VisionAI – Human Eye Fundus Screening System Based on AI Deep Learning Technology

Sungju Park[1]

[1]Seoul Scholars International, Seoul, South Korea

ABSTRACT

The recent development in artificial intelligence contributed to the utilization of information and communication technologies in medical fields. In ophthalmology, the fundus photograph decoding technology is receiving wider attentions as it can easily detect the retinal disorders without having to embrace side effects or inconveniences that follow with pupil dilation test. As AI-based diagnostic technologies can effectively discover disorders in optic nerves, optic layers in retina, retinal vessels, it can be useful in early detection and health checks. This study therefore develops a model which classifies and analyzes 24,000 fundus photographs into four categories (normal, cataract, glaucoma, diabetic retinopathy) based on diagnostic data. The model is further realized into a website which will contribute to effective diagnoses of fundus diseases. Convolutional neural network (CNN, specialized in image processing) is applied as a learning model and EfficientNet was used to configure the network. Hyperparameter optimization was used for tuning, and the developed model is later realized as a public webpage. The designed model proves its excellence by reaching accuracy of 90.8% and other evaluation results. For the enhancement of performance, this model would necessitate extensive datasets and more intricate classifications of fundus diseases through the collaborative research with medical institutions. The author anticipates more prompt diagnosis and treatment for patients with reduced accessibility and quicker diagnosis for medical professionals.

Introduction

Artificial intelligence technology has already been widely utilized in computer-aided diagnosis (CAD). It identifies types of disorder and distinguishes minor differences between similar disorders, by analysing and classifying photographic and video data. In continuation to this trend, this study seeks to enhance the diagnosis of various eye disorders such as cataract, glaucoma, and diabetic retinopathy, by applying deep learning technologies to decoding of fundus photographs.

Ocular fundus refers to an area that is covered with retina, which accounts for two thirds of the rear part of the inner eyeball. This is the only area where major vessels can be directly observed. As ocular fundus is linked to central nerves (brain) via optic nerves, general eye disorders, systemic diseases, and central nervous disorders often accompany ocular fundus diseases. Hence, the observation of ocular fundus is beneficial in diagnosing and determining the severity and progress of a given disorder. Fundoscopic examination observes the surface of retina through pupil. Here, the changes in retina, optic disc, choroid are observed. The general practice is to insert mydriatic into eye and an experienced ophthalmologist observes the ocular fundus with an ophthalmoscopy or a lens, 30 to 40 minutes after inserting eye drop. Ocular fundus can be observed with slit lamp or indirect ophthalmoscope. [1]

Pupil dilation test takes time, hence causes inconvenience to patients and requires a significant involvement of ophthalmologists. As such, this diagnostic solution has limitations in providing medical services to those living in areas with reduced accessibility to eye treatments. Where fundus cameras are used, such limitations can be somewhat overcome. Non-mydriatic fundus cameras can record the post pole of retina without inserting eye drop. This device is widely used for health checks as it is useful for intricate determination and follow-up of diabetic retinopathy, promptly detecting various eye disorders. This study also uses fundus photographs taken with fundus camera. [1]

To minimize the loss of time and labour and to enhance the efficiency and convenience of diagnosis, this paper develops the model which diagnoses fundus diseases based on a given fundus photograph and provides a website which offers diagnosis results. When neurons receive certain stimulus, it reacts to with a behavior. However, when the feedback is negative, the neuron deactivates the neural cell which triggered the behavior. If the feedback is positive, the neural cell becomes even more activated. The network created by the connection between neurons is referred to as neural network, and the computerized version of its realization is artificial neural network (ANN). Deep neural network (DNN) refers to an ANN which is created by inserting numerous hidden layers between input and output layers. On the other hand, convolutional neural network (CNN) contains convolution layers and pooling layers in addition to DNN. This model minimizes the loss of two-dimensional information (spatial and local) in images and videos, improving the shortcomings of DNN which only deals with one-dimensional data.

This paper adopts EfficientNet, a model derived from CNN. The model is trained with 24,000 fundus photographs and provide diagnoses from four categories (normal, cataract, glaucoma and diabetic retinopathy). Further, this model is also synchronized with a webpage for the purpose of supporting ordinary users and medical professionals, so that a user can obtain a diagnosis if they upload a fundus photograph on the webpage.

The purpose of this study is to minimize the side effects of pupil dilation test (acute glaucoma) and costs and obstacles associated with the test. It aims to enhance the accessibility to medical treatments for eye disorders and to provide more accurate diagnoses.

Materials and Methods

Collection of Fundus Photographs

For the purpose of training, this study collected and combined seven data sets of fundus photographs provided by medical institutions and hospitals around the world. A total of 24,213 photographs were collected and each photograph was classified according to the diagnosis of ophthalmologist. Upon categorization, the number of photographs of normal fundus was 3,398, while those of cataract accounted for 1,214, and 2,327 for glaucoma, and 17,274 for diabetic retinopathy. The collected data sets are as follows.

Data set from Bajwa Hospital (Multi-eye Disease Dataset)

This data set consists of fundus photographs from Bajwa Hospital in Dina Nagar, Gurdaspur, India (date: April 20, 2022) The photographs are classified into four categories: 300 for normal, 100 for cataract, 101 for glaucoma and 100 for other retinal diseases. [2] This study only used the normal, cataract, and glaucoma photographs.

ODIR-2019 Data Set

Ocular disease intelligent recognition (ODIR) data sets are structured ophthalmologist data sets of which individual photographs consist of fundus photographs of 5,000 patients (both left and right eyes) as well as medical diagnoses of individual photographs. This data includes patient information collected by Shanggong, a Chinese medical technology company, from various hospitals and medical centers. The photographs were taken with various camera models such

as Canon, Zeiss, and Kowa and come in various resolutions. This data set is classified into eight labels (normal, diabetic eye disorder, glaucoma, cataract, AMD, hypertension, pathological myopia, and other eye disorders). [3] This study only included normal, diabetic retinopathy, glaucoma, and cataract photographs in data set for training.

Eyepacs Diabetic Retinopathy Data Set

Eyepacs, a free diagnostic platform for diabetic retinopathy, provides 35,126 high resolution fundus photographs. Each photograph is assigned with a grade (out of five), and the assignment was carried out by medical professionals according to the condition of fundus. (0- non-diabetic retinopathy, 1- mild diabetic retinopathy, 2-medium diabetic retinopathy, 3-severe diabetic retinopathy, 4- neovascularization diabetic retinopathy). [4] This study did not include the photographs which are categorized as “0- non-diabetic retinopathy”.

OIA (Ophthalmic Image Analysis) DDR Data Set

This data set includes 13,673 fundus photographs of diabetic retinopathy from 147 hospitals in 23 regions in China. Each photograph is assigned with a grade (out of five), and the assignment was carried out by medical professionals according to the condition of fundus. (0- non-diabetic retinopathy, 1- mild diabetic retinopathy, 2-medium diabetic retinopathy, 3-severe diabetic retinopathy, 4- neovascularized diabetic retinopathy). [5] This study did not include the photographs which are categorized as “0- non-diabetic retinopathy”.

RIGA Data Set

This data set contains three separate data sets of different diabetic retinopathies: Bin Rushed, MESSIDOR, and Magrabi. The first data set is comprised of 195 fundus photographs of diabetic retinopathy provided by Bin Rushed eye hospital. The second data set, MESSIDOR, is a data prepared by the French Ministry of Defense in 2004, for the purpose of automatic diagnosis of diabetic retinopathy using artificial intelligence. This includes 460 fundus photographs. The final data set includes 95 fundus photographs collected by Magrabi eye hospital. Each photograph in three data sets is provided with diagnoses of six ophthalmologists and there are a total of 750 photographs. [6]

REFUGE (Retinal Fundus Glaucoma Challenge) Data Set

This data set is provided by the glaucoma diagnosis competition and contains 1,200 fundus photographs. Each photograph is attached with the diagnosis by physicians specialized in glaucoma. The diagnoses were provided by seven glaucoma specialists from Zhongshan eye hospital and Sun Yat- Sen University, and the diagnosed areas are described in detail. [7]

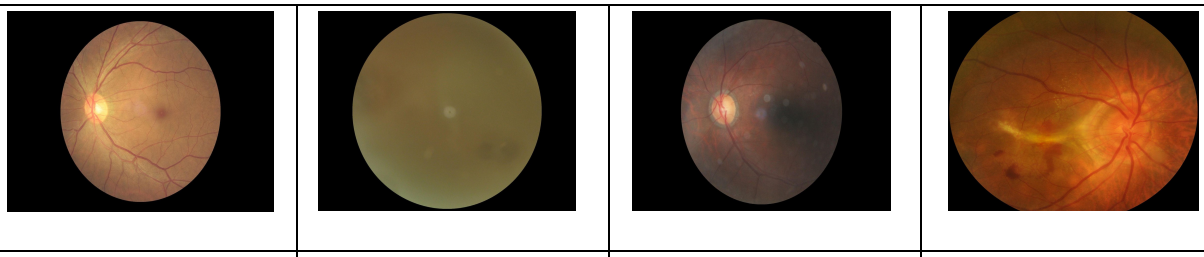
Immature Cataract Fundus Images

This data set includes 800 cataract fundus photographs provided by the database from Telkom University. [8]

Data Processing

Classification of Images

A total of 24,213 fundus photographs (combining seven data sets) were classified into four categories (normal, cataract, glaucoma, and diabetic retinopathy). The photographs that do not fall within the categories were not included. The category features of the four categories are later converted into numerical values through label encoding. Normal photographs are assigned 0, cataract photographs are assigned 1, glaucoma photographs are assigned 2, and diabetic retinopathy photographs are assigned 3. The photographs exhibited different features by each class. The explanation of each class is as follow.



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

Figure 1. Fundus photo classification

Normal fundus photographs (a) have clear retinal surfaces and optic discs and retinal vessels are clearly visible. The cataract photographs (b) exhibit overall media opacity. The glaucoma photographs (c) show changes in various optic discs and focal defects in retinal nerve fibers. Lastly, the diabetic retinopathy photographs (d) gradually show microaneurysm, hard exudate, cotton-wool patch, and new vessels.

Image Processing and Normalization

As the fundus photographs collected from the seven data sets are in different sizes, the images were resized into 192* in height and 256* width, to the extent that do not lose the information. As data sets are integrated at a later stage, the colorized and grey images are combined. In consideration of consistent training, RGB values were adjusted, and all images were converted into grey. Red, green and blue values are divided into 255 (RGB values are expressed into 0 to 254) and normalized in values between 0 to 1 in order to be processed in the model. Finally, the quantity of images increased by image augmentation (e.g. 180° reversion, filling in adjacent empty spaces with reversed images etc.) and improved the consistency of images in data sets.

Model Structure

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 192, 256, 3)]     0

efficientnet-b0 (Functional  (None, 6, 8, 1280)        4049564
)

global_average_pooling2d (G  (None, 1280)              0
lobalAveragePooling2D)

dense (Dense)                (None, 4)                 5124

=================================================================
Total params: 4,054,688
Trainable params: 4,012,672
Non-trainable params: 42,016
_________________________________________________________________
```

Figure 2. Composition of Model Layers

This study adopts a model in which EfficientNet B-0 network was applied to convolutional layer. EfficientNet is a model which was designed to maximize the efficiency with limited resource; it is a state-of-the-art (SOTA) model which find the optimal combination of model scaling (depth, width, resolution, and scaling) through AutoMI. According to [9], the aforementioned three variables are closely linked together, and the article suggests the proportionate scaling. This is called “compounding scaling” and it finds the optimal combination of three factors by allowing the increase of width, depth, and resolution of feature map. The performance will be enhanced if the optional solution derived from this is applied to the baseline network. The network uses mobile inverted bottleneck convolution (MBConV) [10] as a main block and carried out a transfer learning by retrieving the weight that was trained with ImageNet data in advance. The composition of EfficientNet B-0 model is as follows.

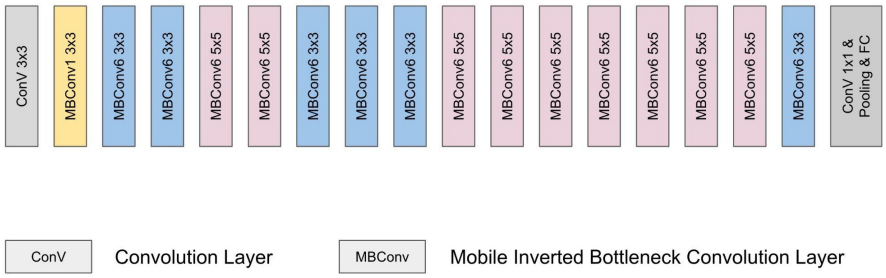


Figure 3. composition of EfficientNet B-0 model layers

The final tensor derived from EfficientNet layer is still large and it creates approximately four million parameters if flattened in one-dimensional form. This may create overfitting. Hence, the average values of each channel were extracted by using global average pooling (GAP) which effectively reduces parameters. The output is conducted in dense layer. Here, softmax function is used to determine the category of diagnosis. To improve the performance, Adam optimizer was used to enhance the performance, and the learning rate was set at 0.00003. Categorical

cross – entropy loss (softmax loss) was used as loss function and label smoothing was set at 0.01. Early stopping function was applied, which terminates the training even though the mode does not reach the designated epoch numbers to avoid the overfitting of data. This helps the model find the adequate epoch and was realized by using Keras Callback modules. Further, ReduceLROnPlateau function was applied to improve the model by adjusting the learning rate. To achieve the optimization of the model, hyper parameter tuning was conducted to find the optimal combination between learning rate and batch size. This study adopts random search tuning techniques and applied the optimal combination of learning rate, batch size, and weight, which can boost the performance, to the model.

Experimental Results

Implementation Details

Training was carried out by MacBook Pro (13-inch, 2019, Two Thunderbolt 3 ports) PC. 1.4 GHz Quad-Core Intel Core i5 was used for CPU and Intel Iris Plus Graphics 645 1536 MB was used for GPU. Jupyter Notebook was executed via anaconda python compiler. In realizing the model, Python and Pandas were used for data storage. Python-OpenCV was used for image conversion. TensorFlow and Keras were used for creating model layers, while Sckit Learn was used for assorting data. EfficientNet library was used for generating a core model. 24,000 fundus photographs were created in dataframe and saved (used Pandas) and to verify the deep learning model, Sckit Learn library was used to set train, test, and validation data at the ratio of 7:2:1. The assorted images were adjusted to batch size 32, 192* height, and 256* width, using Python OpenCV library.

Model Evaluation

Evaluation Metrics

To assess the performance of the trained deep learning model, the following criteria were used to generate the values.

		Predicted	
		Positive (1)	Negative (0)
Actual	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Table 1. Confusion Matrix

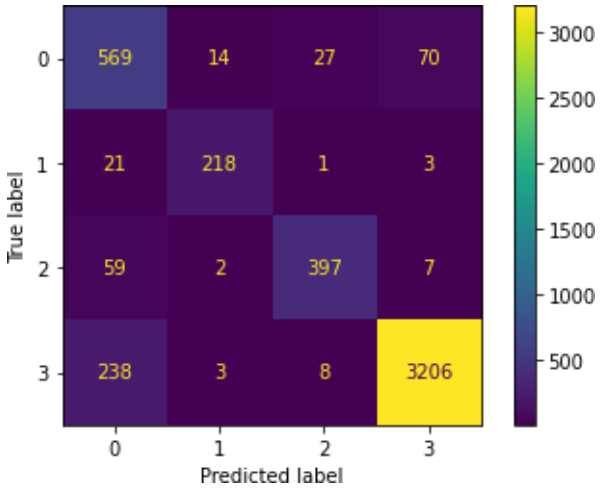


Figure 4. Confusion Matrix on Test Dataset

Confusion Matrix refers to a table that compares the estimate and actual values to measure the anticipated performance. The matrix is used to assess the performance of classification model. TP refers to the values whose category is accurately classified, while FN refers to the values whose category is inaccurately classified. FP refers to the values which incorrectly indicate the category, while TN refers to the values which correctly indicate wrong category. With these four values, accuracy, precision and recall can be evaluated and the explanations of each formula are as follow.

- (1) Accuracy: the frequency of incidents where the predicted value and actual values correspond.
- (2) Precision: the frequency of incidents where the predicted positive values and actual positive values correspond.
- (3) Recall: the ratio of data whose actual and predicted values are both positive. This is also called “true positive rate”.

$$accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$
$$precision = TP/(TP + FP) \quad (2)$$
$$recall = TP/(TP + FN) \quad (3)$$

Accuracy and Loss Rate

Accuracy refers to the rates of correct predictions out of the entire data set and is the most common criterion for assessing the model. The overall accuracy of the model was 90.8% and the accuracy of learning and verification is set out in Figure 5 below.

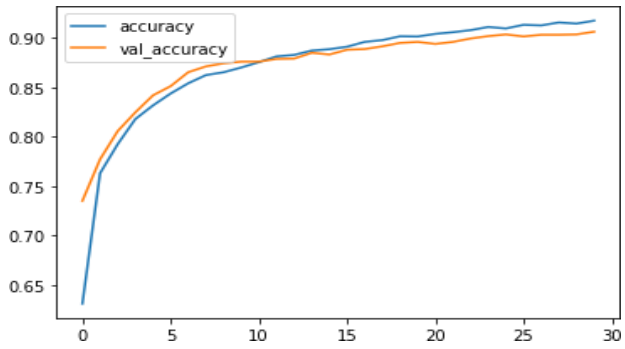


Figure 5. Accuracy of Training and Verificaion dataset

Loss rate refers to the margin of error between the predicted and actual values in model. The bigger the loss is, the larger the margin becomes. The loss rate was 26.2% in this model and from 15 epoch onwards, the change stabilizes.

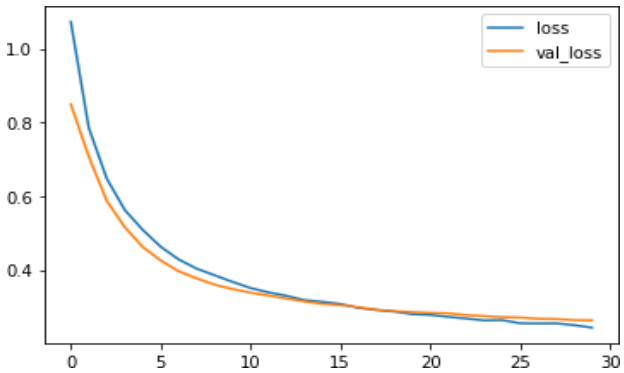


Figure 6. Loss Rate of Training and Verificaion dataset

ROC and AUC

ROC curve is a graph representing the efficiency of the model by using sensitivity and uniqueness. The area size between the curve and x axis is called area under curve (AUC); a model is said to be efficient if the AUC value is closer to 1. The categories in this model (normal, cataract, glaucoma, and diabetic retinopathy) were extracted in 0, 1, 2, and 3; although there are minor differences, the values of all categories are close to 1. However, the classification of 0 (normal fundus photograph) varies from medical opinions and is determined on the basis of minor symptoms and individual constitution. As such, its scope is too wide and the classification shows a relatively low value. In addition, where a photograph has too much noise, minor diagnostic evidence cannot be adequately identified. This also causes an expansion of a normal classification.

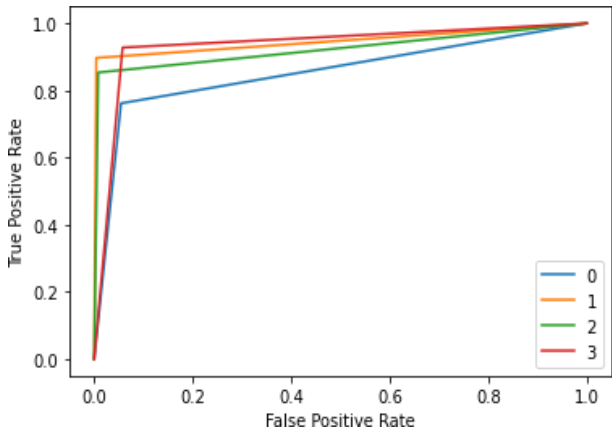


Figure 7. Receiver Operating Characteristic curve

Other Criteria

In case where data is imbalanced, precision, recall rate, and F1 scores are also assessed. In case where data is imbalanced, precision, recall rate, and F1 scores are also assessed. Precision refers to the rate of true classifications which were classified by the model as 'true', while recall rate refers to the rate of those classified by the model as 'true', which are true classifications. Hence, this can be used as an independent metric and the harmonized average of F1 score can be derived to be affected by smaller values.

$$F1\ score = 2 * \frac{precision * recall}{precision + recall} \quad (4)$$

	precision	recall	f1-score	support
0	0.70	0.76	0.73	680
1	0.92	0.90	0.91	243
2	0.92	0.85	0.88	465
3	0.98	0.93	0.95	3455
micro avg	0.92	0.90	0.91	4843
macro avg	0.88	0.86	0.87	4843
weighted avg	0.93	0.90	0.91	4843
samples avg	0.90	0.90	0.90	4843

Figure 8. Table of Precision, Recall, F1, and Support by Classification

As shown in the table, the margins of error between precision and recall in each classification are not large, hence confirms that the model can reliably process complex and imbalanced data. However, as it can be seen from 3.2.3, 0 (normal fundus) category involves vagueness which relies on individual medical opinion. Hence, its performance is relatively lower than other classifications.

Web Development

To support medical services to wider populations, the model was realized into a website. The overall structure is as follows. For the realization of server, Flask (a web server library that can be realized by Python) was used. By uploading the form that was introduced to the first page, the website received fundus photographs from users. By executing h5 model file program within the server, photographs were analysed. The results were then delivered to the form in the second page and as suggested in the page, the results are then delivered to the user. HTML and CSS were used for the realization of website, and Javascript framework was realized by using bootstrap.

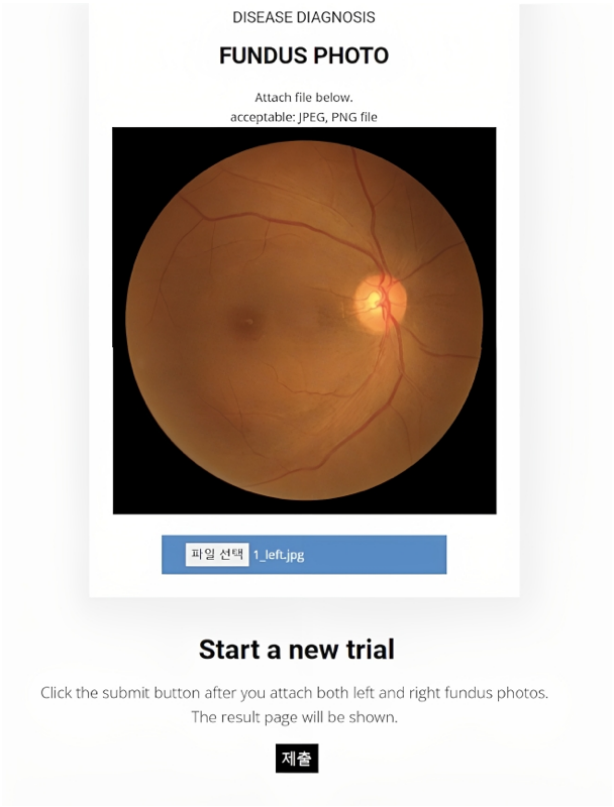


Figure 9. first page of the website

The first page has an upload function. Once clicking the ‘submit’ button after selecting a fundus photograph, the image is saved in img folder in the server. Subsequently, the image is analysed by the internal deep learning model and the results are forwarded to the result page. The result page is comprised of fundus type, treatment, and analysis labels. In fundus type label, the name of diagnosis and its explanation is printed out. In treatment label, the treatment method is described and the results can be found in analysis label.

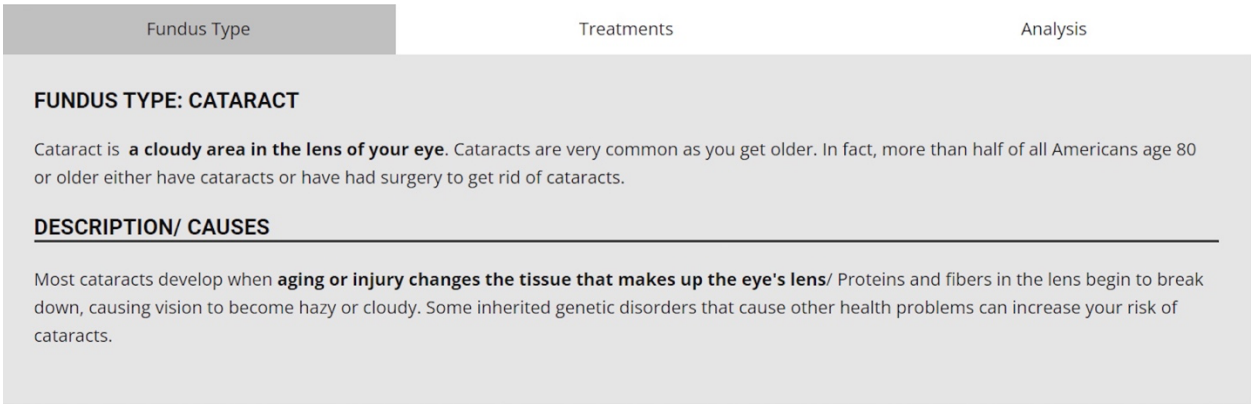


Figure 10. result page 1 in case of cataract

This page shows which category that a given photograph falls under: normal, cataract, glaucoma, and diabetic retinopathy. It also provides the cause of the disease and its definition. The figure above is an example of description displayed when a photographic diagnosis indicates cataract. It explains the media opacity in the lens and that the cause of the disease is related to ageing, trauma, systemic disease, and inflammation in the eye.

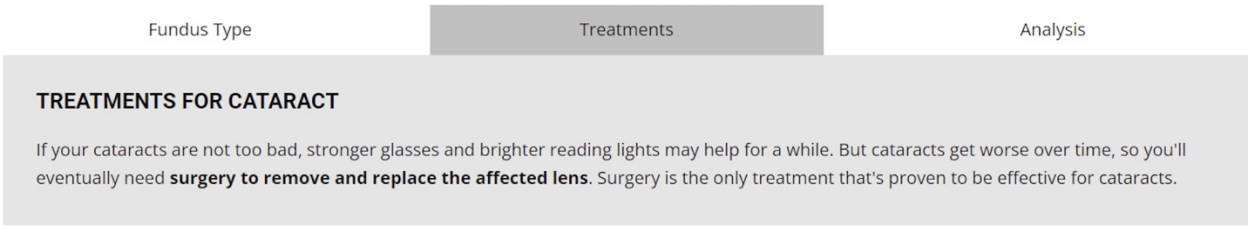


Figure 11. result page 2 in case of cataract

This page explains potential treatments for the disease. In case of cataract, the opaque lens can be removed by operation and an artificial lens can be inserted.

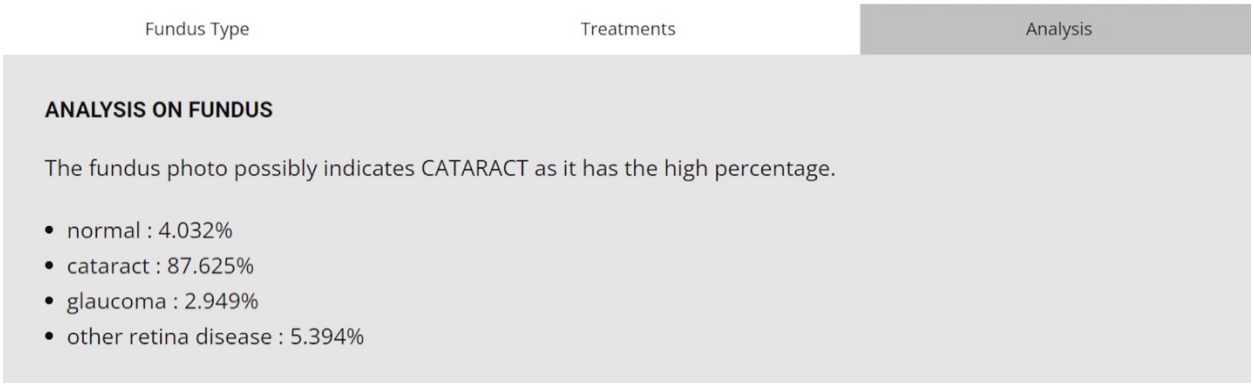


Figure 12. result page 3 in case of cataract

The analysis of deep learning model is carried out in the last page. The photograph here is determined as cataract at the rate of 87.6%, hence the diagnosis is ‘cataract’. The possibility of normal fundus is 4%, 2.9% for glaucoma, and 5.4% for diabetic retinopathy.

Discussion

The development of machine learning based fundus photograph classifying models is on progress. [11] [12] However, most research only concerns with determination of specific disease, and this suggests the limitation of fundus photographic diagnosis.

‘Development of AI-based Cataract Diagnosis Platform’ set up a diagnosis webpage based on the algorithm trained with online image crawling. [11] Published in March 2022, this paper collected data through image crawling which uses Selenium and Beautifulsoup. By using the data, it trained the image-classifying AI model via Teachable Machine supported by Google. Afterwards, it also developed a web page that provides diagnostic results.

However, there was no point of improvement in the model as it employed the basic format supported by Teachable Machine. Besides, the model only dealt with the classification of cataracts and did not include other eye disorders. The diagnosis accuracy of cataract was 82.83%, while the rates of diagnosing normal eyes and red eyes were 61% and 56.38% respectively, which was not as competent as similar studies. The final point to note is that this study did not use the verified fundus photographs and used the crawled data obtained by image search. As cataract cannot be easily diagnosed by investigating the surface photograph, this method reduces the accuracy of diagnosis.

In October 2019, ‘Classification of RoP Using Deep Learning’ was published in Journal of Korean Institute of Information Technology. [12] This paper produced a classification device using CNN algorithms, for the purpose of diagnosing retinopathy of prematurity (RoP) via fundus photographs. The accuracies were compared by applying AlexNet, Inception-V3, Xception 3, and ReLu function was used. Smartphone cameras were used to take fundus photographs as the recording of ocular fundus image was impossible on premature infants and taking the photographs on those patients necessitated expensive imaging device (RetCam) and experienced photographer. The classification device produced 99% of accuracy.

Nonetheless, the limitation of this paper was clear. The paper only classified images into abnormal and normal; although the paper trained the model with 25,812 images using image augmentation, a small data set of 478 fundus photographs from 239 patients can be a critical shortcoming. Hence, the reliability of the model cannot be estimated as over 90%, and a larger data set needs to be verified for more users.

This study improved a significant number of issues in preceding research and provides a systematic web service which only uses verified fundus photographic data. The model can diagnose four types of diseases (normal, cataract, glaucoma, and diabetic retinopathy), and as it demonstrated over 90% of accuracy under the multiclass classification, rather than diagnosing a specific disease. As such, the model’s comparative competence can be said to be superior. In addition, this model also developed data processing technology which can uniformly size the images (regardless of color). This minimizes the loss of information and reduces the processing time and allows an easy application to other data sets, which is convenient for scaling.

Conclusion

Despite the limitations of this study, the Efficient Net, CNN deep learning model trained with a limited amount of fundus photographic data showed 90.8% accuracy and corresponding evaluation results in classifying four types of fundus diseases.The author anticipates this model to be utilized beyond its primary purpose of diagnosing various retinal diseases. Ultimately, this model seeks to contribute to the speedy treatment of the populations with reduced accessibility to medical services, by utilizing the internet as a means of increasing accessibility.

This study needs to be improved by the verification of model, which can handle the processing of real-time data and extensive data sets. If more photographic data is collected from domestic and international medical institutions to offer correct diagnoses and prevent overfitting, the performance of this model can be boosted up to SOTA level. In addition, if the diagnostic solution can be improved in a way that can diagnose a variety of retinal diseases such as AMD, retinal vascular diseases, inflammatory disease, and intraocular tumor, a wider range of retinal disorders can be diagnosed using this model.

References

[1] Hyung-Gon Yoo, Fundus Examination, Seoul National University School of Medicine, Ophthalmology Seminar
[2] Kaur, Palwinder (2022), “Bajwa Hospital (Multi Eye Disease Dataset)”, Mendeley Data, V3, doi: 10.17632/rgwpd4m785.3
[3] “Peking university international competition on ocular disease intelligent recognition (odir-2019),” https://odir2019.grandchallenge.org/
[4] Eyepacs, LLC. (n.d.). Eyepacs. Retrieved Feb 21,2017, from http://www.eyepacs.com/eyepacssystem/
[5] Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H. (2019): Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences 501, 511–522.
[6] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, Vasudevan Lakshminarayanan (2018), “Retinal fundus images for glaucoma analysis: the RIGA dataset”, Proc. SPIE 10579, Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, 105790B (6 March 2018); https://doi.org/10.1117/12.2293584
[7] Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunović, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, Xiulan Zhang. REFUGE: Retinal Fundus Glaucoma Challenge. IEEE Dataport. (2019) https://dx.doi.org/10.21227/tz6e-r977
[8] Sa'idah, Sofia; Magdalena, Rita; Nur Fuadah, Yunendah, 2022, “Immature Cataract Fundus Images”, https://doi.org/10.34820/FK2/CDWESA, Telkom University Dataverse, V1
[9] Tan, Mingxing, and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks.” International conference on machine learning. PMLR, 2019. https://doi.org/10.48550/arXiv.1905.11946
[10] Sandler, Mark, et al. “Mobilenetv2: Inverted residuals and linear bottlenecks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018 https://arxiv.org/abs/1801.04381
[11] Park & Baek (2022), AI-Based Cataract Detection Platform Development. Journal of Platform Technology, vol.10, no., pp.1725. DOI: 10.23023/JPT.2021.10.1.017
[12] Lee, H. D. & Kim, J. S. & Kwon, Y. H. & Kim, Y. K (2019), Classification of ROP Using Deep Learning. The Journal of Korean Institute of Information Technology (JKIIT), vol.17, no.10, pp. 17-24. DOI: 10.14801/jkiit.2019.17.10.