the Inception-V3+GAP-based classifier, achieved 82.0% QWK. Improving the performance of our models as well as applying our transfer learning models to other medical problems are interesting avenue of future directions.

ACKNOWLEDGEMENTS

This research is partially funded by Jordan University of Science and Technology, Research Grant Number: 20170107 and 20190306.

REFERENCES

[1] J. H. Kempen, B. J. O'Colmain, M. C. Leske, S. M. Haffner, R. Klein, *et al.*, "The prevalence of diabetic retinopathy among adults in the united states.," *Archives of ophthal- mology.*, vol. 122, no. 4, pp. 552-563, 2014.
[2] C. Wilkinson *et al.*, "Proposed international clinical diabetic retinopathy and dia- betic macular edema disease severity scales," *Ophthalmology*, vol. 110 , no. 9, pp. 1677–1682, 2003.
[3] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436-444, 2015.
[4] P. Kamavisdar, S. Saluja, S. Agrawal, "A survey on image classification approaches and techniques," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 1, pp. 1005-1009, 2013.

[5] LeCun, Yann *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[6] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
[7] M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks," *European conference on computer vision*, pp. 818-833, 2014.
[8] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
[10] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[11] S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, pp. 448-456, 2015.
[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for com- puter vision," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818-2826.
[13] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
[14] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
[15] S. Sabour, N. Frosst, G. E. Hinton, "Dynamic routing between capsules," *arXiv preprint arXiv:1710.09829*, pp. 3856-3866, 2017.
[16] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.
[18] M. Tan, Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2020.
[19] Kaggle, Aptos 2019 blindness detection, [Online] 2019, Avaible: https://www.kaggle.com/c/aptos2019-blindness-detection.
[20] Z. Lai, H. Deng, "Medical image classification based on deep features extracted by deep model and statistic feature fusion with multilayer perceptron," *Computational intelligence and neuroscience*, 2018.
[21] G. Gardner, D. Keating, T. H. Williamson, A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool.," *British journal of Ophthalmology*, vol. 80, no. 11, pp. 940–944, 1996.
[22] J. David, R. Krishnan, S. Kumar, "Neural network based retinal image analysis," *2008 Congress on Image and Signal Processing*, vol. 2, 2008, pp. 49-53.
[23] R. Priya, P. Aruna, "Svm and neural network based diagnosis of diabetic retinopathy," *International Journal of Computer Applications*, vol. 41, no. 1, 2012.
[24] R. Priya, P. Aruna, "Diagnosis of diabetic retinopathy using machine learning techniques," *ICTACT Journal on soft computing*, vol. 3, no. 4, pp. 563–575, 2013.
[25] S. Roychowdhury, D. D. Koozekanani, K. K. Parhi, "Dream: diabetic retinopathy analysis using machine learning," *IEEE journal of biomedical and health informatics*, vol. 18, no. 5, pp. 1717-1728, 2013.
[26] S. Somasundaram, P. Alli, "A machine learning ensemble classifier for early prediction of diabetic retinopathy," *Journal of Medical Systems*, vol. 41, no. 12, pp. 1-12, 2017.
[27] Q. Abbas, I. Fondon, A. Sarmiento, S. Jime´nez, P. Alemany, "Automatic recognition of severity level for diagnosis of diabetic retinopathy using deep visual features," *Medical & biological engineering & computing*, vol. 55, no. 11, pp. 1959-1974, 2017.
[28] R. Ka¨lvia¨inen, H. Uusitalo, "Diaretdb1 diabetic retinopathy database and evaluation protocol," *Medical Image Understanding and Analysis*, vol. 2007, p. 61, 2007.
[29] K. Bhatia, S. Arora, R. Tomar, "Diagnosis of diabetic retinopathy using machine learning classification algorithm," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 347-351.
[30] D. Doshi, A. Shenoy, D. Sidhpura, P. Gharpure, "Diabetic retinopathy detection using deep convolutional neural networks," in: *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, 2016, pp. 261-266.
[31] S. Dutta, B. C. Manideep, S. M. Basha, R. D. Caytiles, N. Iyengar, "Classification of diabetic retinopathy images by using deep learning models," *International Journal of Grid and Distributed Computing*, vol. 11, no. 1, pp. 89–106, 2018.
[32] Krishnan, Arvind Sai *et al.*, "A transfer learning approach for diabetic retinopathy classification using deep convolutional neural networks," in: *2018 15th IEEE India Council International Conference (INDICON)*, 2018, pp. 1–6.
[33] A. B. Jung *et al.*, [Online], Avaible: https://github.com/aleju/imgaug, 2019.
[34] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in: *Workshop on Challenges in Representation Learning, ICML*, vol. 3, no. 2, 2013.

[35] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285-1298, 2016.
[36] M. Lin, Q. Chen, S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, "Learning deep features for discriminative localization," in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921-2929.
[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929-1958, 2014.
[39] K. He, X. Zhang, S. Ren, J. Sun, "Identity mappings in deep residual networks," in: *European conference on computer vision*, Springer, 2016, pp. 630-645.
[40] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological bulletin*, vol. 70, no. 4, 1968.
[41] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, "A survey on deep transfer learning," in: *International Conference on Artificial Neural Networks*, Springer, 2018, pp. 270-279.
[42] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sa´nchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60-88, 2017.
[43] L. Perez, J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.
[44] N. Japkowicz, S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429-449, 2002.