STATS 402 - Interdisciplinary Data Analysis
〈Anomaly Detection in Fundus Photographs〉

<Yufei Dong>
<Ziying Ye>

Abstract

How to detect anomalies is a broad concern in ophthalmology, and anomaly detection is often used to identify unexpected items such as lesions and deformities in fundus images or OCT [1]. In our project, we intend to detect eye diseases in fundus photographs, namely, to distinguish anomalous images from healthy eye images. This problem can be addressed as a classification problem which can be resolved through supervised and unsupervised learning models. In our project, we selected CNN as a representative for the supervised learning model, and GAN for the unsupervised learning model to perform anomaly detection on the dataset collected by Tsukazaki Hospital in Japan. For the CNN model, the probability outcome was generated when multiple diseases are detected to see whether it can successfully detect different diseases. We compared the efficiency and performance of models, where CNN models achieved over 80% accuracy generally, and the GAN achieved a high accuracy but low AUC score.

1. Introduction

Our eyes are significant as a sense organ since we perceive approximately 80% of the environment through vision [2]. However, according to World Health Organization (WHO), vision impairment affects over 2.2 billion people worldwide, and among those impairment cases, almost half of them are either preventable or have not yet been coped with [3]. Moreover, as reported by WHO, the most prevalent causes of eye impairment are age-related macular degeneration, cataract, diabetic retinopathy, glaucoma, and uncorrected refractive errors. Therefore, taking care of our eyes and paying attention to eye and retina diseases is vital in our daily life.

To preserve the health of our eyes, routine eye exams are recommended to find eye diseases and maintain one's vision, according to CDC [4]. If a person has experienced eye trauma, is suspected of certain eye diseases, or has an eye disease history and needs to track the progress of the disease, taking a fundus photograph is medically necessary and highly recommended. The fundus is the inside back surface of the eyes, and fundus imaging refers to a color picture of the back of the eye taken by a special fundus camera through a dilated pupil [5]. Finding the anomaly, which is different from the normal eye pictures and can imply potential diseases, is vital in retinopathy diagnosis.

Thus, the aim of the project is to perform classification tasks on fundus photographs of healthy or diseased eyes. Machine learning techniques, including GAN and CNN, were utilized to conduct anomaly detection and classification in fundus images. Evaluations of the supervised (CNN) and unsupervised (GAN) learning approaches were executed to compare their accuracy and efficiency.

In this paper, the organizations are presented as below. The second part will discuss the related research in the area of anomaly detection in medical images. In the proposed methods, we will introduce the methodologies (including dataset selection, data processing, models and algorithms implementation) adopted in this project in detail. The fourth part will demonstrate the performance evaluation of the models we implemented. For the last part conclusion and future work, the project results will be summarized, and potential future improvements will be proposed.

2. Related work/literature review

Identifying and diagnosing diseases based on medical images like fundus photographs and CT mainly depends on doctors' or technicians' manual checks [6]. The traditional diagnosis method, which is still prevalent today for its high accuracy, largely relies on the experience of doctors. Still, it may be of low efficiency when encountering a large number of medical images. Therefore, machine learning techniques are introduced into this industry to tackle problems like massive image classification to share the burden of doctors. Supervised learning models like Decision tree and SVM are simple and widely used, though their performance may not be ideal [7]. Afterward, more complex deep learning models like CNN are applied in this industry, which can achieve relatively higher accuracy [8].

However, supervised learning models tend to perform well on balanced and well-labeled data. In real-life scenarios, anomalous images are significantly fewer, leading to the imbalance of the dataset [9]. Moreover, there is no clear definition of an anomaly in medical images. Even the same diseases' anomaly photos may differ significantly, let alone some rare or unknown illnesses. To resolve those difficulties, a weakly-supervised learning method was implemented [10]. This technique would train the model with normal images only and would be fed with some abnormal images later from time to time. By doing so, the model's performance can be greatly enhanced. In recent years, Generative Adversarial network (GAN), a fully unsupervised learning method, has also been introduced into medical image analysis. The training set for GAN only consists of normal images. The GAN model can generate new images based on the training set, which is

considered as a learned norm. When processing the abnormal image in the testing set, the anomaly will deviate significantly from the learned norm, thus being detected as an anomaly. This method has higher generalizability and can be utilized in a wider variety of anomaly detection problems. As the most used technique in eye disease detection, in previous work, researchers have tested the performance of commonly used GAN models like GANomaly, Auto Encoder, and AnoGAN models [11], [12].

In our work, we applied both supervised models like CNN and unsupervised learning models GAN to compare their performances. We aim to figure out which model outperforms in anomaly detection on this dataset. Besides the outcome of detecting the presence of disease, we also want to investigate whether the two models have divergent performances on various diseases. In our method, the probability of each disease was predicted independently.

3. The proposed method

The goal of this project is to detect anomalies in fundus photographs. Here we will break the methodology into the following three parts, the EDA on the dataset, data preprocessing, and model implementation.

3.1 EDA:

The dataset used in this project was collected by Tsukazaki Hospital in Japan, including 13,047 normal and abnormal fundus images in total [13]. The disease includes diabetic retinopathy, glaucoma, age-related macular degeneration, retinal vein occlusion, macular hole, retinal detachment, retinitis pigmentosa, artery occlusion, and Diabetic Macular Edema respectively abbreviated as DR, Gla, AMD, RVO, MH, RD, RP, AO, DM in the dataset. The dataset has no NA value; all pictures have labels, including id, left or right eyes, age, sex, and the specific disease(s). Fig. 1 displays the distribution of age and sex of the patients in the dataset. Given that eye diseases are more prevalent among the elderly, the age of the patients displayed an intuitive left-skewed distribution with a mean of 65.1 and a median of 67, in which the most populated age ranges are 61-79 and 71-80. The sex of the patients is approximately balanced, with males slightly more than females. Fig. 2 shows the distribution of eye diseases in the dataset. The dataset contains 4000+ healthy eye images, which is of the highest quantity among all categories. The following disease categories are DM, DR, and Gla. The remaining disease categories have much fewer images contained in this dataset, indicating the imbalance of the data set. Fig. 3 is a list of sample fundus images.

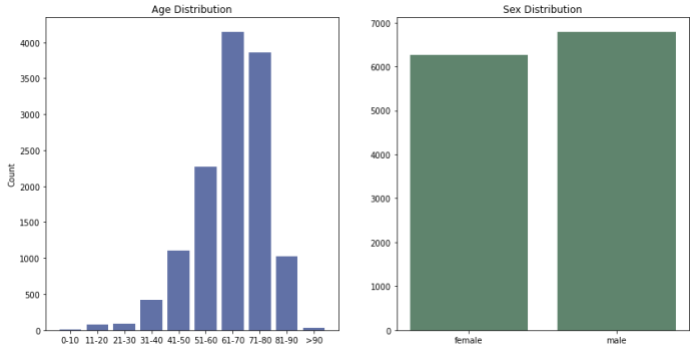

Fig. 1. Distribution of age and sex

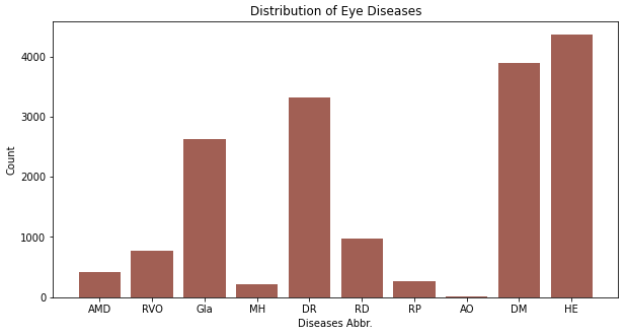

Fig. 2. Distribution of various eye diseases in this dataset

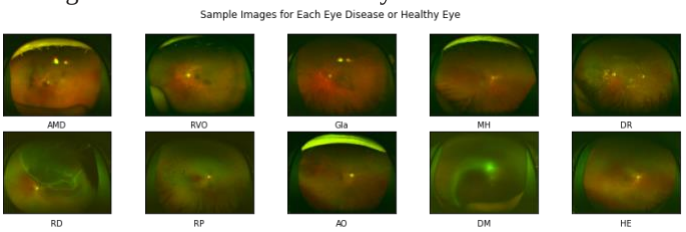

Fig. 3. Sample images of different eye disease categories or healthy eyes from the dataset

3.2 Data preprocessing

For preprocessing, we performed image resizing and train-test split for our dataset.

Normalization and Resizing:

We performed RGB normalization using the Normalize() transform in the GAN model. This process normalized an image with mean and standard deviation. Here we used the mean and standard deviation from ImageNet, with mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225] [14]. The formula applies the normalization: img = (img - mean * max_pixel_value) / (std * max_pixel_value), where max_pixel_value is the maximum value on the pixel range of [0, 255].

Image resizing was implemented for both GAN and CNN models. The original images have 768 rows, 1024 columns, and 3 channels if read in array format. Such images may be too large and may take too much time to train the model. To run the model efficiently, we resized the images to 256*256 for the GAN, the Inception V3, and the ResNet model. For the simple CNN model, we input the original size to train the model.

Training, validation, testing set split:

All fundus photographs are mingled together as the original dataset, along with a CSV file with the label marking each image as healthy, or the specific disease(s) found in the image.

Since GAN models take a long time to train, we only trained the model with 2608 images, which contains 872 healthy eye images and 1736 anomalous eye images. Since GAN models only need normal images to train, all 1736 anomalous eye images were categorized into the test set. Then we applied a train-test split on healthy eye images with a split ratio of 0.2. The validation test is created within the test set. Also, we set both the training batch size and test batch size to 32.

For CNN, we used all 13,047 images. We have done two kinds of categorization. We split the dataset into normal and abnormal fundus images. We also classified fundus images into different folders according to the specific diseases they belong to. When performing various classes classification, we can

directly extract the corresponding images from the disease folder. Then for both categorizations, we perform the train-test split. We split the dataset into training and testing sets with a split ratio of 0.2, and split the training set into training and validation sets with a split ratio of 0.25. The batch size is either 32 or 16 based on the performance of the models.

3.3 Model implementation:

For the GAN model, we merged all disease classes into a single category of abnormal eye images and performed the binary classification task to identify abnormal fundus images from normal images. For CNN, we experimented with three different models, the simple CNN, Inception V3, and ResNet. We performed distinct classification tasks with the three models. For some models, cross-validation was also utilized.

GAN:

We intended to apply a binary classification task with Generative Adversarial Network to distinguish anomalous eyes from normal eyes. GANomaly follows the common logic of GAN methods to train the generator to generate new images based on the input images. The model has an encoder-decoder-encoder architecture. The generator learns how to encode features of the generated image for normal samples, which means that the model will report a low loss when inputting a normal image. However, when an anomalous image is encoded, the reconstruction in the model will fail to function properly, and thus can be recognized by the model [15]. GANomaly uses this architecture to achieve anomaly detection, while at the same time, the loss enables the model to carry out a self-supervision. The loss consists of three parts, the encoder loss, the adversarial loss, and the contextual loss. The adversarial loss indicates the difference between the feature representation of the original images with the generated images. While with the assistance of contextual loss, the model can learn more contextual information and thus be optimized. The encoder loss represents the distances between the encoded features of the input and generated images [15]. Each loss is assigned a weight to adjust the impact that each loss contributes to the total loss. In our project, the weight remained as default.

CNN:

In addition to GAN, we also proposed using different CNN models for performing various classification tasks on fundus images. Basically, we used three kinds of CNN models, which are the simple CNN model, the Inception V3, and the ResNet. The simple CNN model originated from the dataset, while the other two models are utilized in previous studies, and had achieved reliable results [16]. The code for inception V3 and ResNet were from the Keras package. Flatten layers and dense layers were added at the end of the Inception V3 and ResNet model. All other layers were not frozen for tuning the supervised learning model. This allowed the model to have plenty of trainable parameters, helping the model achieve reasonable predictions. Table one shown below displays various fundus image classification tasks, the models applied for the corresponding tasks, and whether cross-validation was implemented. All CNN models can generate the probability outcome.

Classification Tasks	CNN Models Applied	Cross Validation
Normal and Abnormal	Simple, Inception V3, ResNet	No
Normal, Glaucoma, Diabetic Retinopathy	Inception V3, ResNet	Yes
10 classes, normal and all fundus diseases	ResNet	No
Normal, Each Separate Disease (with fewer images)	ResNet	Yes
5 classes, diseases with fewer images	ResNet	Yes

Table 1. List of classification tasks for CNN models

We first implemented normal and abnormal on fundus image classification by using simple Inception V3 and ResNet CNN model. Then, since we want to figure out whether CNN can distinguish between different retinopathy, and the dataset is imbalanced as discussed before, we selected the disease categories with more than 1,000 images to perform the classification. Moreover, we tried to implement the classification for all the disease classes and normal fundus images using ResNet. To ensure the CNN model can learn the feature of various retinopathy, we also performed the classification for each separate disease with fewer images (fewer than 1,000 images) and the corresponding amount of normal fundus images. Additionally, to solve the case that some images may not be classified well in the three categories classification model, we created the model which aims to classify the five retinopathy with fewer images with ResNet to see whether the image would belong to other specific categories.

4. Performance evaluation

GAN:

After preprocessing the data, due to GAN's long runtime and large GPU consumption, we ran the GANomaly model on the part of our initial dataset, which contains 1736 anomalous eye images and 872 normal eye images. Since the GANomaly model is trained only with normal images, we fed the healthy eye images into the model. After running five epochs, we achieved an accuracy score of 0.908, an F1 score of 0.95, but an Area Under Curve (AUC) score of 0.46. The unexpectedly low AUC score made us wonder why there was an inconsistency between two evaluation methods. Therefore, we compared our results with those of the authors on another dataset [17]. We found a similar pattern in their result, where the AUC score for

most categories was around 0.5, but the accuracy was over 0.9. Since the imbalance of the dataset may cause a high accuracy score but a low AUC score, we examined the composition of our dataset [18]. We discovered that our test set contained 1736 anomalous eye images but only 174 normal eye images (20% of the 872 normal eye images were split into the test set), which indicates that our model could also reach a 90% accuracy by classifying most image (nearly every) as anomalous. To test our hypothesis, we reduced the number of anomalous eye images to 866 while the number of normal images remained the same. The model returned an accuracy of 0.83 and an AUC below 0.5, verifying our hypothesis that the model classified almost every test image as anomalous. This indicated that the model often raised type I error, a false alarm of categorizing normal eyes as anomalous eyes. The overestimation of this model and extreme imbalance in the test set also explained the high F1 score. The calculation of the F1 score is given by

$$F1 = 2 \times \frac{PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

The F1 score is the harmonic mean of precision and recall, with lower FP and FN resulting in a higher F1 score [19]. In our case, both FP and FN only accounted for a small number in the test set due to the uneven distribution of healthy images and anomalous images. We discovered that the author's dataset also had an imbalance in the test set, with a significantly lower number of normal images than anomalous images, which explained why they got a similar pattern [20].

Test metric	DataLoader 0
image_AUROC	0.4554399251937866
image_Accuracy	0.9089004993438721
image_F1Score	0.9522764682769775

```
[{'image_F1Score': 0.9522764682769775,
  'image_AUROC': 0.4554399251937866,
  'image_Accuracy': 0.9089004993438721}]
```

Fig. 4. Results after five epochs

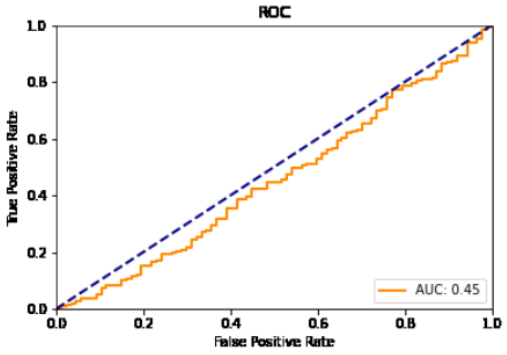

Fig. 5. The ROC after five epochs

We wonder whether having poor performance in recognizing normal images is a common problem in GAN models when performing classification tasks. Therefore, we generated one sample from normal pictures for testing. The image is labeled as anomalous with a 25% confidence score, which is contrary to our expectation, indicating that the error may occur in the classification threshold. In our previous work, the model ran with an adaptive threshold which is automatically generated based on the input dataset. However, when we tried to adjust the threshold, the model failed to recognize the manually set threshold.

We tried to fix the problem through tuning parameters like image size and batch size, but the model did not show significant improvements. We wonder whether the bad performance is related to the fundus images or the encoder-decoder structure. We attempted to output the generated image to see whether the model was functioning properly in the reconstruction process. However, the images are encoded and decoded in the form of latent vectors. Therefore, we could not evaluate the model's credibility by comparing the generated image and the original image.

Therefore, to see whether other autoencoder models can successfully generate images, we implemented Variational Autoencoder to test the image reconstruction process [21]. We ran the VAE model for 100 epochs to see the reconstructed images, which are shown in Fig.6. The generated images displayed hazy and blurry images, which eliminated many important eye structures and features to make the diagnosis.

In terms of why the model has such bad performance on our dataset, one possible assumption is that the learning of the model with an encoder-decoder structure largely depends on identifying the shape and color of the object. However, our images have vague edges and low color contrast among pixels, which may hinder the learning process. Because the majority of the diseases in our dataset have very slight differences from normal eyes, the model may struggle to find detecting regions with such blurry images.

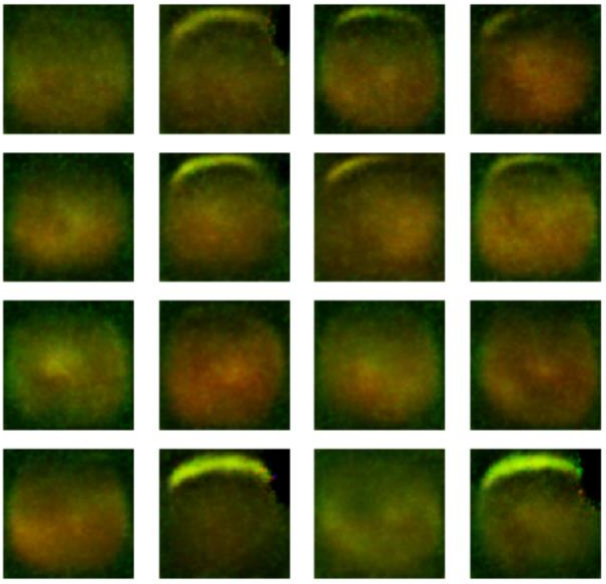

Fig.6. Samples of reconstructed images

CNN:

First, we applied both the simple model, the Inception V3, and the ResNet model to classify normal and abnormal fundus images, where normal images refer to healthy fundus photographs, and abnormal images refer to fundus photographs with retinopathy.

Fig. 7.1, 7.2, and 7.3 display the training and validation accuracy for simple CNN, Inception V3, and the ResNet models of normal and abnormal fundus image classification. We can see

that the ResNet model in Fig. 7.3 and the Inception V3 model in Fig. 7.2 attained over 90% training accuracy and approximately 80% validation accuracy. While for the simple model, both training and validation accuracy were slightly over 70%.

Overall, the Inception V3 and ResNet models had similar performance, and their performance was better than the simple CNN model. These results indicate that the CNN model is qualified for performing fundus classification tasks, and the Inception V3 and ResNet can be more suitable for further categorization tasks.

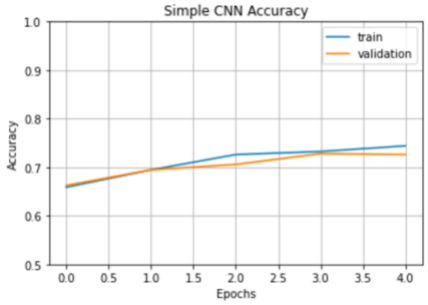

Fig. 7.1 Accuracy for Normal and abnormal fundus image classification with simple CNN model.

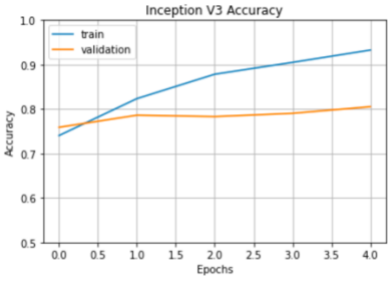

Fig. 7.2 Accuracy for Normal and abnormal fundus image classification with the Inception V3 model.

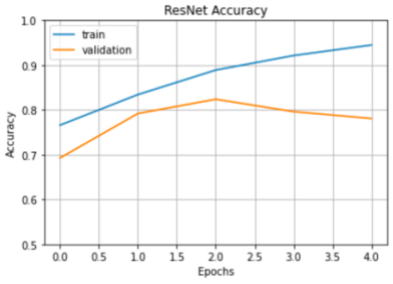

Fig. 7.3 Accuracy for Normal and abnormal fundus image classification with ResNet model.

The second classification task is distinguishing healthy, glaucomatous fundus images or pictures with diabetic retinopathy. Based on the result of the first task, we applied the Inception V3 and the ResNet model. Fig. 8.1 and 8.2 display the outcome of the model without cross-validation. The training accuracies were over 90% for both models, and the validation accuracy was slightly lower than 80% for Inception V3 and marginally higher than 80% for ResNet. Generally, the training and validation accuracy of the ResNet was better than the Inception V3 model.

Based on Fig. 8.1 and 8.2, we can find that the difference between training and validation accuracy is fairly large for both models, over 10% for Inception V3 and around 20% for ResNet. These two points indicate the potential problem of overfitting. Thus, we applied cross-validation to tackle this issue.

The bar plot of different cross-validation sets' accuracy in Fig. 9 was generated by the Inception V3 model. Fig. 10 exhibits the validation accuracy of different cross-validation sets of the ResNet model for the three categories classification. Both of them show that there were cross-validation sets with validation accuracy more than 90%, which is close to the training accuracy. This result reflects the problem of overfitting for CNN models and demonstrates the necessity of performing cross-validation to solve the problem. It also reveals that Inception V3 and ResNet can have high accuracy in performing the fundus photographs classification.

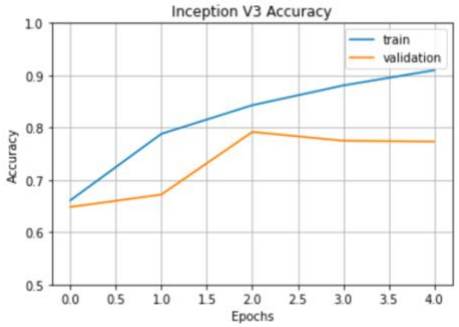

Fig. 8.1 Accuracy for fundus images that are healthy, with glaucoma, or diabetic retinopathy classification with the Inception V3 model.

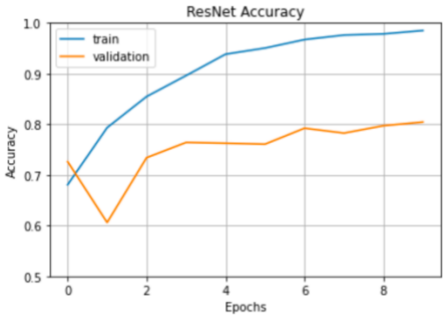

Fig. 8.2 Accuracy for fundus images that are healthy, with glaucoma, or diabetic retinopathy classification with ResNet model.

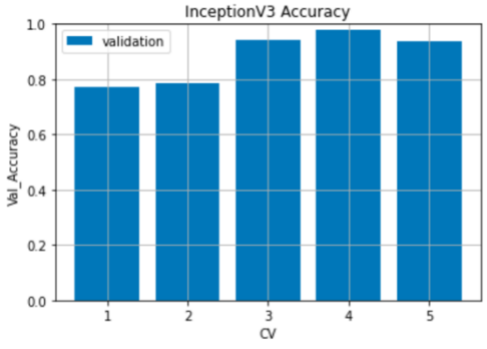

Fig.9. Cross-validation Accuracy for fundus images that are healthy, with glaucoma, or diabetic retinopathy classification with the Inception V3 model.

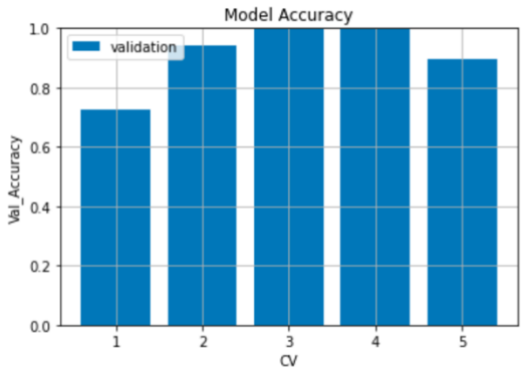

Fig. 10. Cross-validation Accuracy for fundus images that are healthy, with glaucoma, or diabetic retinopathy classification with the ResNet model.

As stated before, we also implemented the ResNet CNN model to perform classification for all nine fundus diseases and healthy fundus images contained in the dataset. Fig. 11 showed that the training accuracy kept increasing and reached over 90% at the last epoch, whereas the validation accuracy fluctuated around 50%.

This low validation accuracy might be due to several problems. One is still the problem of overfitting. As the training accuracy increased through the training process, the gap between training and validation accuracy became huge. The other possible reason is that the dataset is highly imbalanced. In addition to the photographs belonging to normal, glaucoma, diabetic retinopathy, and Diabetic Macular Edema classes, which have over 2000 images, the fundus photographs belonging to other groups are all fewer than 1000 pieces. Moreover, since in certain categories, the number of images is small (like artery occlusion which only has 21 images), we wonder whether CNN can succeed in learning to classify the diseases. Additionally, some disease categories have considerable overlapping. For instance, almost all the fundus images with diabetic retinopathy in the dataset also have diabetic macular edema. This can also confuse the model since it may not be able to spot the discrepancy between these two categories.

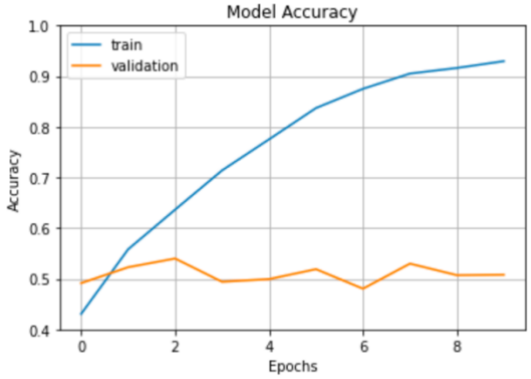

Fig. 11. Accuracy for classification of all fundus diseases of ResNet

We also applied separate models to classify each retinopathy and healthy fundus to determine whether the ResNet CNN model can learn to categorize the specific disease and normal fundus photographs. Fig. 12 displays the test accuracy for different fundus diseases using the ResNet model with cross-validation. We can conclude that for diseases other than artery occlusion, CNN can successfully differentiate between the specific illness and normal fundus.

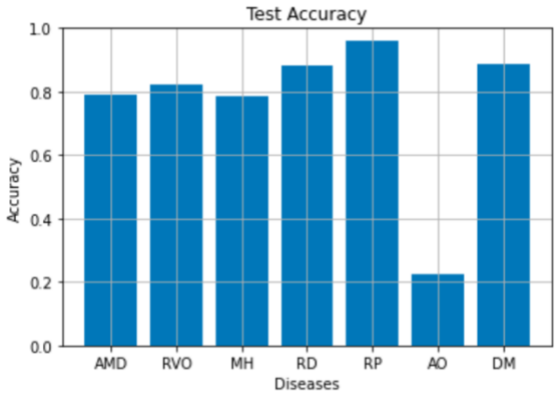

Fig. 12 Test accuracy after cross-validation for the classification of each separate disease (not included in 3 class classification) using the ResNet

Last, for five disease categories (AMD, RVO, MH, RD, RP), we carried out classification using ResNet and cross-validation as well. Through training, we found that the training accuracy was high, which is even close to 100%. However, the validation accuracy varies tremendously. Fig. 13 displays the validation accuracy for the 4 different cross-validation sets. The highest accuracy is over 80% while the lowest is only around 20%. This indicates that the classification highly depends on the images since the images were randomly distributed to different cross-validation sets.

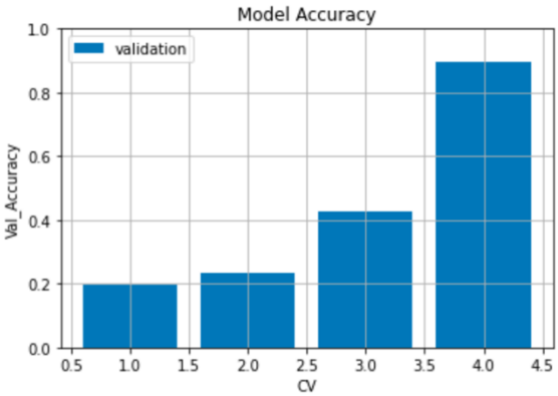

Fig.13 Cross-validation Accuracy for fundus images with AMD, RVO, MH, RD, RP using the ResNet model.

5. Conclusion and future work

We achieved a result of high accuracy but low AUC score using GAN for binary classification, indicating a bad performance on our dataset. The possible assumption of this result is that the structure of the GAN model cannot effectively classify fundus images without clear edges and high color contrast between pixels.

The CNN models had distinct performances in different classification tasks. For the categorization of normal and abnormal images, classification of normal images, photographs with DR or Gla, differentiation of each separate retinopathy, and normal fundus pictures, the CNN models obtained positive outcomes. However, for the classification of all retinal diseases and normal fundus, and categorization of the five diseases with fewer images, the results were not ideal.

For future work, the imbalance of the dataset can also be solved through data augmentation through more advanced GAN that can generate pictures with higher resolution and better quality [22]. Then, we can apply the CNN model again to see whether it can learn to classify among those illnesses with fewer images. Moreover, if we are able to obtain well-labeled groundtruth image sets, lesion localization can be performed by the GAN model. In addition, with a more advanced GPU and larger RAM, the GAN model can be trained with more epochs and larger image sizes to see whether the low AUC score problem can be solved.

References

[1] Y. Han et al., "Application of an Anomaly Detection Model to Screen for Ocular Diseases Using Color Retinal Fundus Images: Design and Evaluation Study," PubMed Central (PMC), Jul. 13, 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8317033/ (accessed Nov. 12, 2022).

[2] V. Wolf, "Medical Eye Center | Importance of Eye Care | Medford," Medical Eye Center, Jun. 20, 2016. https://www.medicaleyecenter.com/2016/06/20/importance-eye-care/ (accessed Nov. 12, 2022).

[3] "Vision impairment and blindness," Blindness and vision impairment, Oct. 13, 2022. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed Nov. 12, 2022).

[4] CDC, "Keep an Eye on Your Vision Health," Centers for Disease Control and Prevention, Oct. 01, 2020. https://www.cdc.gov/visionhealth/resources/features/keep-eye-on-vision-health.html (accessed Nov. 12, 2022).

[5] "Color Fundus Photography | Department of Ophthalmology," Color Fundus Photography | Department of Ophthalmology. https://ophthalmology.med.ubc.ca/patient-care/ophthalmic-photography/color-fundus-photography/ (accessed Nov. 12, 2022).

[6] A. Boyle et al., "Can middle grade and consultant emergency physicians accurately interpret computed tomography scans performed for head trauma? Cross-sectional study," Emergency Medicine Journal, Jul. 22, 2009. https://emj.bmj.com/content/26/8/583 (accessed Nov. 12, 2022).

[7] C. L. Chowdhary, M. Mittal, K. P., P. A. Pattanaik, and Z. Marszalek, "An Efficient Segmentation and Classification System in Medical Images Using Intuitionist Possibilistic Fuzzy C-Mean Clustering and Fuzzy SVM Algorithm," MDPI, Jul. 13, 2020. https://www.mdpi.com/1424-8220/20/14/3903 (accessed Nov. 12, 2022).

[8] Cen, LP., Ji, J., Lin, JW. et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nat Commun 12, Aug. 10, 2021. https://doi.org/10.1038/s41467-021-25138-w

[9] A. Cook, "Global Average Pooling Layers for Object Localization," Global Average Pooling Layers for Object Localization, Apr. 09, 2017. https://towardsdatascience.com/anomaly-detection-in-images-777534980aeb(accessed Nov. 12, 2022).

[10] T. Finck et al., "Faster and Better: How Anomaly Detection Can Accelerate and Improve Reporting of Head Computed Tomography," PubMed Central (PMC), Feb. 10, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8871235/ (accessed Nov. 12, 2022).

[11] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised Anomaly Detection via Adversarial Training," GANomaly: Semi-supervised Anomaly Detection via Adversarial Training | SpringerLink, May 29, 2019. https://link.springer.com/chapter/10.1007/978-3-030-20893-6_39 (accessed Nov. 12, 2022).

[12] K. Zhou et al., "Encoding Structure-Texture Relation with P-Net for Anomaly Detection in Retinal Images," Encoding Structure-Texture Relation with P-Net for Anomaly Detection in Retinal Images | SpringerLink, Nov. 12, 2020. https://link.springer.com/chapter/10.1007/978-3-030-58565-5_22 (accessed Nov. 12, 2022).

[13] DateCazuki, "GitHub - DateCazuki/Fundus_Diagnosis: ツカザキ病院から提供されている眼底画像データセットを用いた分類機/Classifier using fundus image data set provided by Tsukazaki Hospital.," GitHub. https://github.com/DateCazuki/Fundus_Diagnosis (accessed Nov. 12, 2022).

[14] S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, and U. Genc, "Anomalib: A Deep Learning Library for Anomaly Detection," GitHub, Feb. 01, 2022. https://github.com/openvinotoolkit/anomalib (accessed Nov. 26, 2022).

[15] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training," arXiv:1805.06725 [cs], Nov. 2018, [Online]. Available: https://arxiv.org/abs/1805.06725

[16] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning," PubMed Central (PMC), May 26, 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8198489/ (accessed Dec. 15, 2022).

[17] S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, and U. Genc, "Anomalib: A Deep Learning Library for Anomaly Detection," GitHub, Feb. 01, 2022. https://github.com/openvinotoolkit/anomalib (accessed Nov. 26, 2022).

[18] R. Rajamani, "High on accuracy but low on ROC score?," Medium, Mar. 16, 2021. https://ranjani-rajamani.medium.com/high-on-accuracy-but-low-on-roc-score-a40f2053b6c4

[19] J. Korstanje, "The F1 score," Medium, Aug. 31, 2021. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

[20] P. Bergmann, K. Batzner, M. Fauser, D. Sattelager, and C. Stegar, "The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," www.mvtec.com, Jan. 06, 2021. https://www.mvtec.com/company/research/datasets/mvtec-ad (accessed Nov. 26, 2022).

[21] A. Mousavi, "Anomaly Detection using Variational autoencoder," GitHub, Nov. 29, 2022. https://github.com/amousavi9/Anomaly-Detection-using-VAE (accessed Dec. 15, 2022).

[22] S. Motamed, P. Rogalla, and F. Khalvati, "Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images," Informatics in Medicine Unlocked, vol. 27, p. 100779, 2021, doi: 10.1016/j.imu.2021.100779.