A Multiclass Retinal Diseases Classification Algorithm using Deep Learning Methods

R. Behbahani Nejad[a], J. Khoramdel[b], A. Ghanbarzadeh[a], M. Sharbatdar[a], E. Najafi[a]

[a]Faculty of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran

[b]Faculty of Mechanical Engineering, Tarbiat Modares University, Tehran, Iran

Emails: {bnreza98, ghanbarzadeh.armin}@gmail.com, J.khoramdel@modares.ac.ir, {m.sharbatdar, najafi.e}@kntu.ac.ir

***Abstract*—Medical image classification plays a crucial role in monitoring and detecting diseases. This paper presents deep learning methods to distinguish images taken by the Optical Coherence Tomography technique from the normal eye and three eye-related diseases named Diabetic Macular Edema (DME), Choroidal neovascularization (CNV), and DURSEN. To achieve this aim, the images undergo a patch extraction process; then, the extracted patches are treated as sequences, and Recurrent Neural Networks are implemented to classify the images. Four pre-trained models, including VGG16, ResNet152V2, NasnetMobile, and Densnet169, and a vision transformer model are also applied and compared. Based on the results, The proposed model has achieved 99.38% test accuracy, higher than other models.**

***Index Terms*—Deep learning, Optical Coherence Tomography, Medical image classification, Recurrent Neural Networks, Vision transformer, Transfer learning**

I. Introduction

One of the global chronic diseases is diabetes which has significant side effects on the heart, eyes, nerves, etc. According to World Health Organization (WHO), approximately 422 million people have diabetes all around the world [1]. A recent study also shows that in Iran, the number of people with diabetes and prediabetes was 15.0 % and 25.4 % of the population, respectively [2]. One of the diabetes-related diseases is Diabetic Retinopathy (DR). It is an eye condition that blood vessels in the Retina, as shown in Fig. 1 are affected and can lead to vision loss and blindness in 20-65 years old people tackling diabetes [3]. One potential complication of DR is Diabetic Macular Edema (DME), which results from disruption of the blood-retinal barrier due to high blood glucose [3]. The estimations reveal that 28 million people with diabetes are suffered from DME [3]. Moreover, Age-related Macular Degeneration (AMD) is also one of the most common reasons for vision loss in elderly people. It damages the Macula, part of the Retina, shown in Fig. 1, and is responsible for human central vision [4]. Unfortunately, the expectation for the number of people with AMD is about 288 million by 2040 [5]. One common advanced technique to diagnose and detect DME and AMD is Optical Coherence Tomography (OCT). It is a non-invasive imaging technique that creates cross-sectional images of the retina with high-resolution [7]. Although OCT is a very effective method to diagnose eye affections like DME, distinguishing between eye diseases with similar signs can be challenging for ophthalmologists.

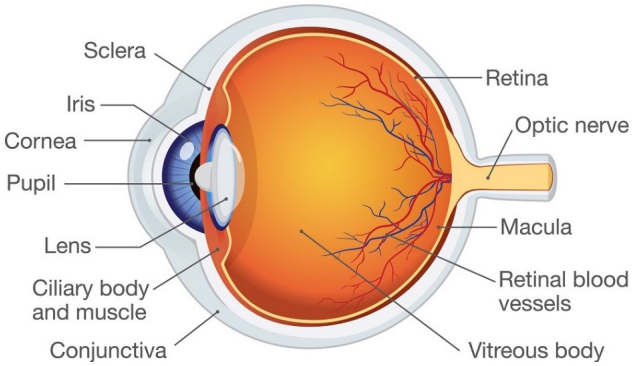


Fig. 1: Human eye anatomy: abnormalities in Retinal and Macula result in Diabetic Macular Edema and Age-related Macular Degeneration, taken from [6]

Fortunately, numerous OCT images are available all around the world, and using Artificial Intelligence methods, especially deep learning methods, can be very helpful in reducing ophthalmologists' burden. Thus OCT image classification has become popular in recent studies. In [8], The InceptionV3 architecture was used and obtained 96.6% classification accuracy. Other architectures like ResNet50 and MobileNet-v2 were also trained and evaluated for this purpose. They achieved 83.4% and 93.8 % test accuracy respectively [9]. A joint attention network has also been proposed in the literature to make the classification more robust and effective. The accuracy of their Joint-Attention-Network MobileNet-v2 was 95.60 % [9]. This paper presents four different architectures including VGG16, ResNet152v2, Densenet169 and NasnetMobile [10, 11, 12, 13] , a vision transformer model and a proposed model to classify the OCT2017 dataset. The obtained results indicate that the proposed architecture performance has the highest accuracy. The outline of this work is as follows: section II is the background. Modeling attitude is presented in section III. The results and discussion are described in section IV, and finally, the conclusion is discussed in section V.

II. Background

This section briefly discusses the main concept related to ANNs, including structures of Convolution neural networks,

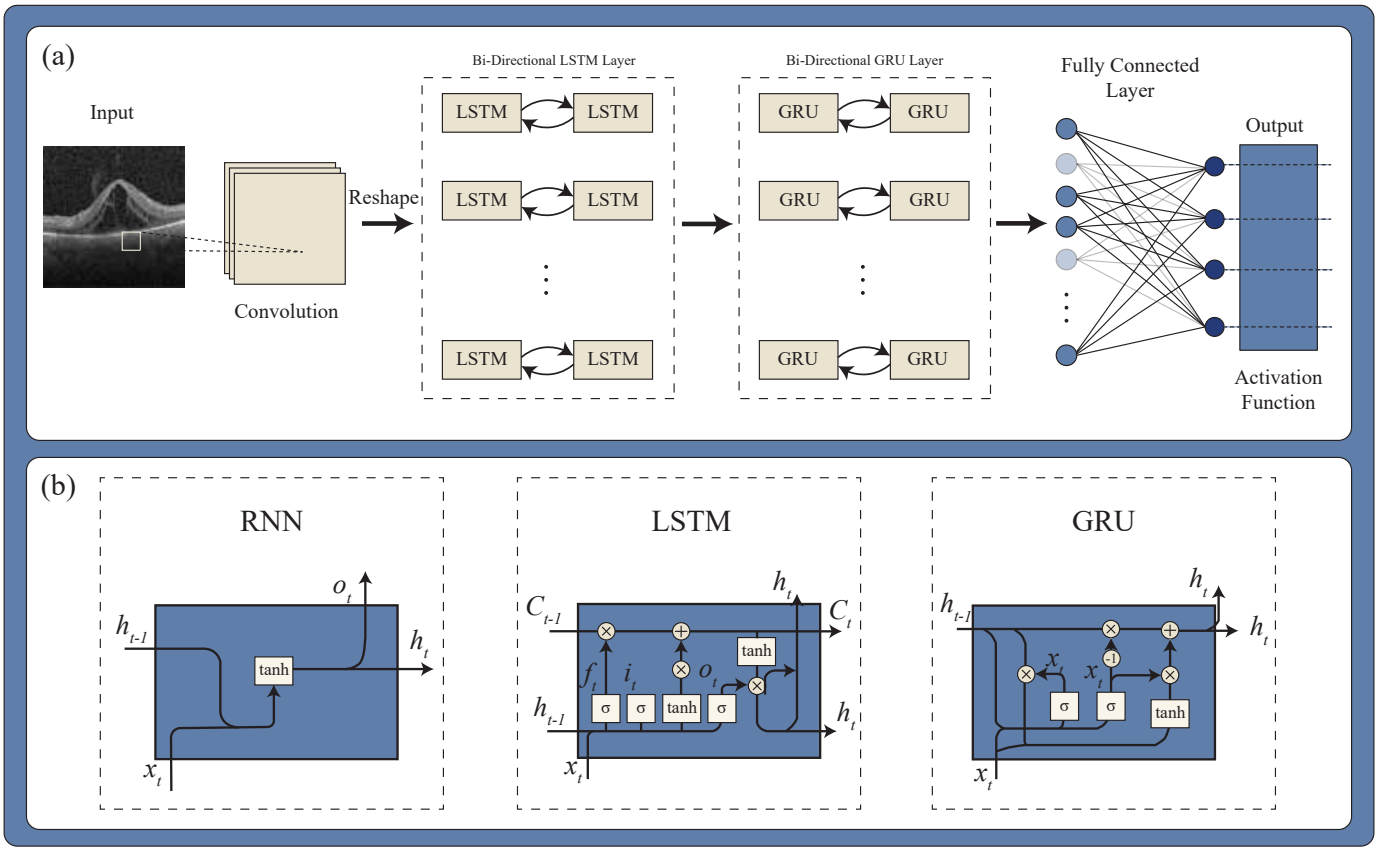


Fig. 2: a. Proposed model architecture: The image is given to a 2D convolution layer, then reshaped to sequences and fed into Bi-directional LSTM and GRU layer, respectively. The output in the next stage is given to a dense layer with dropout, and after that, another dense layer is applied to classify the input image. b. The structure of simple RNN, LSTM, and GRU : LSTM has three gates, while GRU only uses two gates

Transfer learning methods for image classification, and structures of recurrent neural networks.

A. Convolution neural networks

Convolutional Neural Networks (CNN) are widely used for image recognition tasks. Their capability to take an image as an input, break it down into perceptive features, and then classify the images makes them an interesting choice for image classification. CNN's usually consists of three layers: convolutional layers, pooling layers, and fully-connected layers [14, 15].

B. Transfer Learning

Transfer learning is a machine learning method used in image classification, image prediction, and natural language processing. Its main idea is to reuse the learnt knowledge from a problem on another task [16]. Various complex models like VGG16 trained on the Imagenet, a large-scaled dataset with over 14 million images sorted in 1000 categories, and their weights can be used for image classification of new datasets [17]. This process is shown in Fig. 3. Usually, three roadmaps describe the transfer learning process. In the first one, all network weights are frozen, the primary classifier of the network (fully connected layers) is eliminated, CNN pre-trained layers are used as feature extraction, and then a classifier layer is added. The second suggests eliminating the original classifier, using a very small learning rate to fine-tune the network, and then applying a proper classifier. The third strategy is similar to the second one, but only the top layers are fine-tuned and the bottom layers are kept frozen. This work will implement the third strategy because the first layers usually detect general features like edges, and most of the particular data-set features are perceived by the last layers [18].

C. RNN

Recurrent Neural Networks, or RNNs, are used to process sequential data. While feedforward networks cannot maintain state, RNNs can process input sequences using an internal state. Recurrent cells' feedback loops naturally take into account the temporal relationships and order of the sequences. This increases the efficiency of RNNs in sequence modeling tasks [19].

1) LSTM: Long Short-Term Memory (LSTM) network is an enhanced RNN, and Unlike RNNs that have simple layers in the network block, the LSTM block performs some

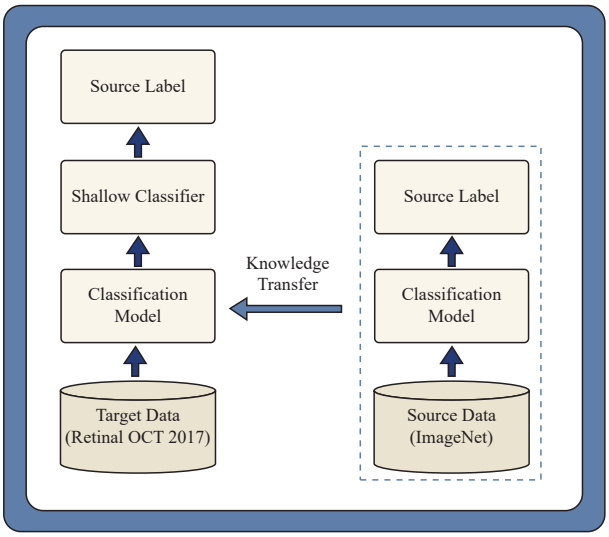


Fig. 3: Transfer learning diagram : transfer learning is a method that reuses the weights from a pre-trained model to increase network performance

additional operations. It uses input, output, and forget gates to remember important information and forget irrelevant information learned across networks [19, 20].

2) GRU: A great RNN variation is a Gated Recurrent Unit, or GRU, similar to LSTM but with only two gates named update gate and reset gate. The update gate decides the amount of prior knowledge that must be transmitted along with the next state. A reset gate governs how much of the previous memory must be retained [21].

3) Bidirectional RNNs: A bidirectional RNN or BRNN allows direct (past) and reversed (future) traversal of inputs. A BRNN consists of two RNNs, one starting at the beginning of the data sequence and moving forward and one starting at the end of the data sequence and moving backward. Simple RNNs, GRUs, or LSTMs, are all acceptable types of network blocks for BRNNs [22].
The structures of Simple RNN, LSTM, and GRU are shown in Fig. 2(b).

D. Vision Transformer

A Transformer is a deep learning model using attentional mechanisms to differentially weigh the importance of each component of the input data. The ViT is a visual representation of a transformer whose architecture was initially created for text-based operations. Similar to the series of word embeddings used when using transformers to convert text to text, the ViT model represents an input image as a set of image patches and predicts class labels for the image. The development of conventional transformer models has been influenced by three main concepts: self-attention, large-scale pre-training, and bidirectional feature encoding [23, 24].

III. Proposed Method

The steps used in this research to classify the retinal OCT2017 dataset are illustrated in this section.

A. Dataset and data visualization

The dataset composed of 84495 X-Rey OCT images in JPEG format gathered by Shiley Eye Institute of the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, the Shanghai First People's Hospital, and Beijing Tongren Eye Center between July 1, 2013 and March 1, 2017. It has four image categories (NORMAL, CNV, DME, and DRUSE) arranged into three folders for train, test, and validation [8]. There are one thousand images in the test set, and a small portion of the test set was chosen for validation. Fig. 4 shows the distribution of data for each class.

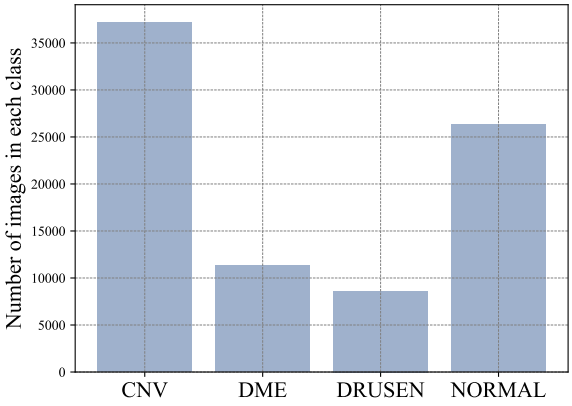


Fig. 4: Number of images in each category of retinal OCT2017 dataset

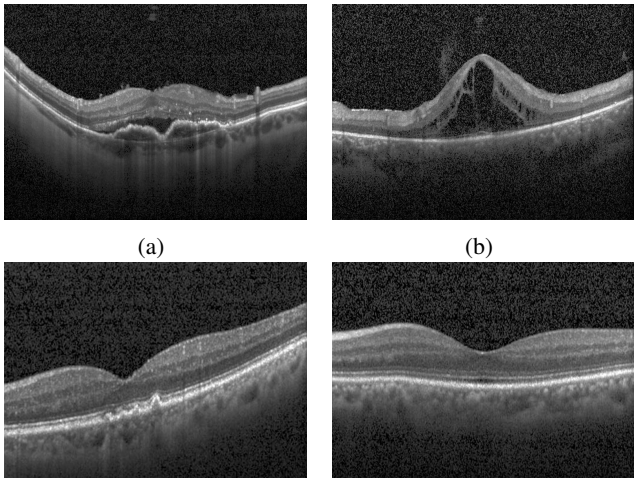
Dataset classes are demonserated as follows and are shown in Fig. 5:

- CNV: Choroidal neovascularization (CNV) is a pathologic change that abnormal blood vessels in the choroid layer of the eye are grown in patients with AMD [25].
- DRUSEN: the yellow deposits under the Retina are called DRUSEN which increases the risk of AMD [26].
- DME: Diabetic macular edema is a situation that irregular blood vessels leak fluid and cause swelling of the macula [27].

The images vary in size, so before using models to classify the classes, all images are resized to 224*224.

B. Pre-trained models

In this subsection, four pre-trained models named VGG16, ResNet152v2, NasnetMobile, and Densnet169 are trained on the Retinal OCT2017 dataset. When transfer learning methods are implemented, one notable point is choosing how to adopt pre-trained models for the new model [18]. In this work, all layers of these models are frozen except the last two layers. Extracted features are given to a four fully connected layers network to classify the inputs. From now on, we will refer to these fully connected layers as MLP. The architecture of the MLP network is shown in TABLE I. All models were trained



(a) (b) (c) (d)

Fig. 5: Visualization of Retinal OCT2017 dataset: the Retinal OCT2017 dataset contains four classes named a. NORMAL, b. DME, c. CNV, d. DRUSEN

on Google Colab's GPU K80, with the Adam optimizer and 0.0001 learning rate. The batch size was 32, and each model trained 15 epochs.

TABLE I: MLP network structure

Fully connected layers		
Layer's number	**Activation function**	**Number of neurons**
1	Relu	200
2	Relu	300
3	Relu	200
4	Sigmoid	4

C. ViT model

As mentioned in previous sections, the ViT model depicts an input picture as a series of image patches and predicts class labels for the image. A ViT model is also implemented to classify the retinal OCT2017 dataset. The parameters of ViT model is shown in TABLE II. The batch size is 256, and a callback with patience ten is also applied.

TABLE II: Parameters of vision transformer model

Parameters of vision transformer model	
Weight Decay	0.0001
Batch Size	256
Patch Size	14
Projection Dimension	256
Number of Heads	4
Transformer Head Layer	2
MLP Head Units	[512,256]

D. The proposed approach

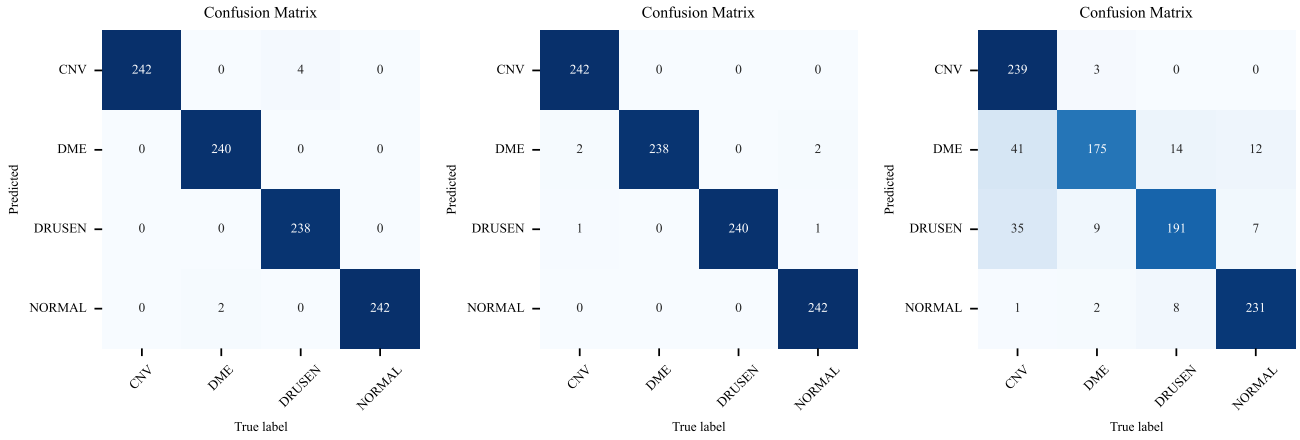
The ViTs concept inspires our proposed approach; however, similar to ViTs for extracting patches from an image, a convolution layer is implemented to produce a vector of features. These features are then fed into a two-layer bi-directional LSTM and GRU, and the output is given to a Dense layer to classify the images. The architecture of the proposed model is shown in Fig. 2(a) and TABLE III. 2D convolution layer composed of 256 filters with kernel size(14,14) ,strides(14,14) and padding='same'. The main reason for choosing a kernel size of 14 by 14 is that in the original ViT paper [24] , it is mentioned that an image worth 16x16 words. Hence, they divided the input image into 256 patches. The reason for choosing 256 filters in this layer is that the projection dim was set to 256 for the ViT model in section C. Consequently, both ViT and the proposed model receive 256 patches of size 14 by 14 and project them to get a 256-dimensional representation for each patch. Bi-directional LSTM and GRU also have 256 and 512 layers, respectively. Moreover, layer dense1 has 512 neurons, and the dropout layer also has a rate=0.25. The batch size is defined as 256, and a callback with patience ten is also applied.

TABLE III: Structure of proposed model

Proposed model architecture		
Layer	**Output Shape**	**Activation Function**
Input Layer	(None, 224, 224, 1)	-
2D Convolution Layer	(None, 16, 16, 256)	-
Reshape	(None, 256, 256)	-
Bi-directional LSTM	(None, 256, 512)	-
Bi-directional GRU	(None, 1024)	-
Dense Layer 1	(None, 512)	Leaky Relu
Dropout Layer	(None, 512)	-
Dense Layer 2	(None, 4)	Softmax

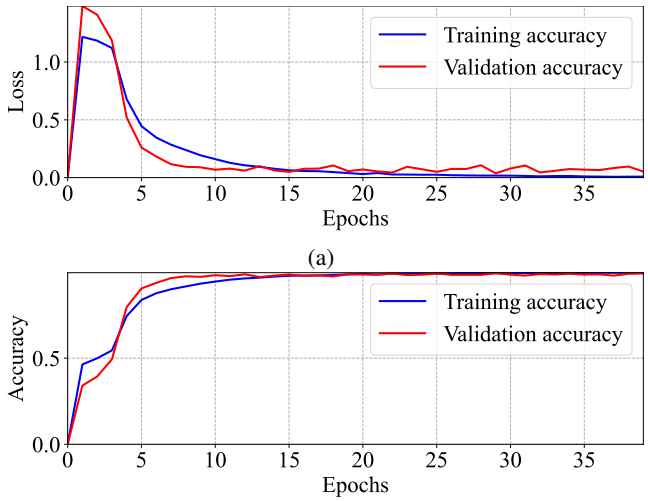
IV. Results and Discussion

This section evaluates the performance of transfer learning models, vision transformer, and our modeling approach. Based on the results, the proposed architecture achieved the highest test accuracy, 99.38% followed by VGG16 99.35 %. The lowest test accuracy belongs to the vision transformer model. The accuracy of other models is also shown in the TABLE IV. Confusion matrices for Proposed Model, VGG16, and vision transformer are shown in Fig. 6. As it can be seen, prediction for CNV and Normal class is 100% for VGG16, While the proposed model predicted DME and DRUSEN class with 100% accuracy. Even though VGG16 has had greater performance than the proposed model, none of the retinal diseases are classified as Normal conditions by the proposed model. as can be seen from Fig. 6(a), predictions for two cases with DME are predicted as Normal. Other models' performances are also satisfactory; all of them predicted CNV more accurately, even the vision transformer model. Maybe the reason why predictions for CNV are more accurate is related to data distribution. Figure. 7 shows how training and validation loss is decreased, and training accuracy is increased for the proposed model based on the number of epochs. As seen, the training Loss is dropped, and its line slope is just about to be straight, so the training is stopped to prevent the model from being overfitted. The training and validation accuracy comes



(a) (b) (c)

Fig. 6: Confusion matrices: a. VGG16, b. proposed model, and c. vision transformer model



(a)

(b)

Fig. 7: a. Training and validation loss for proposed architecture, b. Training and Validation accuracy for the proposed architecture

out to 0.9957 and 0.9917, respectively. The training loss and validation loss decrease and stabilize at a specific point, which is satisfactory. Overall all of these models except the vision transformer achieved more than 95 % accuracy, meaning the deep learning approach is efficient and has the capability to classify the OCT images automatically and accurately.

TABLE IV: Classification report table

Classification report table				
Architecture	**Precision**	**Recall**	**F1 Score**	**Train params**
VGG16	0.99	0.99	0.99	12397620
Resnet152V2	0.99	0.99	0.99	21166324
NasnetMobile	0.97	0.97	0.97	10470304
Densnet169	0.96	0.96	0.96	16468896
Proposed model	0.99	0.99	0.99	4779780
Vision transformer	0.88	0.86	0.86	1434628

V. Conclusion

Advances in Artificial intelligence have made outstanding contributions to the healthcare system. This paper investigated four deep transfer learning methods, a vision transformer model, and a proposed model to distinguish DME, CNV, DURSEN, and NORMAL eyes using retinal OCT images. This work shows that images can be considered sequences and trained by RNNs. This approach is compared to pre-trained and vision transformer models and achieves the highest performance compared to these models and human experts in diagnosing retinal OCT images.

References

[1] W. H. O. (WHO), "Diabetes," in *who.int https://www.who.int/health-topics/diabetes*, accessed 20 July 2022.

[2] M. E. Khamseh, S. G. Sepanlou, N. Hashemi-Madani, F. Joukar, A. H. Mehrparvar, E. Faramarzi, H. Okati-Aliabad, Z. Rahimi, A. Rezaianzadeh, R. Homayounfar *et al.*, "Nationwide prevalence of diabetes and prediabetes and associated risk factors among iranian adults: analysis of data from persian cohort study," *Diabetes Therapy*, vol. 12, no. 11, pp. 2921–2938, 2021.

[3] T. I. D. F. (IDF), "What is macular degeneration?" in *idf.org, https://idf.org/54-our-activities/562-diabetic-macular-edema-dme.html*, accessed 20 July 2022.

[4] K. Boyd, "Diabetes," in *aao.org. https://www.aao.org/eye-health/diseases/amd-macular-degeneration*, accessed 21 July 2022.

[5] K. L. Pennington and M. M. DeAngelis, "Epidemiology of age-related macular degeneration (amd): associations with cardiovascular disease phenotypes and lipid factors," *Eye and vision*, vol. 3, no. 1, pp. 1–20, 2016.

[6] E. R. Group, "Basic anatomy of the retina," in *elmanretina.com https://www.elmanretina.com/the-basic-anatomy-of-the-retina*, accessed 24 July 2022.

[7] J. F. Bille, "High resolution imaging in microscopy and ophthalmology: new frontiers in biomedical optics," 2019.

[8] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[9] S. A. Kamran, A. Tavakkoli, and S. L. Zuckerbrod, "Improving robustness using joint attention network for detecting retinal degeneration from optical coherence tomography images," in *Proceedings of IEEE International Conference On Image Processing (ICIP)*. IEEE, 2020, pp. 2476–2480.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[13] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[14] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[15] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, "Fundamental concepts of convolutional neural network," in *Recent trends and advances in artificial intelligence and Internet of Things*. Springer, 2020, pp. 519–567.

[16] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[17] P. U. Stanford Vision Lab, Stanford University, "Imagenet," in *image-net.org. https://www.image-net.org/about.php*, accessed 24 July 2022.

[18] I. Kandel and M. Castelli, "Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review," *Applied Sciences*, vol. 10, no. 6, p. 2021, 2020.

[19] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, pp. 650–655, 2021.

[20] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[21] Y. Liu, Z. Song, X. Xu, W. Rafique, X. Zhang, J. Shen, M. R. Khosravi, and L. Qi, "Bidirectional gru networks-based next poi category prediction for healthcare," *International Journal of Intelligent Systems*, vol. 37, no. 7, pp. 4020–4040, 2022.

[22] S. Chowdhury, X. Dong, L. Qian, X. Li, Y. Guan, J. Yang, and Q. Yu, "A multitask bi-directional rnn model for named entity recognition on chinese electronic medical records," *BMC bioinformatics*, vol. 19, no. 17, pp. 75–84, 2018.

[23] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] N. J. Y. Yeo, E. J. J. Chan, and C. Cheung, "Choroidal neovascularization: Mechanisms of endothelial dysfunction," *Frontiers in Pharmacology*, vol. 10, p. 1363, 2019.

[26] D. Porter, "What are drusen?" in *aao.org. https://www.aao.org/eye-health/diseases/what-are-drusen*, accessed 19 July 2022.

[27] D. A. K. Kuroiwa, F. K. Malerbi, and C. V. S. Regatieri, "New insights in resistant diabetic macular edema," *Ophthalmologica*, vol. 244, no. 6, pp. 485–494, 2021.