Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Deep Learning Approach for Age-related Macular Degeneration Detection Using Retinal Images: Efficacy Evaluation of Different Deep Learning Models

Ngoc Thien Le [a], Thanh Le Truong [a], Pear Ferreira Pongsachareonnont [b], Disorn Suwajanakorn [b], Apivat Mavichak [b], Rath Itthipanichpong [c,*], Widhyakorn Asdornwised [a], Surachai Chaitusaney [a], Watit Benjapolakul [a,*]

[a] *Center of Excellence in Artificial Intelligence, Machine Learning, and Smart Grid Technology, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, 10330, Thailand*

[b] *Center of Excellence in Retina, Faculty of Medicine, Chulalongkorn University, Bangkok, 10330, Thailand*

[c] *Low vision unit and Glaucoma research unit, Department of ophthalmology, Faculty of Medicine, Chulalongkorn university and King Chulalongkorn Memorial Hospital, Bangkok, 10330, Thailand*

ARTICLE INFO

Keywords:
Retinal image
Age-related macular degeneration (AMD)
Dry AMD
Wet AMD
Deep learning (DL) model

ABSTRACT

Age-related macular degeneration (AMD) is a typical fundus disease that affects the central vision of elderly people. It causes difficulties in everyday activities such as reading and recognizing faces. AMD can progress slowly or rapidly, and it leads to severe vision loss if left untreated. Therefore, early detection and diagnosis of AMD are crucial to prevent or delay vision impairment in the elderly. To handle this requirement, researchers are exploring deep learning-based models as an AI tool to assist ophthalmologist in AMD diagnosis. However, conducting an appropriate deep learning model for the AMD classification is challenging and cost-intensive. This research aims to evaluate the efficacy of various deep learning models for obtaining the best performance results when identifying AMD disease using retinal images. To meet this objective, the retinal images from the Department of Ophthalmology, the King Chulalongkorn Memorial Hospital, Thailand were collected for transfer learning and other publicly available datasets for testing. Then, seven deep learning models VGG19, Xception, DenseNet201, EfficientNetB7, InceptionV3, NASNetLarge, and ResNet152V2 were chosen to training for the 2-labels (Normal vs. AMD) and the 3-labels (Normal vs. Dry AMD vs. Wet AMD) classifications. From the experimental results, the DenseNet201 model with Dense block in its structure showed the best efficacy in both 2-labels and 3-labels AMD classifications since its performance always include in the Top-3 models accuracy and generalization performance measured by total accuracy and total F1-Score, respectively. Furthermore, the accuracy performance of deep learning models in Top-3 are comparable with the performance of retinal specialist. These results contribute consolidated knowledge to the process of implementation effective deep learning as production that detects AMD automatically in the clinical and enhance the quality of healthcare service.

1. Introduction

Age-related macular degeneration (AMD) treatment is one of the critical missions of the eye care programs in many developing countries. In Thailand, the victim of AMD in the Thai population aged above 50 are 12.2% . The population-based research in [1] also concludes that Thailand has a large number of late AMD patients. To diagnose and treat AMD effectively, many researchers have proposed deep learning (DL) models that can detect and classify AMD using retinal images, which are images of the retinal area [2–4]. DL provides a powerful method for analyzing and extracting information from complex and high-dimensional data in retinal images [5–9].

* Corresponding authors.
E-mail addresses: Thien.L@chula.ac.th (N.T. Le), letruongthanh3698@gmail.com (T. Le Truong), pear.p@chula.ac.th (P.F. Pongsachareonnont), disorn.s@chulahospital.org (D. Suwajanakorn), Apivat.M@chula.ac.th (A. Mavichak), rath.i@chula.ac.th (R. Itthipanichpong), Widhyakorn.A@chula.ac.th (W. Asdornwised), Surachai.C@chula.ac.th (S. Chaitusaney), watit.b@chula.ac.th (W. Benjapolakul).

https://doi.org/10.1016/j.eij.2023.100402
Received 19 June 2023; Received in revised form 15 August 2023; Accepted 5 September 2023
Available online 4 October 2023

Convolutional layer is the basis component in deep learning architecture. The features of retinal image are extracted by a set of filters in convolutional layer. A filter represents as a weights matrix that slides over the input and computes a dot product, producing an output value. This creates a feature map that shows where the feature is detected in the input. A convolutional layer also uses multiple filters to detect different features and change the size or resolution of the feature maps with padding, stride, dilation, or pooling. From the basis of convolutional layers, there are many blocks have been developed for different architectures of DL model as below.

- *Convolution block*: main component of most DL models which named as convolutional neural network (CNN). Convolution block consists of multiple layers that perform convolution, pooling, activation, and other operations on the input data. Convolution block is popular in image classification, object detection, and other computer vision tasks.
- *Inception block*: a type of CNN layer that consists of multiple parallel branches with different convolutional filters. The idea is to capture features at various scales and various levels of abstraction. Inception blocks increases the DL network's depth and width without increasing the number of parameters too much. Inception blocks are used in deep learning models, such as GoogLeNet and Inception [10].
- *Residual block*: a type of CNN layer that introduces input-output skip connections of a subnetwork. The idea is to allow the network to learn residual functions that can cope with the gradient vanishing problem and improve feature learning [11]. Residual blocks can increase the network's depth without degrading its performance. Residual blocks are used in deep learning models, such as ResNet [12,13].
- *Reduction block*: a type of CNN layer that reduces the spatial dimensions of the feature maps by using pooling or strided convolution. The idea is to decrease the computational cost and memory usage of the network while preserving the essential information. Reduction blocks can also increase the receptive field of the network and improve its generalization ability. Reduction blocks are often used before or after inception or residual blocks [14].
- *Dense block*: a type of CNN layer that links previous layer to all subsequent layers in a feed-forward structure. The idea is to reuse features and enhance information flow within the network. Dense blocks can increase the network's width and feature diversity without increasing the number of parameters too much. Dense blocks are used in the DenseNet model [15].

From the development of various DL models for AMD classification using fundus image, there are many studies utilized them to achieve well-performance results [16–27]. However, the underlying reason for choosing a particular DL model has not been understood clearly. No previous studies have focused on the question of which DL architecture should be recommended for the task of AMD classification using fundus image. Furthermore, the reason why authors selected DL models in above mentioned literature is somehow arbitrary. Another drawback is the results performance of proposed DL models are only evaluated on the same dataset which was used for the training models. Therefore, there is no guarantee that the trained DL models also perform well on new image dataset that is never used to train the model before. Solving these issues will help researchers save time and resources for selecting appropriated DL model as an effective tool to assist ophthalmologist in AMD diagnosis.

The aim of this study is to figure out the appropriate DL architectures to classify the retinal images into 2-labels (Normal vs. AMD) and 3-labels AMD (Normal vs. Wet AMD vs. Dry AMD) classification in term of the accuracy and the generalization performances. Our hypothesis is that the method of transferring extracted features from the input layer to the output layer in each type of deep learning block effects to the performance of image classification. Therefore, we select seven DL models which are implemented from different block types, named as VGG19 [28], Xception [29], DenseNet201 [15], EfficientNetB7 [30], InceptionV3 [31], NASNetLarge [32], and ResNet152V2 [12] in our study [33,34]. The training process of those DL models is based on the transfer learning technique using our retinal images of Chula-AMD dataset and an independent dataset for evaluating the generalization ability of trained DL models. Finally, our contribution is to provide a useful guideline for researchers about the most appropriate DL model and the Top-3 DL models, in termed of the accuracy and generalization, for AMD classification using retinal image.

The study is organized as follows. The review of using DL models for AMD detection is shown in Section 2. The material and research methodology in our study are given in Section 3. Section 4 shows the result experiments and discussion of trained DL models on 2-labels and 3-labels AMD classifications. Finally, Section 5 concludes our study.

2. Literature review

Most of the recent deep learning approaches are focused on the AMD binary classification problem, for examples AMD vs. non-AMD, early-AMD vs. advanced-AMD. Authors in [23] proposed a 13-layer CNN architecture to classify retinal image as non-AM or AMD with the accuracy from 90% to 99.5%. In [16], authors used a modified VGG16 architecture to classify retinal images from the Age-Related Eye Disease Study (AREDS) dataset [35] and achieved an accuracy value of 91.6% and a sensitivity value of 88.4%. Authors [17] used ten-fold cross validation technique to train a 14-layer CNN for detecting Normal or AMD images from the Department of Ophthalmology, Kasturba Medical College, India (KMC dataset) and achieved an accuracy of 95.45% and a sensitivity of 96.43%. However, their dataset was imbalanced as there were more AMD images than normal images. In [18], authors used fundus images recorded with a viewing angle of 200° to implement a DL model to classify an image as Normal or Wet AMD and achieved about 100% on both accuracy and sensitivity values.

Another challenge in AMD classification is to use multiple AMD labels that reflect the different stages and types of AMD disease. Various studies have been suggested to solve the 3-labels (e.g.: Normal plus Early vs. Intermediate vs. Advanced) and 4-labels (e.g.: Normal vs. Early vs. Intermediate vs. Advanced) classification tasks. Based on the 4-step AMD severity scale of the AREDS dataset, the proposed customized VGG model [19] reached an accuracy of 83.2%. Another proposed self-supervised neuron network [26] obtained a lower accuracy of 65%. In addition, other approaches of deep random forest [21] and support vector machine [20] obtained the accuracy results about 79.4% and 78.34%, respectively. For the 3-step AMD scale of AREDS images, the accuracy performance of support vector machine [20] was 81.5%. For different 3-labels of Normal, AMD, and Tessellated using images from Health Management Center, Shenzhen University General Hospital Shenzhen University, authors [25] proposed InceptionV3 and ResNet50 model with the accuracy of 91.8% and 93.8%, respectively. The recent study applied the particle swarm optimization method [36] solved the 3-labels AMD with Normal, Dry AMD, and Wet AMD labels using KMC retinal images. The obtained accuracy and sensitivity are 85.3% and 82.65%, respectively. However, the KMC dataset had a limited number of Wet AMD images, which may affect the reliability of their results. These results in literature infers that there is a necessary requirement for improving the accuracy and sensitivity of multi-labels AMD classification [37,38].

3. Material and methodology

3.1. Dry and Wet AMD

Based on the clinical classifications in [39,40], the AMD disease includes the Dry and Wet types, which are illustrated as in Fig. 1. Dry

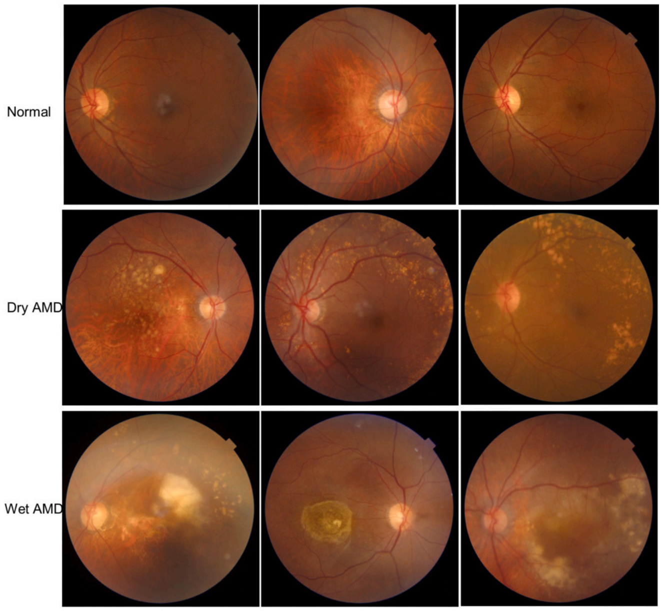


Fig. 1. Retinal images in Chula-AMD dataset: Normal retinal (top row), Dry AMD disease (middle row) and Wet AMD disease (bottom row).

AMD is a slow deterioration of the retina with the drusen syndrome under the macular. This causes the macular to thin and dry out, losing its function and affecting central vision. Minimal vision loss with the changes of large drusen and pigment in the macular are the main symptom in the early stage of Dry AMD. Drusen are debris under the retinal pigment epithelium [41]. Most people over 50 have small drusen, but only large drusen increase the risk of late AMD [42,43].

Wet AMD disease is a rare and severe form of late AMD that damages the macular quickly. It occurs when abnormal blood vessels develop under the retina and macular, and leak fluid or blood. This makes the macular swell or detach, affecting central vision. Early treatment can improve the outcome of Wet AMD.

3.2. Image dataset

The color retinal images using to process the transfer learning for DL models is from Chula-AMD dataset, which was gathered from the Department of ophthalmology, the King Chulalongkorn Memorial Hospital in Bangkok. The retinal images are captured using digital fundus camera in the DICOM format. Then, these retinal images are converted to JPEG to erase the health information of patient. Examples images of our Chula-AMD dataset can be seen in Fig. 1. In our study, the Chula-AMD images is randomly divided into 80% – 20% for training and validating, respectively. Images in testing step are collected from publicly available datasets Retinal Image Bank [44] and iChallenge-AMD [45]. Remark that the images in testing dataset are not used during our training period. The goal of using an independent testing dataset in our experiments is that to test the generalization of DL model, which means the ability to adapt properly to the unseen retinal images of DL models. Moreover, our Chula-AMD dataset has a balanced number of retinal images for all three categories, which minimizes any biased results during the transfer learning. The summarized total number of images in our experiments are shown in Table 1.

3.3. Experiment implementation

The experiment process in our experiments is illustrated in Fig. 2. We utilize the Colab cloud platform, which supports Keras and Python environments, to train DL models. The retinal images are scaled to 224 × 224 pixels and normalized to [0, 1] range before imported into the DL models, as shown in Table 2. We use the ImageDataGenerator function in Keras to prepare the images for the transfer learning process. We import the seven DL models into the Colab platform and fine-tune them with the Chula-AMD dataset. The optimized algorithm is rectified Adam cross-entropy which is common for the image classification task. The loss function for training is based on the categorical cross entropy method.

Table 1
Training, validating, and testing datasets in our study. The Chula-AMD is randomly divided into 80% for training step and 20% for validating step.

Types	Labels	Chula-AMD		Retinal bank and i-Challenge
		Training	Validating	Testing
2-labels AMD	Normal	972	243	100
	AMD	780	195	100
3-labels AMD	Normal	411	103	50
	Dry AMD	379	100	50
	Wet AMD	396	100	50

The architectures of these DL models are presented in Table 2. We select these DL models because they are the latest versions of their respective DL architectures and have achieved the best accuracy results on the ImageNet dataset [47], which is a standard benchmark for image classification with 1,000 different object categories. The number of epochs is 30. We use the accuracy and loss measurements to monitor during the training process and to stop the training DL models. The early-stopping during training is allowed. The accuracy of the trained DL model on the validation dataset is the main criterion for stopping the transfer learning process. If the validating accuracy achieves a maximum value, the transfer learning will be terminated. We also collect the values of precision and recall to measure the accuracy and F1-Score performance of seven DL models in our experiments.

3.3.1. Transfer learning technique

This is a technique where a pre-trained DL model for one task is used for a different but related task. This technique is very common in computer vision [48]. It helps to achieve high performance in the case of small dataset as in Chula-AMD, by training a pre-trained model on similar images. It also saves time and computing resources than implementing a DL model from scratch.

As shown in Fig. 2, the transfer learning starts with the import of seven pre-trained DL models into the Colab Pro. Then, the existing output layer is replaced by the corresponding output layers of 2-labels and 3-labels AMD classification. Indeed, only the output layer of pre-trained model is trainable during the training. Finally, the images are feed into the input layer to process the transfer learning.

3.3.2. Model evaluation

We utilize the confusion matrix table to show how well a deep learning model can classify images after a transfer learning process. It shows the number of correct and incorrect predictions made by the DL model for each category. The meaning of four terms in a confusion matrix are given as below:

- True positive (TP): A experiment result that correctly detects the presence of AMD disease in 2-labels task and Dry or Wet AMD in 3-labels task.
- True negative (TN): A experiment result that correctly indicates the absence of AMD disease in 2-labels task and Dry or Wet AMD in 3-labels task.
- False Positive (FP): A experiment result that wrongly indicates the presence of AMD disease. For example, the model predicts that an image is AMD but it is actually Normal.

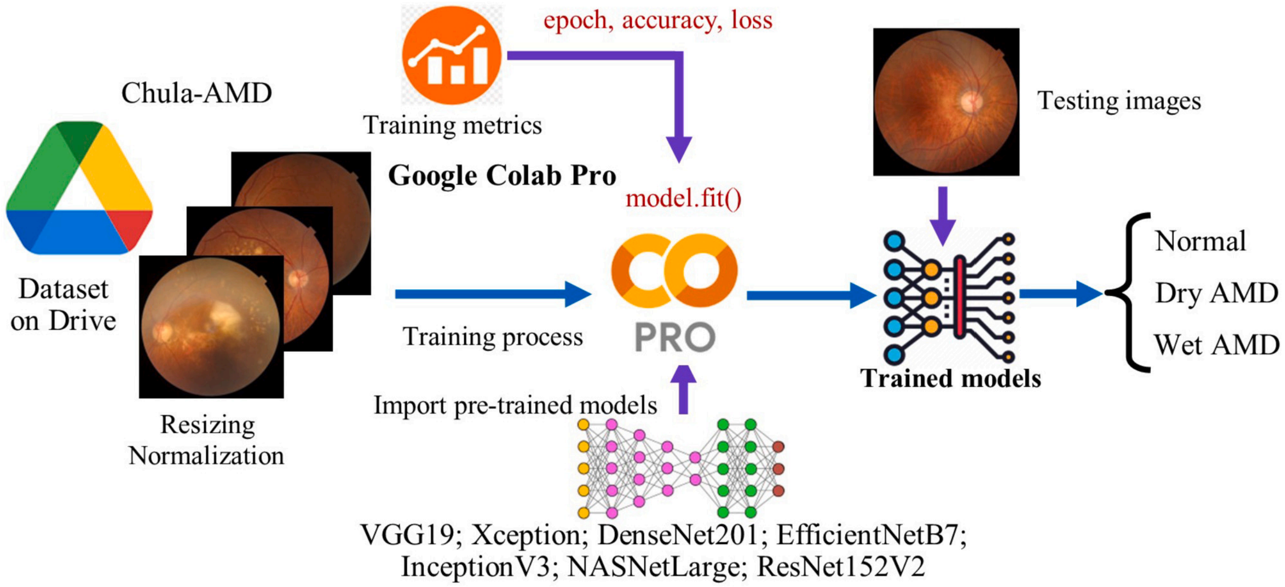


Fig. 2. Flowchart of applying transfer learning technique for the pre-trained DL models to learn the AMD classification using fundus image.

Table 2
Seven pre-trained deep learning networks used in our experiments using transfer learning technique [46]. To eliminate the effect of image resolution to model performance, we keep the same image size for all DL models.

Deep learning Model	Main block in architecture	Parameters (Millions)	Input image (pixels)
VGG19	Convolution block	20	224 x 224
Xception	Depthwise convolution block	23	224 x 224
DenseNet201	Dense block	18.6	224 x 224
EfficientNetB7	Mobile Inverted Bottleneck convolution block	66.7	224 x 224
InceptionV3	Inception block; Reduction block	21.9	224 x 224
NASNetLarg	Neural Architecture Search block	90	224 x 224
ResNet152V2	Residual block	58.6	224 x 224

- False negative (FN): A experiment result that wrongly indicates the absence of AMD disease. For example, the model predicts that an image is Normal but it is actually AMD (Dry or Wet).

From the confusion matrix, the precision value is defined as Eq. (1):

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{1}$$

The recall (or sensitivity) score is calculated as Eq. (2):

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

The accuracy score is defined as Eq. (3):

$$\text{Accuracy}_i = \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \tag{3}$$

The F1-Score metric is defined as Eq. (4):

$$\text{F1-Score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{4}$$

where i is Normal or Dry AMD or Wet AMD labels for the classification task.

The metrics from Eq. (1) to Eq. (4) are used to measure the trained DL model performance in our experiments. Indeed, both Accuracy and F1-Score are the measures of DL model accuracy. However, F1-Score also evaluates the harmonic mean of Precision and Recall values. Hence, a high F1-Score is corresponded to the high values of both Precision and Recall.

Unlike recent studies in literature which only focused on the accuracy aspect of model, our experiments also take into account the generalization performance, which measures how well the trained model adapts to the unseen dataset (not used for training model). We will evaluate the DL models across three datasets (training and validating dataset belonged to Chula-AMD and testing dataset from Retinal Image Bank and iChallenge-AMD) to obtain the generalization performance. Therefore, we define the formulas of the total accuracy in Eq. (5) and total F1-Score value in Eq. (6) to measure the Top-3 DL models performance easily.

$$Acc_{\text{total}} = Acc_{\text{train}} + Acc_{\text{val}} + Acc_{\text{test}} \tag{5}$$

$$\text{F1-Score}_{\text{total}} = \text{F1-Score}_{\text{Normal}} + \text{F1-Score}_{\text{DryAMD}} + \\ + \text{F1-Score}_{\text{WetAMD}} \tag{6}$$

The benefit of using Acc_{total} and F1-Score$_{\text{total}}$ is that they consider the performance of trained DL model on both seen and unseen image datasets.

4. Result and discussion

4.1. 2-labels AMD classification

4.1.1. Trained model results

The experiment results for 2-labels AMD classification is given in Table 3 on the validating dataset. The highest accuracy of 96% is obtained from the trained DenseNet201 model. Furthermore, other trained models also achieved well performance with accuracy larger than 90%. Meanwhile, the lowest accuracy of 83% from trained VGG19 model indicates that this DL architecture is not suitable for the AMD classification using fundus image. The accuracy of trained VGG19 model was even worse than the performance of trained VGG16 model in [16] (83%, compared to 91.6%). However, the low accuracy performance of trained

Table 3
The 2-labels AMD classification results: precision, recall, F1-Score, and accuracy on validating dataset.

Model	Label	Precision	Recall	F1-Score	Accuracy (%)
VGG19	Normal	0.91	0.77	0.84	0.83 (83%)
	AMD	0.76	0.91	0.83	
Xception	Normal	0.93	0.97	0.95	0.94 (94%)
	AMD	0.96	0.91	0.93	
DenseNet201	Normal	0.97	0.95	0.96	0.96 (96%)
	AMD	0.94	0.96	0.95	
EfficientNetB7	Normal	0.94	0.97	0.95	0.95 (95%)
	AMD	0.96	0.92	0.94	
InceptionV3	Normal	0.94	0.96	0.95	0.94 (94%)
	AMD	0.95	0.92	0.93	
NASNetLarge	Normal	0.95	0.97	0.96	0.95 (95%)
	AMD	0.96	0.93	0.95	
ResNet152V2	Normal	0.92	0.96	0.94	0.93 (93%)
	AMD	0.95	0.90	0.92	

Table 4
The 2-labels AMD classification results: precision, recall, F1-Score, and accuracy on testing dataset.

Model	Label	Precision	Recall	F1-Score	Accuracy (%)
VGG19	Normal	0.64	0.76	0.70	0.67 (67%)
	AMD	0.71	0.58	0.64	
Xception	Normal	0.81	0.84	0.82	0.82 (82%)
	AMD	0.83	0.80	0.82	
DenseNet201	Normal	0.74	0.92	0.82	0.80 (80%)
	AMD	0.89	0.67	0.77	
EfficientNetB7	Normal	0.78	0.93	0.85	0.83 (83%)
	AMD	0.91	0.74	0.82	
InceptionV3	Normal	0.75	0.76	0.76	0.76 (76%)
	AMD	0.76	0.75	0.75	
NASNetLarge	Normal	0.72	0.77	0.74	0.73 (73%)
	AMD	0.75	0.70	0.73	
ResNet152V2	Normal	0.78	0.62	0.69	0.73 (73%)
	AMD	0.69	0.83	0.75	

VGG19 model could be interpreted as this is one of the baseline models in deep learning which are used to develop next generation model such as Xception, DenseNet, or Inception, which obtained higher accuracy performance on the ImageNet dataset [47].

Table 4 gives the experiment results of the same trained DL models in Table 3 on testing dataset. It was not surprise that all trained DL models performed lower resulting accuracy on the testing images. In detail, we obtained the well accuracy performance from the EfficientNetB7, Xception, and DenseNet201 with the corresponding accuracy of 83%, 82%, and 80% respectively. Those accuracy values outperformed than the remains. Finally, Fig. 3 visualizes the accuracy results of trained DL models to highlight the variation of accuracy performance across three datasets.

4.1.2. *Top-3 performance*

In term of total accuracy (Acc_{total}) and total F1-Score (F1-Score$_{\text{total}}$), we obtained the Top-3 DL models performance as in Table 5. The trained DenseNet201 model always included the Top-3 in three cases (Total accuracy, F1-Sore on validating, and F1-Score on testing). Therefore, we concluded that the DenseNet201 model obtained best performance for the 2-labels AMD classification in term of accuracy and generalization performance. Another well-performance models are Xception and EfficientNetB7 with the accuracy results are comparable with the level of trained retinal physicians (about 95.2% of accuracy) [37,38]. Finally, the confusion matrix of DL models in Top-3 are given as in Fig. A.1 and Fig. A.2 to show the detail numbers of TP, TN, FP, and FN on validating and testing datasets, respectively.

Table 5
The 2-labels AMD classification results: Top-3 performance of total accuracy and total F1-Score.

Total accuracy (Acc_{total})	Total F1-Score (F1-Score$_{\text{total}}$)	
	Validating	Testing
Xception (2.75)	**DenseNet201** (1.91)	EfficientNetB7 (1.67)
DenseNet201 (2.73)	NASNetLarge (1.91)	Xception (1.64)
InceptionV3 (2.69)	EfficientNetB7 (1.89)	**DenseNet201** (1.59)

4.2. *3-labels AMD classification*

4.2.1. *Trained model results*

Compared to the 2-labels AMD classification, the complexity of 3-labels AMD classification was increased because the DL models have to distinguish small abnormal areas in the retinal, such as drusen and geographic atrophy in Dry AMD or the choroidal neovascularization in Wet AMD. Therefore, the accuracy performance of trained DL models on the validating dataset in Table 6 were lower than those in Table 3. The highest accuracy of 90% was obtained form the trained Xception model. Furthermore, DenseNet201, InceptionV3, and ResNet152V2 also achieved well performance with accuracy larger than 85%. Meanwhile,

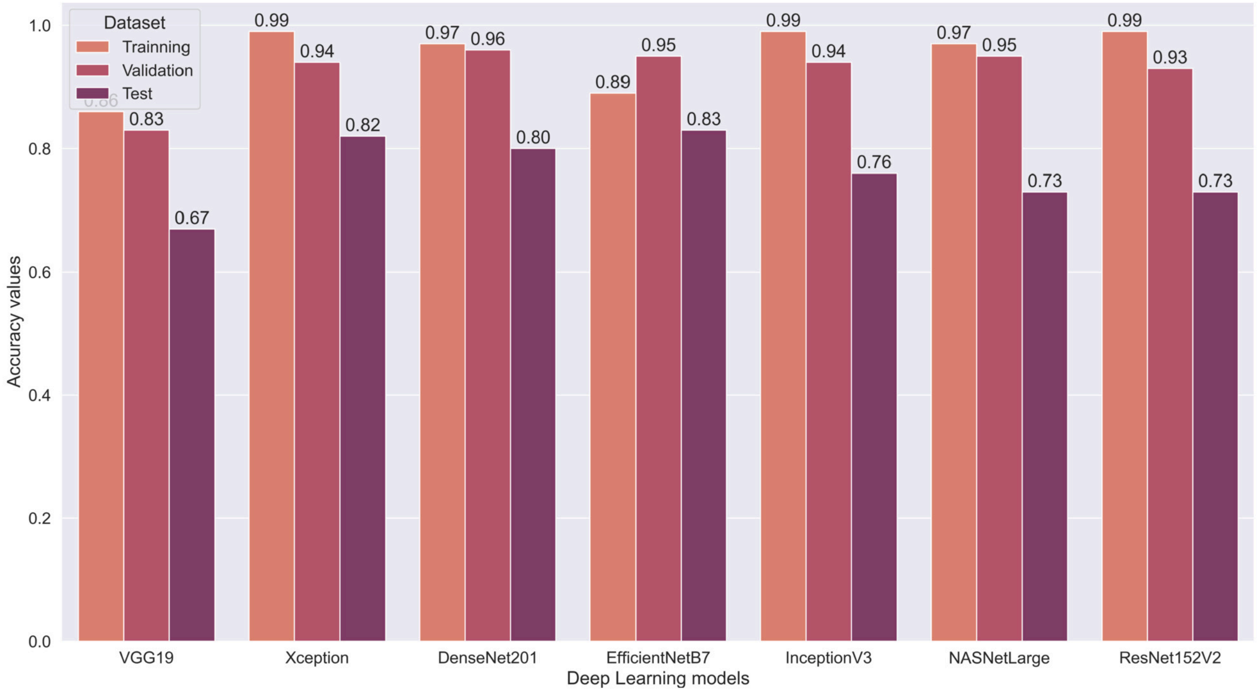


Fig. 3. Accuracy results for 2-labels AMD classification performed on three datasets.

it was not surprise that the VGG19 model also got the lowest accuracy of 70%. This result confirmed again that VGG19 model is not suitable for the 3-labels AMD classification using fundus image.

Table 7 gives the results for 3-labels AMD classification of trained DL models on testing dataset. The accuracy of trained DenseNet201, EfficientNetB7, InceptionV3, and ResNet152V2 models were 67%, 64%, 62%, and 62% which outperformed than others. In addition, the accuracy performance of Xception model reduced significantly on the testing dataset (57%, in Table 7), compared to the highest accuracy performance on validating dataset (90%, in Table 6). Therefore, we concluded that the Xception model did not perform well for the generalization performance. Meanwhile, the accuracy performance of DenseNet201 model were always in the top of three highest scores in Table 6 and Table 7. Therefore, the DenseNet201 model shown the well performance for generalization ability. Finally, Fig. 4 illustrates the accuracy results of trained DL models to highlight the variation of accuracy performance across three datasets.

4.2.2. Top-3 performance

In term of total accuracy (Acc_{total}) and total F1-Score (F1-Score$_{\text{total}}$), we obtained the Top-3 DL models performance as in Table 8. The trained DenseNet201 model always appeared the Top-3 in three metrics (Total accuracy, F1-Sore on validating, and F1-Score on testing). Therefore, we concluded that the DenseNet201 model also obtained best performance for the 3-labels AMD classification in term of accuracy and generalization performance. Another well-performance models are InceptionV3 and ResNet152V2. Finally, the confusion matrix of DL models in Top-3 are given as in Fig. B.1 and Fig. B.2 to show the detail numbers of TP, TN, FP, and FN on validating and testing datasets, respectively.

4.3. Discussion

For the 2-AMD classification task, our trained DL models in Top-3 performance (Table 3 and Table 5) achieved comparative results compared to the modified VGG16 model in [16] and 14-layer CNN model in [17] in term of accuracy on the validating dataset. Furthermore, we also obtained the generalization performance of trained DL models on testing dataset but cannot compare to others in literature.

For the 3-AMD classification task, our trained DL models in Top-3 performance (Table 6 and Table 8) performed better than the most recent in literature [20,36] on the validating dataset. In addition, these results are also comparable with the level of trained retinal physicians (about 85% of accuracy) [37,38]. On the testing dataset, DenseNet201 model obtained highest generalization performance in term of accuracy (67%) and total F1-Score (1.99).

From above results, we conclude that the DenseNet model with Dense block in its architecture achieved the most effective performance for both 2-labels and 3-labels AMD classifications using retinal images than others. In a DenseNet architecture, the previous Dense layers are connected to all subsequent layers. Therefore, all the necessary features for distinguishing AMD is successfully transferred to flatten layer for the classification task. In addition, since the DenseNet201 has less parameters and needs less resources than other DL models, it will convenience to implement the DL-based assistant tool in production. Our study has provided a valuable recommendation for ophthalmologist about what DL models are suitable and effective to implement automatic tool to help them in AMD diagnosis [37,38].

There are two limitations in our experiments. Firstly, the resolution of retinal images is fixed into (224 × 224) pixels due to the resource limitation of Google Colab, which may lead to loss information in fundus image from larger size. Secondly, we have not applied the pre-processing image techniques such as remove black area, blue color channel or border enhancement, which can improve the accuracy of the DL models during training.

5. Conclusion

In this research, we utilized different DL architectures to figure out the advantaged ones for the classification AMD disease using retinal images. The experiments were conducted using transfer learning technique with the retinal images from our Chula-AMD dataset. In addition, the images for testing trained DL models were from Retinal Image Bank and iChallenge-AMD datasets. The results showed that DenseNet201

Table 6
The 3-labels AMD classification results: precision, recall, F1-Score, and accuracy on validating dataset.

Model	Label	Precision	Recall	F1-Score	Accuracy (%)
VGG19	Normal	0.80	0.67	0.73	0.70 (70%)
	Dry AMD	0.60	0.75	0.67	
	Wet AMD	0.75	0.69	0.72	
Xception	Normal	0.91	0.96	0.93	0.90 (90%)
	Dry AMD	0.90	0.84	0.87	
	Wet AMD	0.90	0.91	0.91	
DenseNet201	Normal	0.87	0.95	0.91	0.87 (87%)
	Dry AMD	0.88	0.78	0.83	
	Wet AMD	0.87	0.88	0.88	
EfficientNetB7	Normal	0.82	0.94	0.88	0.81 (81%)
	Dry AMD	0.85	0.58	0.69	
	Wet AMD	0.77	0.90	0.83	
InceptionV3	Normal	0.86	0.92	0.89	0.86 (86%)
	Dry AMD	0.80	0.84	0.82	
	Wet AMD	0.93	0.82	0.87	
NASNetLarge	Normal	0.82	0.95	0.88	0.80 (80%)
	Dry AMD	0.73	0.76	0.75	
	Wet AMD	0.85	0.67	0.75	
ResNet152V2	Normal	0.79	0.97	0.87	0.86 (86%)
	Dry AMD	0.87	0.80	0.83	
	Wet AMD	0.95	0.80	0.87	

Table 7
The 3-labels AMD classification results: precision, recall, F1-Score, and accuracy on testing dataset.

Model	Label	Precision	Recall	F1-Score	Accuracy (%)
VGG19	Normal	0.61	0.54	0.57	0.52 (52%)
	Dry AMD	0.58	0.22	0.32	
	Wet AMD	0.46	0.80	0.58	
Xception	Normal	0.56	0.82	0.67	0.57 (57%)
	Dry AMD	0.56	0.30	0.39	
	Wet AMD	0.60	0.60	0.60	
DenseNet201	Normal	0.64	0.84	0.72	0.67 (67%)
	Dry AMD	0.82	0.46	0.59	
	Wet AMD	0.64	0.72	0.68	
EfficientNetB7	Normal	0.63	0.82	0.71	0.64 (64%)
	Dry AMD	0.56	0.28	0.37	
	Wet AMD	0.68	0.82	0.75	
InceptionV3	Normal	0.72	0.62	0.67	0.62 (62%)
	Dry AMD	0.53	0.72	0.61	
	Wet AMD	0.67	0.52	0.58	
NASNetLarge	Normal	0.56	0.80	0.66	0.59 (59%)
	Dry AMD	0.56	0.50	0.53	
	Wet AMD	0.70	0.46	0.55	
ResNet152V2	Normal	0.59	0.58	0.59	0.62 (62%)
	Dry AMD	0.56	0.66	0.61	
	Wet AMD	0.74	0.62	0.67	

Table 8
The 3-labels AMD classification results: Top-3 performance of total accuracy and total F1-Score.

Total accuracy (Acc_{total})	Total F1-Score (F1-Score$_{\text{total}}$)	
	Validating	Testing
DenseNet201 (2.52)	Xception (2.71)	**DenseNet201** (1.99)
ResNet152V2 (2.47)	**DenseNet201** (2.62)	ResNet152V2 (1.87)
InceptionV3 (2.46)	InceptionV3 (2.58)	InceptionV3 (1.86)

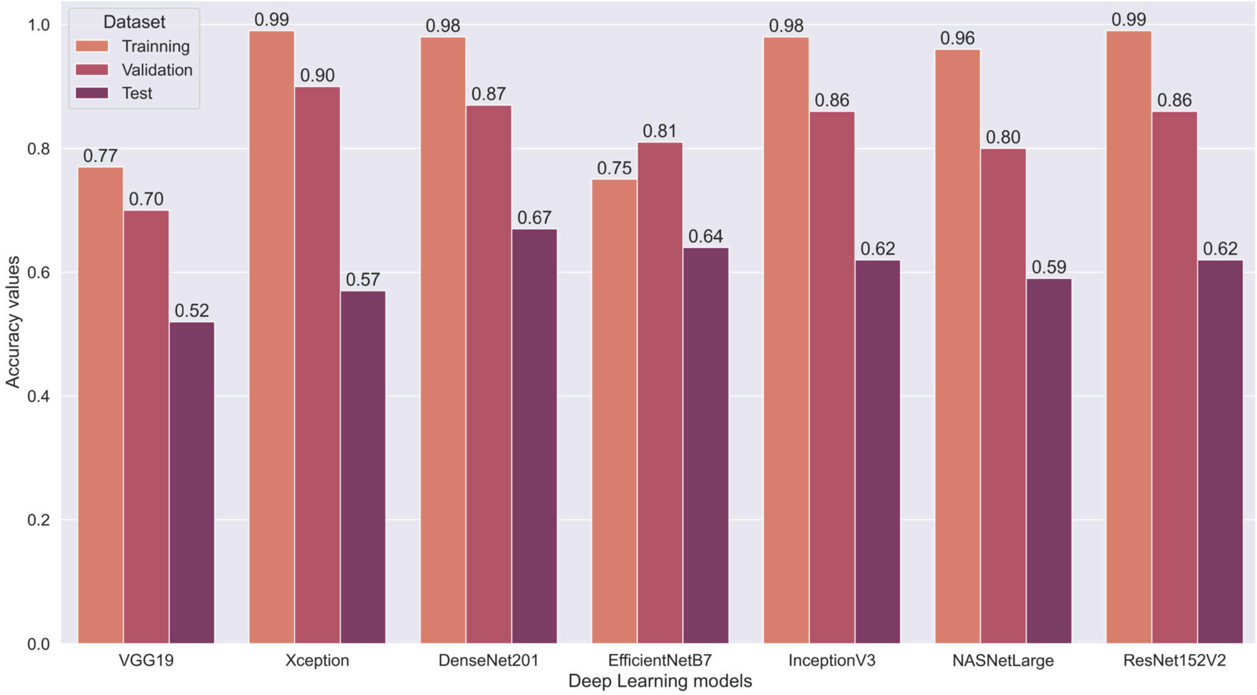


Fig. 4. Accuracy results for 3-labels AMD classification performed on three datasets.

with Dense block in its architecture is most effective DL model for both 2-labels and 3-labels AMD classifications. Furthermore, Xception and EfficientNetB7 are other suitable models for 2-labels AMD classification. InceptionV3 and Resnet152V2 are alternative models for 3-labels AMD classification. All the accuracy performances of trained DL models in Top-3 are comparable with the performance of retinal specialist for both 2-labels and 3-labels AMD tasks.

For future work, we will conduct a dataset of 4-labels AMD classification (e.g.: Normal vs. Early AMD vs. Intermediate AMD vs. Advanced AMD) and train the DenseNet201 deep learning model using this dataset. Another research direction is the performance evaluation of using DenseNet architecture for other eyes diseases classification such as diabetic retinal or glaucoma.

Declarations

The Faculty of Medicine, Chulalongkorn University, Thailand approved the use and storage of retinal images from patients at the Department of Ophthalmology, the King Chulalongkorn Memorial Hospital under Certificate Number 1233/2022.

Funding

This research has received funding support from the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [Grant No. B04G640068].

The research also received support from the Ratchadapisek Somphot fund for Center of Excellence in artificial intelligence, machine learning, and smart grid technology and for postdoctoral fellowships at Chulalongkorn University.

CRediT authorship contribution statement

Ngoc Thien Le:
Conceptualization of this study, Methodology, Software, Original draft preparation, Formal analysis.

Truong Thanh Le:
Experiments, revise and editing.

Pear Pongsachareonnont Ferreira:
Data curation, result validation, revised experiments.

Disorn Suwajanakorn:
Data curation, result validation, revised experiments.

Apivat Mavichak:
Data curation, result validation, revised experiments.

Rath Itthipanichpong:
Data curation, result validation, revised experiments, and supervision.

Widhyakorn Asdornwised:
Data curation.

Surachai Chaitusaney:
Data curation.

Watit Benjapolakul:
Conceptualization of this study, resources, review and editing, visualization, supervision, project administration, funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research has received funding support from the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [Grant No. B04G640068].

The research also received support from the Ratchadapisek Somphot fund for Center of Excellence in artificial intelligence, machine learning, and smart grid technology and for postdoctoral fellowships at Chulalongkorn University.

Appendix A. The 2-labels AMD classification results

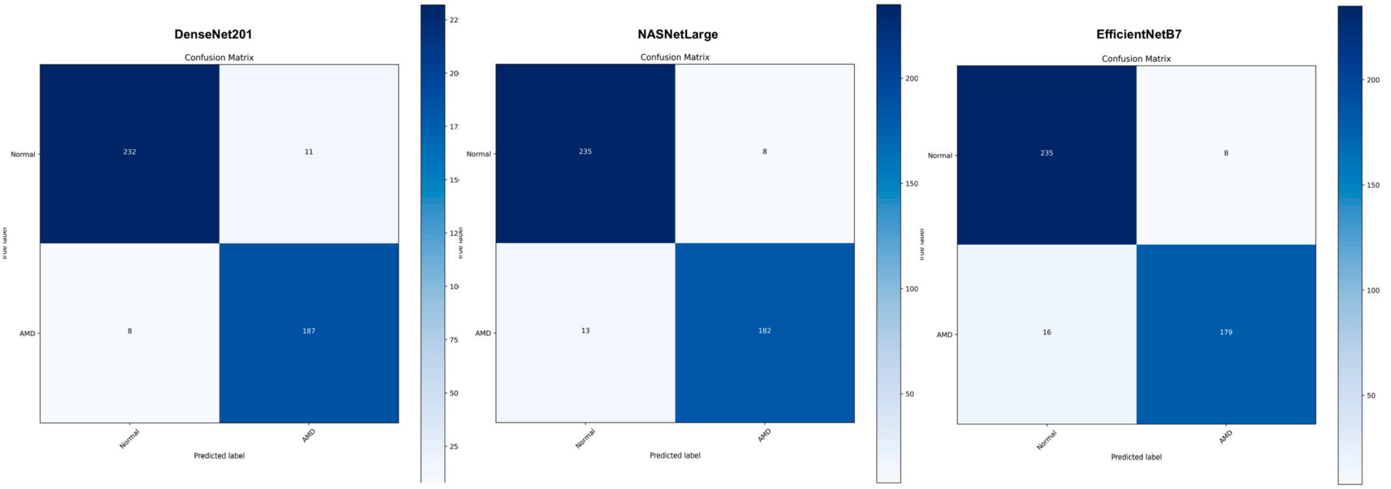


Fig. A.1. The 2-labels AMD classification results: confusion matrix of the Top-3 F1-Score$_{\text{total}}$ performance on validating dataset. First: DenseNet201, Second: NASNetLarge, Third: EfficientNetB7.

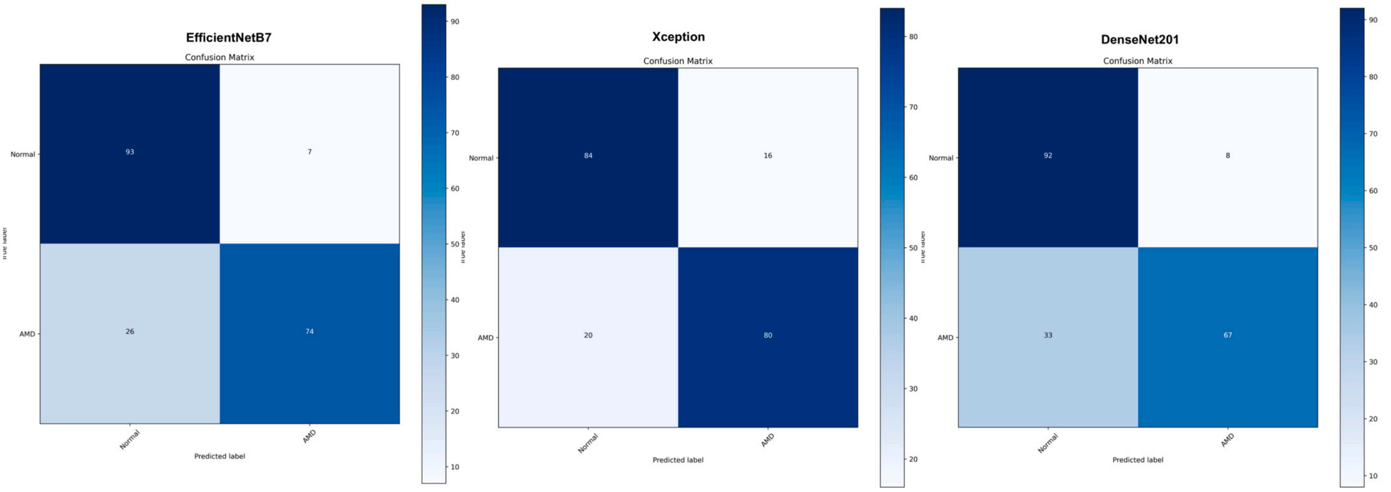


Fig. A.2. The 2-labels AMD classification results: confusion matrix of the Top-3 F1-Score$_{\text{total}}$ performance on testing dataset. First: EfficientNetB7, Second: Xception, Third: DenseNet201.

Appendix B. The 3-labels AMD classification results

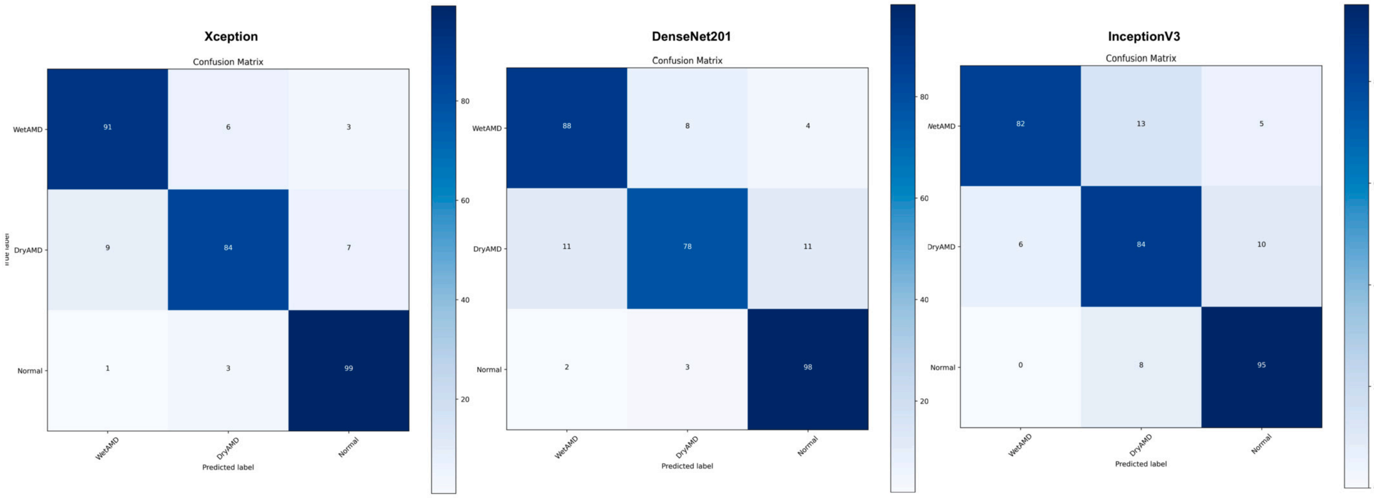


Fig. B.1. The 3-labels AMD classification results: confusion matrix of the Top-3 F1-Score$_{total}$ performance on validating dataset. First: Xception, Second: DenseNet201, Third: InceptionV3.

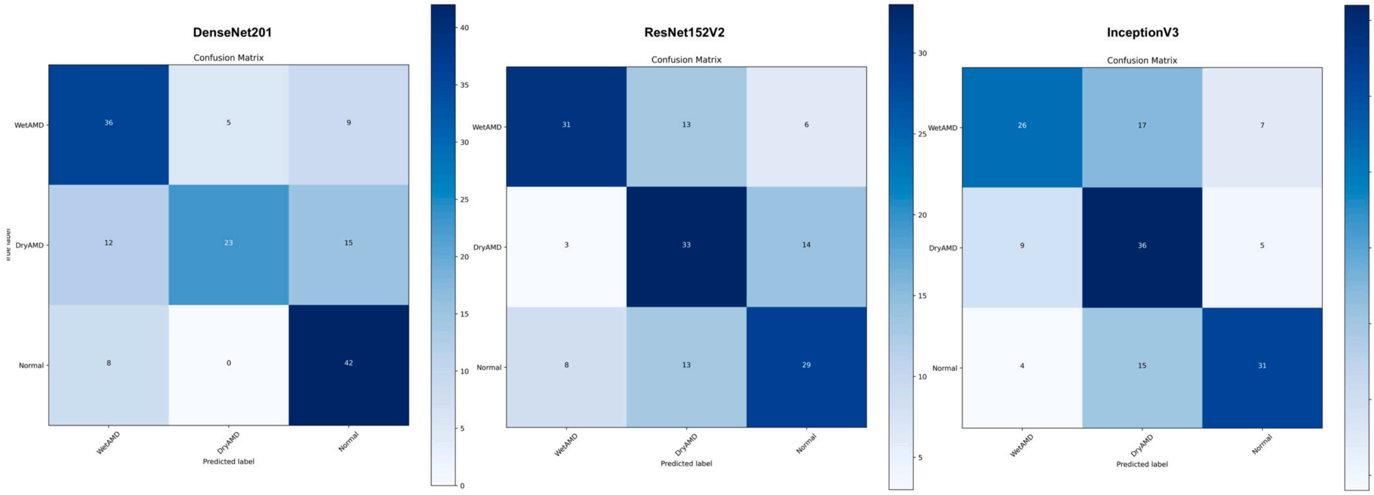


Fig. B.2. The 3-labels AMD classification results: confusion matrix of the Top-3 F1-Score$_{total}$ performance on testing dataset. First: DenseNet201, Second: ResNet152V2, Third: InceptionV3.

References

[1] Jenchitr W, Ruamviboonsuk P, Sanmee A, Pokawattana N. Prevalence of age-related macular degeneration in thailand. Ophthalmic Epidemiology 2011;18:48–52.

[2] Shad R, Cunningham JP, Ashley EA, Langlotz CP, Hiesinger W. Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. Nature Machine Intelligence 2021;3:929–35.

[3] Le WT, Maleki F, Romero FP, Forghani R, Kadoury S. Overview of machine learning: Part 2: Deep learning for medical image analysis. Neuroimaging Clinics of North America 2020;30:417–31.

[4] Kim KM, Heo T-Y, Kim A, Kim J, Han KJ, Yun J, et al. Development of a fundus image-based deep learning diagnostic tool for various retinal diseases. Journal of Personalized Medicine 2021;11.

[5] Leng X, Shi R, Wu Y, Zhu S, Cai X, Lu X, et al. Deep learning for detection of age-related macular degeneration: A systematic review and meta-analysis of diagnostic test accuracy studies. PLOS ONE 2023;18:1–20.

[6] Jeong Y, Hong Y-J, Han J-H. Review of machine learning applications using retinal fundus images. Diagnostics 2022;12:134.

[7] Gong D, Kras A, Miller JB. Application of deep learning for diagnosing, classifying, and treating age-related macular degeneration. Seminars in Ophthalmology 2021;36:198–204.

[8] Perepelkina T, Fulton AB. Artificial intelligence (ai) applications for age-related macular degeneration (amd) and other retinal dystrophies. Seminars in Ophthalmology 2021;36:304–9.

[9] Kumar H, Goh KL, Guymer RH, Wu Z. A clinical perspective on the expanding role of artificial intelligence in age-related macular degeneration. Clinical and Experimental Optometry 2022;105:674–9.

[10] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9.

[11] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–8.

[12] Targ S, Almeida D, Lyman K. Resnet in resnet: Generalizing residual architectures. preprint. arXiv:1603.08029, 2016.

[13] Wu S, Zhong S, Liu Y. Deep residual learning for image steganalysis. Multimedia Tools and Applications 2018;77:10437–53.

[14] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence, vol. 31.

[15] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–9.

[16] Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmology 2017;135:1170–6.

[17] Tan JH, Bhandary SV, Sivaprasad S, Hagiwara Y, Bagchi A, Raghavendra U, et al. Age-related macular degeneration detection using deep convolutional neural network. Future Generation Computer Systems 2018;87:127–35.

[18] Matsuba S, Tabuchi H, Ohsugi H, Enno H, Ishitobi N, Masumoto H, et al. Accuracy of ultra-wide-field fundus ophthalmoscopy-assisted deep learning, a machine-learning technology, for detecting age-related macular degeneration. International Ophthalmology 2019;39:1269–75.

[19] Govindaiah A, Hussain MA, Smith RT, Bhuiyan A. Deep convolutional neural network based screening and assessment of age-related macular degeneration from fundus images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). p. 1525–8.

[20] Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading amd: A study in using universal deep features and transfer learning for automated amd analysis. Computers in Biology and Medicine 2017;82:80–6.

[21] Horta A, Joshi N, Pekala M, Pacheco KD, Kong J, Bressler N, et al. A hybrid approach for incorporating deep visual features and side channel information with applications to amd detection. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 716–20.

[22] Kim KM, Heo T-Y, Kim A, Kim J, Han KJ, Yun J, et al. Development of a fundus image-based deep learning diagnostic tool for various retinal diseases. Journal of Personalized Medicine 2021;11.

[23] Chakraborty R, Pramanik A. DCNN-based prediction model for detection of age-related macular degeneration from color fundus images. Medical & Biological Engineering & Computing 2022;60:1431–48.

[24] Morano J, Hervella Álvaro S, Rouco J, Novo J, Fernández-Vigo JI, Ortega M. Weakly-supervised detection of amd-related lesions in color fundus images using explainable deep learning. Computer Methods and Programs in Biomedicine 2023;229:107296.

[25] Pan Y, Liu J, Cai Y, et al. Fundus image classification using inception v3 and resnet-50 for the early diagnostics of fundus diseases. Frontiers in Physiology 2023;14:160.

[26] Yellapragada B, Hornauer S, Snyder K, Yu S, Yiu G. Self-supervised feature learning and phenotyping for assessing age-related macular degeneration using retinal fundus images. Ophthalmology Retina 2022;6:116–29.

[27] Skevas C, Weindler H, Levering M, Engelberts J, van Grinsven M, Katz T. Simultaneous screening and classification of diabetic retinopathy and age-related macular degeneration based on fundus photos—a prospective analysis of the retcad system. International Journal of Ophthalmology 2022;15:116–29.

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

[29] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–8.

[30] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, PMLR. pp. 6105–14.

[31] 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. pp. 2818–26.

[32] Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8697–710.

[33] Bengio Y. Learning deep architectures for ai. Foundations and Trends® in Machine Learning 2009;2:1–127.

[34] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review 2020;53:5455–516.

[35] AREDS, National eye institute (nei) age-related eye disease study (areds), 2023. Accessed 20 April 2023.

[36] Acharya UR, Hagiwara Y, Koh JE, Tan JH, Bhandary SV, Rao AK, et al. Automated screening tool for dry and wet age-related macular degeneration (armd) using pyramid of histogram of oriented gradients (phog) and nonlinear features. Journal of Computational Science 2017;20:41–51.

[37] Burlina P, Pacheco KD, Joshi N, Freund DE, Bressler NM. Comparing humans and deep learning performance for grading amd: A study in using universal deep features and transfer learning for automated amd analysis. Computers in Biology and Medicine 2017;82:80–6.

[38] Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks. JAMA Ophthalmology 2017;135:1170–6.

[39] Malik Usman. Age related macular degeneration – understanding the difference between dry amd and wet amd. https://irisvision.com/difference-between-dry-vs-wet-age-related-macular-degeneration/, 2018. [Accessed 20 April 2023].

[40] Retina Health Series. Age-related macular degeneration. 2016.

[41] Bressler NM, Silva JC, Bressler SB, Fine SL, Green WR. Clinicopathologic correlation of drusen and retinal pigment epithelial abnormalities in age-related macular degeneration. Retina 1994;14:130–42.

[42] Klein R, Klein BE, Linton KL. Prevalence of age-related maculopathy: The beaver dam eye study. Ophthalmology 1992;99:933–43.

[43] Klein R, Klein BE, Jensen SC, Meuer SM. The five-year incidence and progression of age-related maculopathy: The beaver dam eye study. Ophthalmology 1997;104:7–21.

[44] image bank Retinal. http://imagebank.asrs.org/home, 2023. [Accessed 20 April 2023].

[45] iChallenge-AMD. Baidu research open-access dataset. https://ai.baidu.com/broad/introduction?dataset=saoke, 2023. [Accessed 20 April 2023].

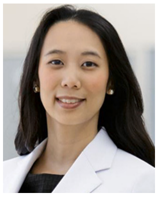
[46] Keras API. Deep learning models. https://keras.io/api/applications/, 2023. [Accessed 20 April 2023].

[47] ImageNet. The ImageNet project. https://www.image-net.org/index.php, 2023. [Accessed 20 April 2023].

[48] Brownlee Jason. A gentle introduction to transfer learning for deep learning. https://machinelearningmastery.com/transfer-learning-for-deep-learning/, 2017. [Accessed 20 April 2023].

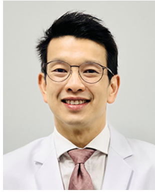
Ngoc Thien Le received the B.E. degree in electrical engineering from HCMC University of Technology and Education, Vietnam, in 2006, and the M.E. degree in telecommunication from Ho Chi Minh City University of Technology, Vietnam, in 2008. He got the Ph.D. degree at the Department of Electrical Engineering, Chulalongkorn University, Thailand in 2016. His research interests include data analysis for smart grid and smart grid communication.

Truong Thanh Le received the B. Eng. degree in Electrical and Electronic Engineering from the Ho Chi Minh City University of Technology, Viet Nam, in 2020, and the M.Eng. degree in Electrical Engineering from Chulalongkorn University, Thailand, in 2022. His research interests are IoT, AI, embedded system, cloud/edge computing.



Pear Pongsachareonnont Ferreira is Assistant Professor of Ophthalmology at Chulalongkorn University. She is the head of the Research Unit, department of Ophthalmology, Chulalongkorn University. She graduated with her vitreoretinal fellowship from the University of Toronto, Canada, and MPH in Epidemiology from Harvard T.H. Chan School of Public health. Her main research area is retinal imaging, vitreoretinal surgery, and the quality of life in visual impairment patients.

Disorn Suwajanakorn is with Center of Excellence in Retina, Faculty of Medicine, Chulalongkorn University. His current research interest includes vitreo-retinal diseases. His current research interest includes vitreo-retinal diseases.



Apivat Mavichak Clinical Instructor / Faculty Staff Retina and Vitreous Research Unit, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand, Ophthalmologist / StaffDepartment of Ophthalmology, King Chulalongkorn Memorial,Pramongkutklao College of Medicine, Bangkok, Thailand Doctor of Medicine (1998 – 2004) Hospital, Thai Red Cross Society, Bangkok, Thailand, Specialist / ConsultantService Center of Retinopathy of Prematurity, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand.

Rath Itthipanichpong received the Doctor of Medicine degree at Siriraj Hospital, Mahidol University, in 2006. He has completed his Ophthalmology Residency Training and Glaucoma Fellowship Training at King Chulalongkorn Memorial Hospital, in 2014 and 2015, respectively. He is currently an Instructor in ophthalmology at the Department of Ophthalmology, Chulalongkorn University, Bangkok, Thailand. His work has been related to both glaucoma and low vision field. From his passion in technology and innovation, he is also developing new technology to help improve quality of care in ophthalmology patients.

Widhyakorn Asdornwised received M.Sci. Electrical Engineering, Drexel University in 1999 and the D.Eng. Electrical Engineering, Chulalongkorn University in 2005. He is now Assistant Professor with the Department of Electrical Engineering, Chulalongkorn University. He does research in Artificial Intelligence, Data Mining and Artificial Neural Network. The current projects are ‘5G Chula’, Micro-operator, and 5G/6G Private Network.

Surachai Chaitusaney obtained the Ph.D. degree from The University of Tokyo, Japan, with JICA scholarship in 2007. At present, he is an Assistant Professor and the Head of Power System Research Laboratory (PSRL) at Chulalongkorn University. His research interests include renewable energy, solar PV, distributed generation, power system planning and reliability, and smart grid regulation. In addition, he has served as a member of subcommittee, committee and working group in several professional organizations, such as Engineering Institute of Thailand (EIT), Energy Policy and Planning Office (EPPO), Energy Regulatory Commission (ERC), and Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Association.

Watit Benjapolakul received the D. Eng. degree in electrical engineering from the University of Tokyo in 1989. He is now Professor with the Department of Electrical Engineering, Chulalongkorn University, Thailand. At present, he is the head of the Center of Excellence in Artificial Intelligence, Machine Learning and Smart Grid Technology, Faculty of Engineering, Chulalongkorn University. His current research interests include mobile communication systems, broadband networks, applications of artificial intelligence in communication systems, wireless networks, and applications of communication networks in smart grids.