Anomaly Detection in Red Reflex Images Using Deep Learning Approaches

Setnipat Kriangsakdachai, Srisupa Palakvangsa Na Ayudhya, Worapan Kusakunniran, Wongchanok Devakula Na Ayudhya, Chayapon Chantrasagul, Rungsimun Manasboonpermpool, Kanchalika Sathianvichitr, Prapasson Sangsre and Thammanoon Surachatkumtonekul

November 4, 2022

Anomaly Detection in Red Reflex Images Using Deep Learning Approaches

[1]Setnipat Kriangsakdachai, [1,*]Srisupa Palakvangsa Na Ayudhya, [1]Worapan Kusakunniran, [1]Wongchanok Devakula Na Ayudhya,
[1]Chayapon Chantrasagul, [1]Rungsimun Manasboonpermpool, [2]Kanchalika Sathianvichitr, [2]Prapasson Sangsre,
[2]Thammanoon Surachatkumtonekul
[1]Faculty of Information and Communication Technology
[2]Department of Ophthalmology, Siriraj Hospital
Mahidol University, [1]Nakhon Pathom, [2]Bangkok, Thailand
Email: setnipat.kra@student.mahidol.ac.th, srisupa.pal@mahidol.edu, worapan.kun@mahidol.edu,
wongchanok.dev@student.mahidol.ac.th, chayapon.cha@student.mahidol.ac.th, rungsimun.man@student.mahidol.ac.th,
kanchalika030@gmail.com, prapasson_n@hotmail.com, si95thim@gmail.com

Amblyopia is a noteworthy disease in children leading to visual loss. This work focuses on creating a deep learning model for the detection of Amblyopia factors in patients wearing masks under the COVID-19 pandemic.

Keywords—Amblyopia, Red reflex, Deep learning, Mask R-CNN, MTCNN, Convolutional neural network

I. Introduction

In 2020, the Centers for Disease Control and Prevention reported that approximately 6.8% of people younger than 18 years old in the United States have an eye condition, with nearly 3% being blind or visual impaired, which is one of the most prevalent disabling conditions in children [1]. There are several common pediatric eye problems that can lead to visual loss, which is a serious issue, and one of the most common is Amblyopia [2, 3]. Research shows that 1–6% of children experience a monocular visual loss, and 2.9% of adults suffer a permanent visual loss caused by Amblyopia when they were young [3].

Amblyopia or Lazy eye is a disease that occurs due to insufficient visual stimulation of the brain during the process of visual development, which occurs around the first 6 – 9 years of life. Any inequality in the image received during this period will result in the progression of Amblyopia. There are several key factors, including Strabismus, Refractive Errors, and Cataracts, which cause both eyes to receive different focused images simultaneously and lead to Amblyopia, also known as Amblyopia factors. Amblyopia factors can start to occur between the ages of 0 – 6 years. Prevention can be done by helping children use both eyes by patching the preferred eye or eliminating the Amblyopia factors [2-4]. Prevention must happen during the visual development process, before the Amblyopia is developed, which makes early detection vital.

Even though early detection has been deemed a crucial process, it is not always easy since visual development occurs at an age before children can communicate well with their parents [4, 5]. In infants and preverbal children, the diagnosis focuses mainly on the fixation and asymmetry of both eyes. Several methods, including Corneal light reflex testing, red reflex testing, and cover tests are used for vision screening [5, 6]. Red reflex testing is usually the most viable option with the advantage of being suitable for newborns and mainly focusing on asymmetric eyes, which include all of the Amblyopia factors [2, 3, 6].

Red reflex testing, also known as the Bruckner test, is a vision screening method using bright light to illuminate the eyes from a distance of 1 meter [7]. The fundus reflects the color and is collected for examination to determine the fixation based on the intensity, brightness, and opacity of the color represented between both eyes. Several issues, including Strabismus and unequal refraction, which represent Amblyopia factors, can be determined by comparing the color reflex between both eyes [2, 5].

There are limitations to the red reflex testing procedure administered in hospitals due to lack of expertise and proper equipment. With the advancement of video capture devices, red reflex can be easily captured, and many studies have focused on how to automate this process in order to create a preliminary diagnosis tool for parents. The automated process usually consists of two main parts, which are pupil localization and Amblyopia classification. The pupil localization creates a region of interest where the red reflex occurs before the classification. Amblyopia classification categorizes the red reflex image into a preliminary diagnosis result.

The COVID-19 pandemic has introduced a new-normal lifestyle where many people still choose to wear a mask even when it is no longer required. Wearing a mask directly affects the automated process of pupil localization. This is because information features must be collected from the red reflex area but the mask reduces the accuracy of the localization.

Due to this obstruction, the model for Amblyopia classification is being introduced and implemented as the back end of the Amblyopia detection system for preliminary diagnoses during the new-normal lifestyle.

II. Related Work

Amblyopia detection and Amblyopia factors detection consist of two main parts, which are pupil localization and classification, and there have been several proposals to improve this procedure. In 2008, Jonathan et al. [8] introduced an artificial intelligence technique to automate the screening process using Automated video vision development assessment (AVVDA) to capture the pupil when foveating. After capturing the pupil area, many classification methods are applied to the red reflex images with the highest accuracy being 77% using a decision tree for classification. Over the years, many new techniques in image processing and classification have been introduced, one of the most known being Convolutional Neural Network (CNN).

* Corresponding Author

Bin Li et al. [9] proposed real-time eye detection using CNN with different trained networks to select the candidate's eye region and the center of the eye. The research highlights the potential of the CNN technique with image processing by using DenseNet201, which has an accuracy of 96.36%, sensitivity of 96.95%, and specificity of 95.84%. Chun et al. [10] introduced Amblyopia factors detection using a classification model using the architecture of ResNet-18 with 4 layers. The research used a dataset of participants with an average age of 4.32 years obtained from a smartphone and used 305 images to train and test the model. The result, after a 5-fold validation, yielded an average accuracy of 83.2% in the validation dataset and 81.6% in the test dataset. Murali et al. [11] proposed a method of pupil localization using Dlib to detect facial landmarks which can be used to calculate the pupil area. Combined with the Kanna algorithm, which is claimed by Murali et al. [11] as a deep learning technique and is involved in the classification process, the results provided an accuracy of 79.6%, sensitivity of 88.2% specificity of 75.6%, and F-Score of 73.2%. However, the accuracy of the data may be questionable since the researchers had a lower number of datasets, merely 54 optometry students aged between 18 – 23 years. Murali et al. [12] continued their research using the same algorithm but increased the number of datasets to 654 patients aged below 18. This yielded better results, with an accuracy of 90.8%, sensitivity of 83.6%, specificity of 94.5% and F-Score of 85.9%. Ma et al. [13] introduced an approach using OpenFace [14], which is a facial landmark detection library used for pupil localization. Ma's research uses both red reflex and Corneal light reflex to classify the Amblyopia risk factors.

Research on the pupil localization process usually calculates the position of the area from the facial landmark detection library. However, for patients wearing a mask, the facial landmark detection does not perform as well since parts of the landmarks are being covered by the mask.

For the classification process, many researchers have proposed different techniques, which usually involves state-of-the-art technology. Even if the techniques are different, the researchers usually use a pre-trained weight in order to reduce the issue of insufficient training images. However, the Convolution Neural Network (CNN) architecture has been developed further which means the classification process will continue to improve.

III. METHODOLOGY

A. Overview of the proposed framework

The proposed framework consists of three main parts, which are data preparation, pupil localization and Amblyopia classification, as shown in Fig. 1

The experiment was approved by the institutional review board of Siriraj Hospital, Mahidol University (certificate of approval number: 459/2563(IRB1)).

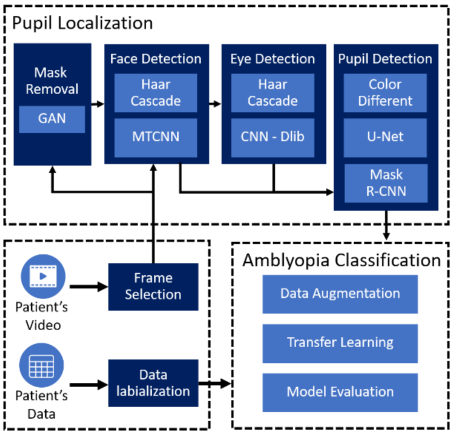

Fig. 1 Overview of the framework process

B. Dataset preparation

As part of the preparation, the characteristics of the data must be explored and determined in order to select a suitable method for arranging the data. The deep learning approach usually requires a large number of datasets in order to generate a good learning model, however, there are no public datasets available. Therefore, the dataset used in this research was provided by the partner hospital, which consists of videos and data collected from patients.

The patients' videos are in different lengths, with a resolution of 3,840*2,160 pixels. The process begins with measuring the distance, then turning on the lights (9–12 lux) so the patients can focus on the light source, and finally capturing the video with an iPhone X or iPhone 7, from a distance of 83–166 centimeters or 110–140 centimeters, respectively.

Preparing the patients' videos was challenging because the pupil size is very small after extraction, less than 100*100 pixels. This can negatively impact the accuracy of the classification process. Additionally, there is no specific frame where the red reflex appears in the video, which also posed a challenge. The sizing issue for the images was solved by using the interpolation technique and the frame with the red reflex in the video was selected manually. This issue was also solved by creating a RGB histogram in order to detect the frame where the red reflex occurs.

All of the data is collected from Thai people, who can be classified as Asian. Research from Li shows that ethnicity can affect the red reflex in pupil size and number of pigments [15].

The patients' data consists of raw information that must be labeled. A total of 322 pupil data samples were labeled since two patients only had one pupil data. The data was categorized into 57 normal cases and 265 abnormal cases. The data consists of Myopia diopters, Hyperopia diopters, Astigmatism diopters, and Cataracts, which were classified into abnormal cases when one of the diopters exceeded the threshold, since this represents an Amblyopia risk factor. The following thresholds were used to classify abnormal cases.

- Myopia diopters over 0.5
- Hyperopia diopters over 0.5
- Astigmatism diopters over 0.5
- Patient had a cataract

C. *Pupil localization*

Pupil localization focuses on image processing techniques and machine learning in order to extract the pupil area where the red reflex occurs. For this work, several approaches were utilized including Haar cascade, Facial landmark detection, Generative Adversarial Networks (GANs), Multi-Task Cascaded Convolutional Neural Networks and Mask Regional - Convolutional Neural Network (Mask R-CNN).

For Haar cascade and MTCNN, the pre-trained model was used, which is commonly used for face and eye detection. For facial landmark detection, Dlib was introduced, which is a library used to create 68 points of facial landmarks, including marks around the eyes.

U-Net and Mask R-CNN require labeled data. The mask labeling process was done manually using Photoshop to create a mask layer of pupil and VGG Image Annotator (VIA) to create a position mask file to use in Mask R-CNN, as shown in Fig. 2

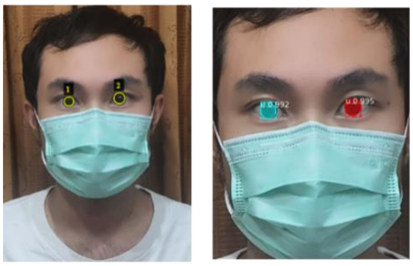

Fig. 2 Mask image and result using Mask R-CNN

The models of U-Net and Mask R-CNN were carried out using 20 training images to test the potential of both methods. The results indicate that Mask R-CNN performed better, which will be elaborated in the experiment section. The process is outlined in Fig. 3

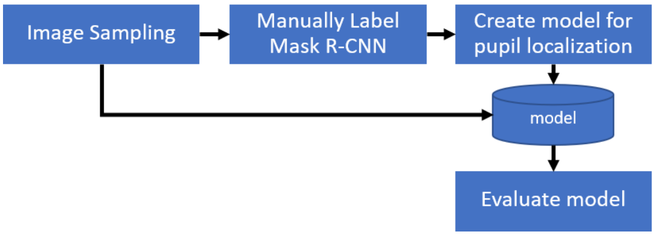
Image Sampling
Manually Label Mask R-CNN
Create model for pupil localization
model
Evaluate model

Fig. 3 Process of creating the Mask R-CNN model

D. *Amblyopia Classification*

The amblyopia classification process was developed by combining the pupil image from the pupil localization process with labeled data, which would result in a classification model that can be used on masked patients. In this work, densely connected convolutional networks (DenseNet) and EfficientNet are selected based on their state--of-the-art performance in their relevant domains. DenseNet121 was selected due to the compactness of the models with similar performance to ResNet [16]. On the other hand, EfficientNet was introduced as a model that scales all dimensions, including depth, width, height, and resolution aspects of the neural network, which directly affect the performance [17]. The process of Amblyopia classification is shown in Fig. 4.

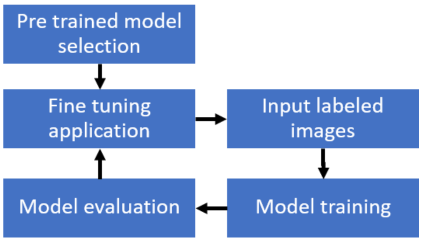

Fig. 4 Amblyopia classification process

Normally, a single aspect of the model will increase in order to improve its performance. However, EfficientNet introduced a new idea of expanding many aspects at once in order to improve the performance. This technique is currently being used by many researchers in this area due to its positive results. In this research, the part concerning EfficientNet is still in the experimental process and will continue to be fine-tuned for future implementation.

IV. Experiment & Results

This research consists of four experiments, with the purpose of generating a model of Amblyopia factors detection with data from patients wearing masks. The first experiment is on pupil localization by creating regions of interest and extracting a pupil until the last process of classification.

A. *Experiment #1: Face detection and eye detection with Haar cascade and facial landmark detection*

The experiment focuses on creating regions of interest for the pupil localization process to remove the environment background from the image processing. Two techniques were used in this experiment, which are Haar cascade and facial landmark detection.

Haar cascade was introduced as an object detection algorithm in 2001. The downside of Haar cascade is that the accuracy is lower compared to current techniques but since the algorithm does not require a lot of computational power it is the most suitable method for a device with lower computational power and low speed of the detection [18].

Apart from the Haar cascade, this experiment also compared results with Dlib, which is an ERT facial landmark detection algorithm. The ERT algorithm is a cascade template and its weakness is the quality of detection when compared to neural network techniques. However, the library and algorithm are still used by many researchers due to the speed of the detection compared to other methods [19, 20].

This experiment deployed Haar cascade as a face detection algorithm to create a region of interest from 162 patient images, which was then compared to the algorithm of the Haar cascade and Dlib in the eye detection process. The result and the example results are shown in TABLE I. , Fig. 5 and Fig. 6.

TABLE I. RESULT OF EXPERIMENT #1

Process	Method	Can detect	Correctly detect
Face detection	Haar Cascade	151	151
Eye detection	Haar Cascade	140	128
	Facial Landmark detection	151	110

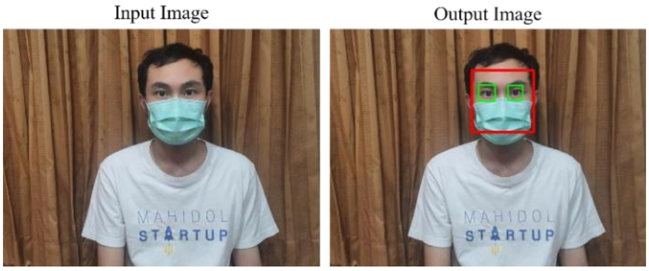

Fig. 5 Example results from experiment #1 with Haar cascade as an eye detection algorithm

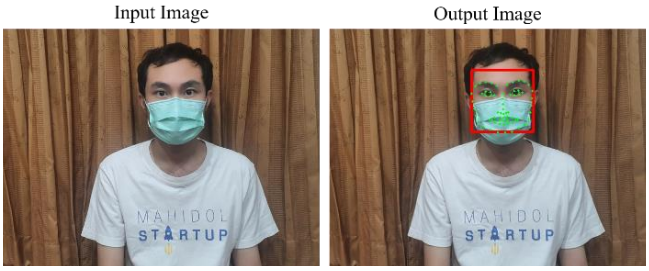

Fig. 6 Example results from experiment #1 with facial landmark detection as an eye detection algorithm

The experiment confirms that masks affect the pupil localization process, especially during the facial landmark detection algorithm since some parts cannot be detected and the position of some landmarks around the eyes are shifted out of position. Under this hypothesis, the experiment continues with the removal of the mask in order to enhance the accuracy of the facial landmark algorithm.

B. *Experiment #2: Mask removal using generative adversarial networks (GANs)*

Generative adversarial networks (GAN) are used specifically to solve image-to-image translation problem. GAN functions by learning the mapping between input and output images so it requires an image to train and learn from. After learning the features, the model will know which feature must be modified in order to create and output images.

In this experiment, images of the human face were used for the training model and an image of a mask was layered on top of the 994 images from LFW_DeepTunnel and tested with images of the patients. The example results from LFW_DeepTunnel are shown in Fig. 7.

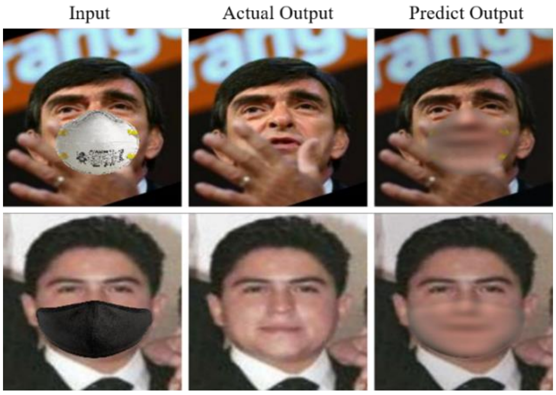

Fig. 7 Example results from experiment #2

Even though the result using LFW_DeepTunnel data is acceptable, experimenting with images of patients did not yield good results. This indicates that the GAN technique cannot be used with patients' images due to the brightness of the environment and the size of the image. Since GAN cannot be used for the experiment, pupil localization must be done with the mask on.

C. *Experiment #3: Pupil localization*

Since the mask removal process has proven to be impossible, this experiment will mainly focus on finding an approach that works while keeping the mask on. There are three methods in this process, which include facial landmark detection with pupil position calculation, U-Net and Mask R-CNN.

Based on the results of experiment 1, facial landmark detection may not be accurate in 100% of the cases, nevertheless, 67.90% was accurately detected. The process of the facial landmark detection can be combined with a position detection by calculating the color to localize the pupil area. However, since the accuracy of the facial landmark detection is quite low, the position of the pupil is often miscalculated rendering the experiment unacceptable. Example results of the detection are shown in Fig. 8

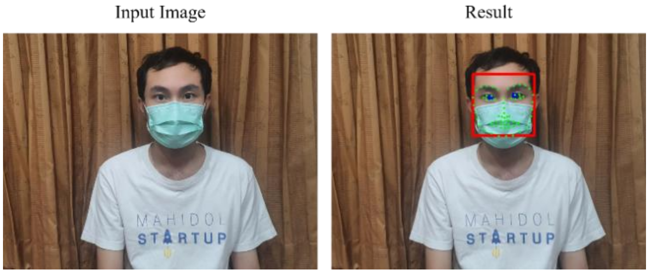

Fig. 8 Example results from experiment #3 with pupil position calculation

The next approach is U-Net, an architecture of neural networks that fit the image segmentation and classification domain, which is both fast and precise [21]. Since the training data of U-Net needs to be labeled manually, the number of the data training and test is lower than the previous approach.

The final approach is Mask R-CNN, which is a state-of-the-art image segmentation architecture. The architecture has an advantage in terms of performance, efficiency and flexibility. Results are shown in Fig. 9 [22, 23].

The data labeling process of Mask R-CNN and U-Net both require manual labeling so only 20 images were used as training data for both models. The number of training images can be increased if the result of the model reveals a potential to detect the pupil. After testing Mask R-CNN, results show that it can detect all of the test images, so the team decided to continue the process using this model with a full set of data.

Compared to all the other approaches, Mask R-CNN yielded the best results because the face was extracted from all the images. Multi-task cascaded convolutional networks (MTCNN) can extract the pupil with 78.09% accuracy as the confusion matrix shows in TABLE II. The example of results and the overall process are shown in Fig. 9 and Fig. 10

TABLE II. THE CONFUSION MATRIX OF EXPERIMENT #3

	Successful Segmented	Fail segmented
Abnormal	53	10
Normal	367	82
Total	420	92

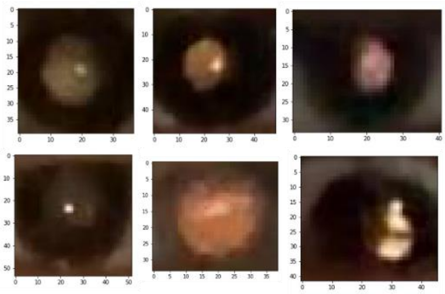

Fig. 9 Example results from experiment #3 with Mask R-CNN

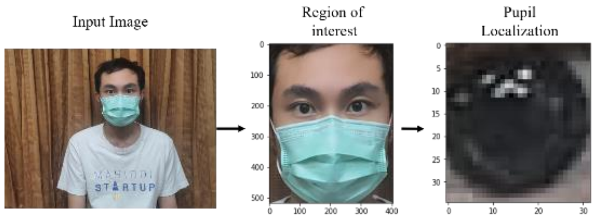

Fig. 10 Final process in experiment #3

D. *Experiment #4: Classification using DenseNet*

The combination of the normal class has fewer images than the abnormal class and this unequal dataset can lead to a biased model. An augmentation technique was used to sample pupil image into 1,000 images per class, which include flip, contrast and rotation images.

Subsequently, transfer learning was used in order to maximize the efficiency of the model. The data was separated into two segments with 80% for training and 20% for validation. The layer of the model consists of pooling layers, dense layers and a dropout on top of the pre-trained layer to reduce the complexity of the model.

TABLE III. THE EVALUATION TABLE FOR EXPERIMENT #4

	Accuracy		Loss	
Experiment 4.1: Base model	Train	Validate	Train	Validate
Option 1: Use of preprocess_input	0.7275	0.7525	0.5414	0.5288
Option 2: No use of preprocess_input	0.8669	0.8325	0.2847	0.4557
Experiment 4.2: Learning Rate				
Option 1: Learning Rate 0.01	0.6956	0.3650	0.8539	1.0241
Option 2: Learning Rate 0.001	0.7150	0.7125	0.5414	0.5395
Option 3: Learning Rate 0.0001	0.7275	0.7525	0.5414	0.5288
Option 4: Learning Rate 0.00001	0.7481	0.7550	0.5217	0.5053
Experiment 4.3: Image size				
Option 1: 32*32 pixels	0.7150	0.7125	0.5414	0.5395
Option 2: 64*64 pixels	0.6831	0.5725	0.6099	0.5987
Option 3: 128*128 pixels	0.6513	0.7075	0.6219	0.5693
Option 4: 224*224 pixels	0.5194	0.5000	0.7042	0.7098
Experiment 4.4: Optimizer				
Option 1: Adam	0.7987	0.8125	0.4242	0.3991
Option 2: SGD	0.7219	0.7525	0.5546	0.4910
Option 3: RMSDrop	0.7944	0.7575	0.4366	0.4898
Experiment 4.5: Non-trainable layers				
Option 1: 419 non-trainable layers	0.9525	0.4925	0.1141	0.9249
Option 2: 422 non-trainable layers	0.8456	0.8050	0.3497	0.4211
Option 3: 426 non-trainable layers	0.7987	0.8125	0.4242	0.3991

Based on TABLE III. the components for boosting the model performance can be concluded. The research uses a preprocess function, a learning rate of 0.001, an image size of 32*32 pixels, the Adam optimizer and 426 non-trainable layers. The results of accuracy and loss are shown in Fig. 11 and Fig. 12. The confusion matrix for the validating set with 200 images are shown in TABLE IV.

The model performance has an accuracy of 72.25%, sensitivity of 75.5% and specificity of 0.69%. The research proves that the model has a potential for Amblyopia classification.

TABLE IV. THE CONFUSION MATRIX USING MODEL FROM EXPERIMENT #4

		Actual Class	
		Abnormal	Normal
Predict Class	Abnormal	151	62
	Normal	49	138

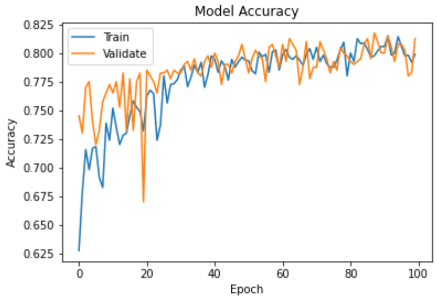

Fig. 11 Final model accuracy

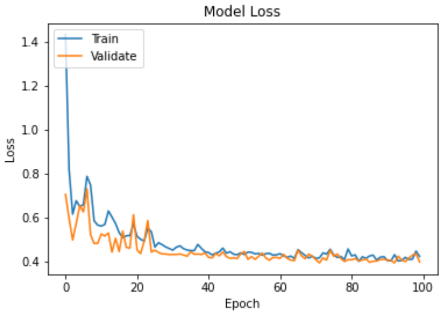

Fig. 12 Final model loss

V. CONCLUSION

Amblyopia is a disease in newborn children, which can lead to visual impairment. Early detection is crucial and can be done during the visual development process in order to prevent amblyopia. The research proposes a model to work as a back-end of the Amblyopia factors detection system, which can be used on patients wearing a mask to suit the new-normal lifestyle. The process consists of three parts, which are data preparation, pupil localization, and Amblyopia classification. The data preparation process includes the video frame capture, which was a challenge in that the image size is small, combined with data that must be labeled with the information of patients. In pupil localization, many approaches were tested, which include Haar cascade, CNN, U-Net, Mask R-CNN to create region of interest and scope down to the pupil area. The results of the experiment show that Mask R-CNN is the best approach, which worked on every image. The Amblyopia classification used DenseNet to test, and the several options were compared, which include the base model, learning rate, image size, optimizer and number of non-trainable layers, to determine the model with the best performance, which offered an accuracy of 72.25%, sensitivity of 75.5%, and specificity of 0.69%. The research confirms its potential of being a back-end model in the system and can be enhanced by improving the complexity of the model; testing the experiment with other CNN architecture, which includes the ongoing research regarding EfficientNet; and increasing the number of datasets to train the model.

REFERENCES

[1] Centers for Disease Control and Prevention. (2020, 26 June). *Fast Facts of Common Eye Disorders*. Available: https://www.cdc.gov/visionhealth/basics/ced/fastfacts.htm#:~:text=Approximately%2012%20million%20people%2040,due%20to%20uncorrected%20refractive%20error.

[2] Magramm Irene, "Amblyopia: etiology, detection, and treatment," vol. 13, no. 1, pp. 7-14, 1992.

[3] J. R. McConaghy and R. McGuirk, "Amblyopia: detection and treatment," vol. 100, no. 12, pp. 745-750, 2019.

[4] M. Pascual *et al.*, "Risk factors for amblyopia in the vision in preschoolers study," vol. 121, no. 3, pp. 622-629. e1, 2014.

[5] A. L. Bell, M. E. Rodes, and L. C. Kellar, "Childhood eye examination," vol. 88, no. 4, pp. 241-248, 2013.

[6] C. Kara and İ. S. Petriçli, "Comparison of photoscreening and autorefractive screening for the detection of amblyopia risk factors in children under 3 years of age," vol. 24, no. 1, pp. 20. e1-20. e8, 2020.

[7] J. M. Miller, H. L. Hall, J. E. Greivenkamp, and D. L. Guyton, "Quantification of the Brückner test for strabismus," vol. 36, no. 5, pp. 897-905, 1995.

[8] J. Van Eenwyk, A. Agah, J. Giangiacomo, and G. Cibis, "Artificial intelligence techniques for automatic screening of amblyogenic factors," vol. 106, p. 64, 2008.

[9] B. Li and H. Fu, "Real Time Eye Detector with Cascaded Convolutional Neural Networks," *Applied Computational Intelligence and Soft Computing*, vol. 2018, p. 1439312, 2018/04/22 2018.

[10] J. Chun *et al.*, "Deep learning–based prediction of refractive error using photorefraction images captured by a smartphone: model development and validation study," vol. 8, no. 5, p. e16225, 2020.

[11] K. Murali, V. Krishna, V. Krishna, and B. Kumari, "Application of deep learning and image processing analysis of photographs for amblyopia screening," vol. 68, no. 7, p. 1407, 2020.

[12] K. Murali *et al.*, "Effectiveness of Kanna photoscreener in detecting amblyopia risk factors," vol. 69, no. 8, p. 2045, 2021.

[13] S. Ma, Y. Guan, Y. Yuan, Y. Tai, and T. Wang, "A one-step, streamlined children's vision screening solution based on smartphone imaging for resource-limited areas: design and preliminary field evaluation," vol. 8, no. 7, p. e18226, 2020.

[14] Brandon Amos, Bartosz Ludwiczuk, and M. Satyanarayanan. (2016). *OpenFace*. Available: https://cmusatyalab.github.io/openface/

[15] Y. Li and D. Huang, "Pupil size and iris thickness difference between asians and caucasians measured by optical coherence tomography," *Investigative Ophthalmology Visual Science*, vol. 50, no. 13, pp. 5785-5785, 2009.

[16] C. Zhang *et al.*, "Resnet or densenet? introducing dense shortcuts to resnet," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3550-3559.

[17] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105-6114: PMLR.

[18] A. Rosebrock. (2021). *OpenCV Haar Cascades*. Available: https://pyimagesearch.com/2021/04/12/opencv-haar-cascades/

[19] D. Zhang, J. Li, and Z. Shan, "Implementation of Dlib Deep Learning Face Recognition Technology," in *2020 International Conference on Robots & Intelligent System (ICRIS)*, 2020, pp. 88-91: IEEE.

[20] K. Khabarlak and L. Koriashkina, "Fast facial landmark detection and applications: A survey," 2021.

[21] A.-L.-U. Freiburg. (27 March). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Available: https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.

[23] E. Odemakinde. (27 MArch). *Mask R-CNN: A Beginner's Guide*. Available: https://viso.ai/deep-learning/mask-r-cnn/#:~:text=Mask%20R%2DCNN%20was%20built,that%20outputs%20the%20object%20mask.