



# Hierarchical Machine Learning Models for Ophthalmologic Disease Detection and Classification

An, Guangzhou

---

(Degree)

博士 (工学)

(Date of Degree)

2020-09-25

(Date of Publication)

2022-09-25

(Resource Type)

doctoral thesis

(Report Number)

甲第7885号

(URL)

<https://hdl.handle.net/20.500.14094/D1007885>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



# Doctoral Dissertation

## Hierarchical Machine Learning Models for Ophthalmologic Disease Detection and Classification

階層的機械学習モデルによる眼疾患の検出及び分類

July 2020

Graduate School of System Informatics

Kobe University

AN GUANGZHOU

安 光州

# Abstract

Aging population is causing a steep increase in number of people with vision impairment. Selecting proper treatments at the early stage is important to prevent vision loss. Conventional diagnose methods tremendously depend on the doctors' experience to detect and classify diseases to determine the treatment plan. Recently, there are lots of researches and social implementations trying to assist the clinical decision making via machine learning. Machine learning especially deep learning requires a large amount of labelled data, and the more data they are provided, the better the machine learning models typically perform. However, the amount of labelled data in medical field is limited.

The major theme of this dissertation is focused on building high accurate machine learning models for ophthalmologic disease detection and classification with small-size dataset. To achieve this goal, the doctors' diagnostic processes are analyzed first as follows. First, doctors perform disease classification after discriminating diseased cases from healthy cases (disease detection). Second, doctors classify diseases into subcategories by reusing the knowledge of classifying healthy and diseased cases. Third, doctors use multiple information to make an optimal diagnosis.

Based on the characteristic of doctors' diagnostic processes analyzed, a framework for building machine learning models of ophthalmologic disease detection and classification is proposed and implemented with newly created training methods. First, hierarchical classification method of two-step is used to build the machine learning models for disease classification, after building the models for disease detection. Second, hierarchy transfer learning method is used to build the machine learning model for disease classification by reusing parameters of the disease detection model in the hierarchical model above. Third, stacking ensemble method is used to extend the two-step framework to handle multiple input data by combining the machine learning models trained separately on single input data with the hierarchical classification method and hierarchy transfer learning method.

A series of experiments are performed to demonstrate the proposed framework for training machine learning models of ophthalmologic disease detection and classification in achieving high accuracy and the applicability to small size dataset.

In chapter 3, based on quantified parameters extracted from the medical images and demographic data, a two-step hierarchical classification with feature selection in

each step is proposed. This method is evaluated by building machine learning classification models of detection and classification for an ophthalmologic disease (glaucoma).

In chapter 4, deep learning technique is applied to handle medical images directly, and the deep learning models are trained with hierarchical classification method and hierarchy transfer learning method. Concept of this framework is first demonstrated with a natural image dataset, labels of which are hierarchically related. Then the framework is evaluated using a real clinical image dataset relevant with an ophthalmologic disease (age-related macular degeneration) diagnosis.

In chapter 5, extension method for the proposed two-step framework using hierarchical classification method and hierarchy transfer learning method is proposed to build machine learning models based on multiple medical images, trying to combine models built with each type of input images in a stacking manner. The effectiveness of this extended two-step framework is demonstrated using a retinal image dataset consisting four types of images extracted from volumetric data by building machine learning models for glaucoma detection and classification.

All the proposed methods enable machine learning models to achieve high accuracies and have good applicability to small size dataset. Conclusively, this dissertation demonstrates that the two-step hierarchical framework has a high potential of deploying high accurate machine learning models for ophthalmologic disease detection and classification with limited labelled data, to assist the clinical decision making of selecting the proper treatments for the patients to prevent the vision loss.

# Contents

|   |            |
|---|------------|
| <b>List of Figures</b>  | <b>v</b>   |
| <b>List of Tables</b>   | <b>vii</b> |
| <b>Chapter 1. Introduction</b>  | <b>1</b>   |
| 1.1. Ophthalmologic Disease Detection and Classification.....   | 1          |
| 1.2. Artificial Intelligence for Disease Detection and Classification.....  | 2          |
| 1.3. Proposed Framework of Developing AIs for Disease Detection and Classification<br>Based on Analysis of Diagnostic Processes ..... | 3          |
| 1.4. Structure of This Dissertation .....   | 5          |
| <b>Chapter 2. Literature Review of Artificial Intelligence for Disease Detection<br/>and Classification</b>                           | <b>7</b>   |
| 2.1. Artificial Intelligence in Medical Field.....  | 7          |
| 2.2. Traditional Machine Learning Techniques to Develop Medical AIs .....   | 9          |
| 2.3. Deep Learning Techniques to Develop Medical AIs .....  | 12         |
| <b>Chapter 3. Hierarchical Machine Learning Models Using Feature Selection<br/>Based on Quantified Parameters</b>                     | <b>17</b>  |
| 3.1. Overview.....  | 17         |
| 3.2. Hierarchical Machine Learning Models Using Feature Selection .....   | 18         |
| 3.3. Experiments Using a Dataset of Quantified Parameters .....   | 19         |
| 3.3.1. Datasets.....  | 20         |
| 3.3.2. Built Machine Learning Models and Training Details.....  | 24         |
| 3.3.3. Evaluation Metrics .....   | 29         |
| 3.3.4. Results and Discussion.....  | 33         |
| 3.4. Conclusions.....   | 37         |
| <b>Chapter 4. Hierarchical Deep Learning Models Using Hierarchy Transfer<br/>Learning Based on Single Input Image</b>                 | <b>39</b>  |
| 4.1. Overview.....  | 39         |
| 4.2. Hierarchical Deep Learning Models Using Hierarchy Transfer Learning.....   | 41         |
| 4.3. Conceptual Experiments Using Natural Image Dataset.....  | 43         |

|  |   |            |
|--|---|------------|
| 4.3.1.   | Datasets.....   | 43         |
| 4.3.2.   | Built Deep Learning Models and Training Details.....                          | 45         |
| 4.3.3.   | Evaluation Metrics .....  | 48         |
| 4.3.4.   | Results and Discussion.....   | 49         |
| 4.4.   | Experiments Using Clinical Image Dataset .....                                | 50         |
| 4.4.1.   | Datasets.....   | 50         |
| 4.4.2.   | Built Deep Learning Models and Training Details.....                          | 53         |
| 4.4.3.   | Evaluation Metrics .....  | 56         |
| 4.4.4.   | Results and Discussion.....   | 57         |
| 4.5.   | Conclusions.....  | 63         |
| <b>Chapter 5. Stacked Hierarchical Deep Learning Models Using Hierarchy Transfer Learning Based on Multiple Input Images</b> |   | <b>65</b>  |
| 5.1.   | Overview.....   | 65         |
| 5.2.   | Stacked Hierarchical Deep Learning Models Using Hierarchy Transfer Learning . | 66         |
| 5.3.   | Experiments Using Clinical Image Dataset .....                                | 67         |
| 5.3.1.   | Datasets.....   | 67         |
| 5.3.2.   | Built Deep Learning Models and Training Details.....                          | 69         |
| 5.3.3.   | Evaluation Metrics .....  | 75         |
| 5.3.4.   | Results and Discussion.....   | 76         |
| 5.4.   | Conclusions.....  | 84         |
| <b>Chapter 6. Conclusions and Future Work</b>  |   | <b>87</b>  |
| 6.1.   | Summary of This Dissertation.....   | 87         |
| 6.2.   | Discussion of the Future Work.....  | 90         |
| <b>Acknowledgements</b>  |   | <b>92</b>  |
| <b>Bibliography</b>  |   | <b>93</b>  |
| <b>Publications</b>  |   | <b>102</b> |

# List of Figures

|   |    |
|---|----|
| Fig. 1.1. Analysis of doctors' diagnostic processes for disease detection and classification .....  | 4  |
| Fig. 2.1. Relationship between AI, machine learning and deep learning .....   | 8  |
| Fig. 2.2. Medical image analysis with traditional machine learning techniques.....  | 10 |
| Fig. 2.3. Relationship between AI classification performance and the number of features .....   | 11 |
| Fig. 2.4. An example of a neural network with a single hidden layer.....  | 13 |
| Fig. 2.5. Medical image analysis with deep learning techniques .....  | 13 |
| Fig. 2.6. An example of convolutional neural network .....  | 14 |
| Fig. 2.7. An example of data augmentation method.....   | 15 |
| Fig. 2.8. Transfer learning from ImageNet pre-trained deep learning model in medical field.....   | 16 |
| Fig. 3.1. Overview of building machine learning models based on quantified parameters .....   | 18 |
| Fig. 3.2. Quantified parameters from volumetric optical coherence tomography images. ....   | 22 |
| Fig. 3.3. Built machine learning models with proposed approaches. ....  | 25 |
| Fig. 3.4. Flowchart of the proposed approach of feature selection applied.....  | 27 |
| Fig. 3.5. Pseudocode of generic algorithm .....   | 29 |
| Fig. 3.6. An example of cross validation (5-fold cross validation).....   | 32 |
| Fig. 3.7. Selected best feature subset for the glaucoma classification using entire dataset .....   | 34 |
| Fig. 3.8. Predicted result by the high-level model of the hierarchical model for glaucoma classification. ....  | 36 |
| Fig. 4.1. Overview of building deep learning models based on single input image.....  | 42 |
| Fig. 4.2. Two public image datasets to create a new dataset for evaluation.....   | 44 |
| Fig. 4.3. Proposed approaches to build deep learning models of classifying cats and four breeds of dogs. ....   | 46 |
| Fig. 4.4. Hierarchy transfer learning in hierarchical deep learning models .....  | 48 |
| Fig. 4.5. Classification performance of deep learning models based on natural images with different training methods and different size dataset. .... | 49 |
| Fig. 4.6. Preprocessing medical image dataset for evaluation of deep learning models built with proposed approaches. ....                             | 52 |

|  |    |
|--|----|
| Fig. 4.7. Proposed approaches to build deep learning models of classifying normal, AMD with fluid, (wet), and AMD without fluid (dry). .....   | 54 |
| Fig. 4.8. Important areas for the deep learning models trained with the proposed approach. ....  | 59 |
| Fig. 4.9. Differentiation of heatmaps created by built deep learning models to identify AMD OCT image with fluid. ....   | 61 |
| Fig. 4.10. Classification performance for deep learning models classifying normal, AMD with fluid, and AMD without fluid, with different training methods and different size dataset ..... | 62 |
| Fig. 5.1. Overview of building deep learning models based on multiple input images....   | 66 |
| Fig. 5.2. Image extraction from the volumetric OCT data. ....  | 68 |
| Fig. 5.3. Proposed approaches for building deep learning models with single input images separately before applying stacking for the overall result. ....                                  | 70 |
| Fig. 5.4. Proposed approaches for building deep learning models with multiple input images directly.....   | 72 |
| Fig. 5.5. Multiple input CNN models for glaucoma detection and classification .....  | 74 |
| Fig. 5.6. Cohen's kappa of deep learning models built with the proposed approaches. ...  | 76 |
| Fig. 5.7. Performance change with different training dataset sizes.....  | 78 |
| Fig. 5.8. ROC curves for stacked hierarchical classification models built using a small training dataset.....  | 81 |
| Fig. 5.9. Heatmaps for en face image of FI-type glaucoma (disease subtype) case.....   | 83 |



## List of Tables

|   |    |
|---|----|
| Table 3.1. Collected cases for evaluation of training methods .....   | 21 |
| Table 3.2. Extracted quantified parameters as the input data for machine learning .....   | 23 |
| Table 3.3. Parameters used in genetic algorithm based feature selection.....  | 29 |
| Table 3.4. An example of confusion matrix.....  | 30 |
| Table 3.5. An example of confusion matrix for a 3-class problem to calculate Cohen's kappa<br>.....                                       | 31 |
| Table 3.6. The interpretation of the Cohen's kappa.....   | 31 |
| Table 3.7. Classification performance of machine learning models based on quantified<br>parameters (Cohen's kappa of 5-fold CV).....      | 33 |
| Table 3.8. Classification performance of machine learning models based on quantified<br>parameters (weighted accuracy of 5-fold CV) ..... | 33 |
| Table 4.1. A newly created natural image dataset for conceptual experiments .....   | 45 |
| Table 4.2. Training dataset to build deep learning models .....   | 52 |
| Table 4.3. Test dataset to evaluate deep learning models.....   | 53 |
| Table 4.4. Classification performance of built deep learning models using entire dataset<br>.....   | 57 |
| Table 4.5. Training and prediction time for each deep learning model.....   | 63 |
| Table 5.1. Number of collected data for evaluation .....  | 68 |
| Table 5.2. Performance change of stacking models with different training dataset sizes<br>.....   | 77 |

# Chapter 1.

## Introduction

### 1.1. Ophthalmologic Disease Detection and Classification

The world is experiencing growth in the number of older persons in the population [1]. Compared to 2017, the number of persons aged 60 or above is expected to more than double by 2050 and to more than triple by 2100. The number of persons aged 80 or over is growing even faster, and is projected to triple by 2050, and nearly seven times by 2100 [1]. Aging of the world's population is causing a steep increase in number of people with vision impairment, which affects their ability to interact with the surroundings severely [2]. At present, there are more than 250 million people with vision impairment, of which 36 million are suffering with blindness, and these number would increase to about 700 million vision impaired and 115 million of these blind, after 30 years from now [2]. The worldwide societal costs of visual impairment have been estimated at \$3 trillion in 2010, most of which was direct health costs, and this burden is projected to increase by approximately 20% by 2020 [3]. Moreover, nearly 90% of the patients suffering vision impairment are from low and middle income countries (LMICs) [3], and the need for solving this social problem in LMICs is more acute.

Early detection and proper treatment at the early stage of the ophthalmologic diseases are important to prevent vision loss and improve the patients' quality of life [4], which are useful for solving the problem of increasing social costs of visual impairment. Conventional diagnose methods are tremendously depend on the ophthalmologists' knowledge and experience to detect and classify diseases subjectively to determine the proper treatment plan [5]. Major shortage and maldistribution of trained ophthalmologists are the main obstacle of realizing global eye healthcare [6]. Many of the interventions for the ophthalmologic diseases remain out of reach for millions living in underserved regions in LMICs as well as disadvantaged populations in high-income countries [6]. The appropriate distribution of the eye care workforce and the development of comprehensive eye care delivery systems are needed to ensure that eye care needs are universally met [6].

### 1.2. Artificial Intelligence for Disease Detection and Classification

In the healthcare industry, huge amount of data including hospital records, medical records of patients, results of medical examinations, and medical image data from different devices are now being accumulated in everyday clinical practice. Among these medical data, medical imaging plays a critical role in establishing the diagnoses for innumerable conditions and it is used routinely in nearly every branch of medicine [7]. With the development of technologies, many medical imaging techniques help doctors to understand the symptoms more thoroughly by providing more valuable information such as volumetric and time-series data [8], which lead to better diagnosis and more proper treatments for the patients. However, interpreting huge amount of volumetric and time-series image data is a big burden for the doctors including ophthalmologists, resulting in the big volume of data is not fully used [8].

Recently, there are lots of researches and social implementations trying to process the massive data to extract useful information via machine learning, a subset of artificial intelligence (AI) for assisting doctors in the clinical decision making [6]–[8]. There are two main features of AI as below. The first feature of AI is the ability to handle enormous amounts of information instantaneously with stable performance. The second feature is that AI can improve its accuracy by learning, which has progressed rapidly especially deep learning enters the stage. Because of these features of AI, AI is being paid more and more attention, and is being developed positively for solving many various social problems.

There are mainly two types of machine learning: one is supervised learning and the other is unsupervised learning. Supervised learning is to train a model from already labelled data, and until now methods adopted in most medical field research are performed in supervised manner because the accuracy and efficacy of supervised learning are better than unsupervised learning [9]. There have been lots of studies for diagnosis using medical data in supervised conditions, most of which have aimed at automatic detection of some diseases, not classifying diseases into subtypes that is relevant with determining proper treatment plans [7].

Machine learning, especially deep learning requires a large amount of supervised data, and the more input data they are provided, the better a machine learning-based model typically performs [10]. However, it is difficult to collect many supervised data by doctors in medical field [6]–[8]. Regarding the disease classification which is more needed in clinical use for treating patients properly, it is more difficult, because it requires much more experience for the doctors who label the data, compared with the disease detection. In most studies, the size of the data used for training machine learning models is from hundreds to thousands [6]–

[8]. To overcome the obstacle of insufficient labelled data in the medical field, there are many machine learning techniques are being proposed. There are mainly two aspects to consider for machine learning achieving good performance, data and optimization, according to its definition: to solve an optimization task using collected sample data. First, some approaches are focusing on making good data for building machine learning models, such as synthetically increasing the number of available samples or decreasing unnecessary information (features) of the sample data [9], [10]. Second, some approaches are focusing on how to optimize, such as designing better optimization method of machine learning classifiers, or decreasing difficulty of optimization for specific tasks. Recently, ‘not-so-supervised’ learning methods, which include semi-supervised, multi-instance, and transfer learning, among which transfer learning inspired by human thought processes has become the most popular [10], [11].

### **1.3. Proposed Framework of Developing AIs for Disease Detection and Classification Based on Analysis of Diagnostic Processes**

The major theme of this dissertation focused on building high accurate machine learning models for ophthalmologic disease detection and classification with small-size dataset.

To achieve this goal, first I analyzed the doctors’ diagnostic processes of disease detection and classification as follows (Fig.1.1). Note that diseases here are not limited to ophthalmologic diseases.

- 1) Doctors perform disease classification after classifying healthy and diseased cases (disease detection), because the disease classification is difficult, which is based on complex symptoms, and it should be performed in the status of normal cases excluded thoroughly.
- 2) Doctors classify diseases into subcategories by reusing the knowledge of classifying healthy and diseased cases.
- 3) Doctors use multiple information (e.g., multiple types of medical images) to make optimal diagnosis.

### 1.3 Proposed Framework of Developing AIs for Disease Detection and Classification Based on Analysis of Diagnostic Processes

---

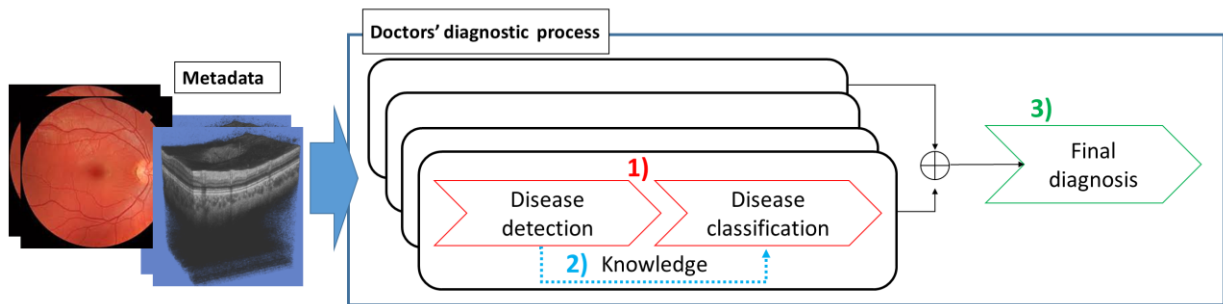


Fig. 1.1. Analysis of doctors' diagnostic processes for disease detection and classification

Based on the analysis of the doctors' diagnostic processes above, a framework of building machine learning models for disease detection and classification in two steps was proposed. In the first step, the model for classifying healthy and disease (disease detection) was built. In the second step the model for disease classification was built by reusing parameters of the model for disease detection. This two-step framework was further extended to handle multiple input data by combining the machine learning models trained separately on single input data.

In detail, in this dissertation, I proposed a training framework composing of three different methods, those were newly created based on the analysis of the doctors' diagnostic processes to build high accurate machine learning models, those can be used for disease detection and classification with limited training data, as below.

1) Hierarchical classification method, to build machine learning models for disease classification, after building the models for classifying healthy and disease (disease detection).

2) Hierarchy transfer learning method, to build machine learning models for disease classification by reusing parameters of the model for disease detection.

3) Stacking ensemble method, to build machine learning models handling multiple input data by combining the machine learning models trained separately on single input data with the training method of 1) hierarchical classification and 2) hierarchy transfer learning.

To demonstrate the effect and efficiency of the proposed framework for building machine learning models of disease detection and classification, a series of experiments using labelled medical datasets have been devised.

### 1.4. Structure of This Dissertation

The rest of this dissertation is organized as follows.

Chapter 2 gives a review on the principal AI techniques relevant with disease detection and classification, prior to the delineation of the three methods those form the framework of building machine learning models for ophthalmologic disease detection and classification based on the analysis of the doctors' diagnostic processes to achieve high accuracy with small size datasets.

Chapter 3 presents the first approach of building machine learning models handling quantified parameters in two steps: building machine learning models for classifying diseases after building the disease detection models. It is implemented by the two-step hierarchical classification method. In each step, feature selection that is one field of feature engineering is used to build machine learning models with the optimized subset of quantified parameters.

Chapter 4 presents the second approach attempting to apply deep learning technique for handling medical images directly without feature engineering to build machine learning models for ophthalmologic disease detection and classification. A two-step framework to build deep learning models for ophthalmologic disease detection and classification is proposed. In the first step, the machine learning model for disease detection is built, and in the second step the machine learning model for disease classification is built by reusing parameters of the model for disease detection. In detail, besides the two-step hierarchical classification method introduced in Chapter 3, hierarchy transfer learning method is used.

Chapter 5 presents an extension approach for the proposed two-step framework introduced in Chapter 4 to handle multiple input data. The machine learning models are trained separately based on each type of input image data extracted from a volumetric data with the two-step framework using hierarchical classification and hierarchy transfer learning method. Then the separate models were combined by the superior model in stacked manner (stacking ensemble method) to build machine learning models handling multiple images for ophthalmologic disease detection and classification.

Chapter 6 summarizes overall contribution of this dissertation, and discusses the necessary works in the future.

On the path to building high accurate machine learning models for ophthalmologic disease detection and classification with small-size dataset, three novel methods based on

## 1.4 Structure of This Dissertation

---

hierarchical classification method are presented. I wish this dissertation will steer away from current research orientation of flat classification to hierarchical classification in building machine learning ophthalmologic disease detection and classification for high accuracy.

## Chapter 2.

# Literature Review of Artificial Intelligence for Disease Detection and Classification

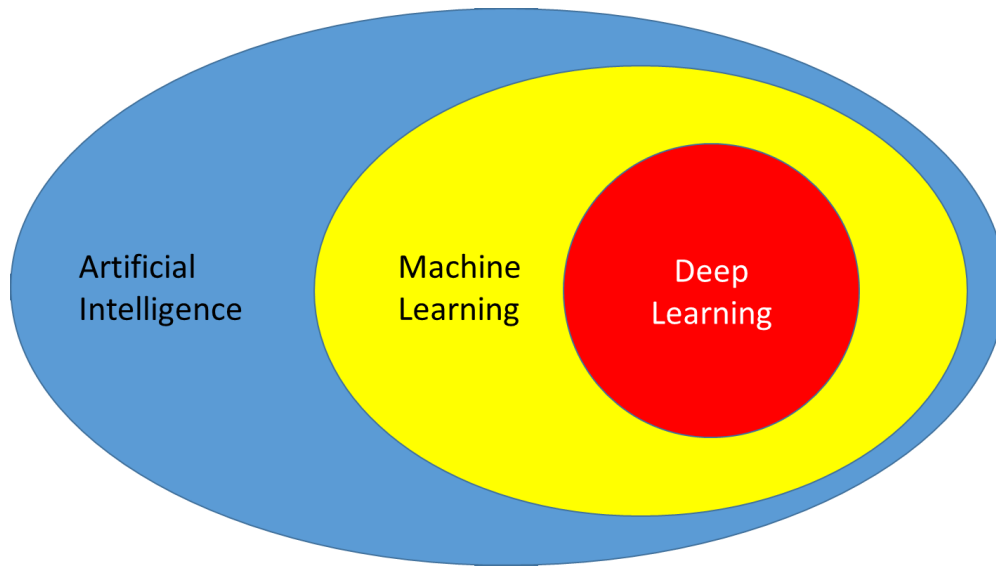
As mentioned in Chapter 1, the essence of the artificial intelligence (AI) for disease detection and classification has been briefly discussed. This chapter is devoted to give the readers the review of the principal AI techniques applied in medical field mainly ophthalmologic field for disease detection and classification, prior to the delineation of three researches to build high accurate machine learning models with small size dataset for ophthalmologic disease detection and classification.

### 2.1. Artificial Intelligence in Medical Field

In this section, artificial intelligence (AI), machine learning and deep learning are talked, and the current status of AIs in medical field especially ophthalmology is discussed.

AI was first proposed in 1950s, is defined as the intelligence of machines, as opposed to the intelligence of humans or other living species [12], [13]. AI also refers to situations wherein machines can simulate human thoughts in learning and analyzing, and thus can be applied in problem solving [14]. Machine learning (ML) is a subset of AI (Fig.2.1), and is proposed in 1980s. It is defined as a set of methods that automatically detect patterns in sample data and then incorporate this information to predict the data not included in the sample data under uncertain conditions [6], [15]. Deep learning (DL) is a subset of ML (Fig.2.1), and is a revolutionary technology of ML gathering attention occurred in 2000s. These technologies have been used in many aspects of modern society, such as object recognition based on images, language translation, etc.





**Fig. 2.1. Relationship between AI, machine learning and deep learning**

The medical field has been the frontier field of the AI application in recent years. Many studies have shown that DL algorithms achieved a high accurate classification performance when applied to detection and classification for skin, breast and lung cancer [16]–[18]. These inspirable research drives numerous studies to research and develop AI in ophthalmology. Traditional machine learning (TML) techniques not including DL applied in ophthalmologic filed were introduced and reviewed in the review paper by M. Caixinha and S. Nunes [19]. Daniel Shu Wei Ting, et al. detailly introduced and reviewed the DL applications in the ophthalmology field [20]. Moreover, Akkara John Davis, et al. introduced generally about the AI development in ophthalmology [21].

Depending on whether to incorporate the outcomes, there are mainly two types of machine learning: one is unsupervised learning and the other is supervised learning. Unsupervised learning is to train a model with unlabeled data, and has the ability of finding new knowledge in medical field, such as histopathology image analysis [22]. Although it is exciting to find new knowledge, it is very difficult to handle it properly, and validate its achievement. Supervised learning is to train a model from already labeled training data, tuning the parameters of the machine learning to improve the accuracy of its predictions until they are optimized. It may expedite classification process and would be useful for discriminating clinical outcomes of interest [6]. More recently, semi-supervised learning has

been proposed as a combination method of unsupervised learning and supervised learning, which is applicable for scenarios in which the outcome is missing for certain subjects. Supervised learning is to train a model from already labelled data by the specialists, and until now methods adopted in most medical field research are performed in supervised manner, because the accuracy and efficacy of supervised learning are better than unsupervised learning [9]. There have been lots of studies for diagnosis using medical image data in supervised manner, most of which have aimed at automatic detection, not classification aiming to assist treatment [7].

**In this dissertation, I built all the machine learning models in supervised manner to demonstrate my proposed framework of building machine learning models for ophthalmologic disease detection and classification.**

Machine learning, especially deep learning requires a large amount of supervised data, and the more input data they are provided, the better a machine learning-based model typically performs [10]. However, it is difficult to collect many supervised data by doctors in medical field [6]–[8]. Regarding the disease classification which is more needed in clinical use for treating patients properly, it is more difficult, because it requires much more experience for the doctors who label the data, compared with the disease detection. This might be a reason for the number of research relevant with disease classification is much less than the number of the research of disease detection. In most studies, the size of the data is from hundreds to thousands [6]–[8]. The most of research intensively studied are concentrating on the diseases, patients of which are large. In ophthalmology, the most intensively studied are diabetic retinopathy, glaucoma, age-related macular degeneration, cataract, all of which are the leading causes of world-wide blindness [6].

The rest of this chapter is organized as follows. Traditional machine learning techniques to develop medical AIs are presented in Section 2.2, while the deep learning techniques to develop medical AIs are presented in Section 2.3.

## **2.2. Traditional Machine Learning Techniques to Develop Medical AIs**

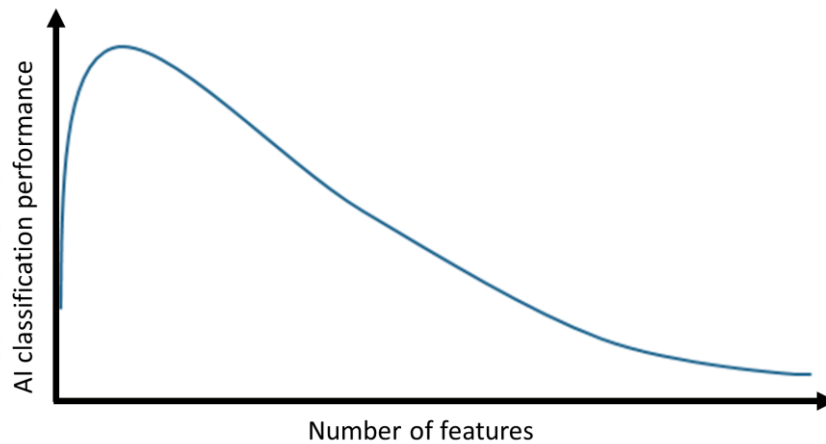
In this section, the basic traditional machine learning (TML) techniques including the overview of achieving high accuracy are introduced.

Existing traditional machine learning (TML) algorithms applied in ophthalmologic field include decision trees [23], random forests (RF) [24], support vector machines (SVM) [25], k-nearest neighbors [26], and neural networks (NN) [27]. TML can get satisfactory outcome with small datasets, but a cumbersome step to select specific effective features manually prior to classification is indispensable [28]. As the steps of building TML is shown in Fig.2.2, all the input data should be processed with the feature engineering method as the key step before being input into the TML algorithms for higher accuracy, for example the manually designed features should be extracted for machine learning.



**Fig. 2.2. Medical image analysis with traditional machine learning techniques**

Let the inputs for the TML algorithms are patients' features ( $X$ ) including the demographic data such as age, gender, etc., and disease-specific data such as medical image, etc., sometimes medical outcomes of interest ( $y$ ) such as treatment or diagnosis and so on. The TML algorithms constructing AI trying to solve optimization problems by extracting effective features from data ( $X$ ), and further finding the relationship function between  $X$  and  $y$ , when medical outcomes of interest are inputted. To fix ideas, the  $j$ th feature of the  $i$ th patient by  $X_{ij}$ , and the outcome of interest by  $Y_i$  was used. Generally speaking, the number of subjects ( $j$ ) increases, the performance of AI will improve. Meanwhile, the number of feature ( $i$ ) increases, the performance of AI will also increase to some extent, but after some points, the performance of AI will decrease, as the difficulty of optimization increases rapidly (Fig.2.3). Thus, it is required to consider the trade-offs between number of features and achieving a good accuracy for the AIs.



**Fig. 2.3. Relationship between AI classification performance and the number of features**

A critical part of the success of a machine learning model needs a good set of features to train on, in other words find the proper features  $j$  in  $X$  described above. In the past dozens of years, the dimensionality of the data involved in machine learning tasks has increased explosively, such as from 2D images to 3D images. With the presence of a large number of features, performance of a machine learning model tends to degenerate. To address the problem, dimensionality reduction techniques including feature extraction and feature selection subset of feature engineering have been studied in the machine learning research field.

Feature engineering, involves:

- Creating new features: create more powerful features based on the knowledge of specialists
- Feature extraction: methods that select or combine features into more powerful ones, effectively reducing the amount of input data, while keeping the ability of presenting the data accurately and completely
- Feature selection: selecting the most useful features to train on existing features

Feature selection is a widely employed technique for reducing dimensionality. It aims to choose a small subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to better classification performance (e.g.,

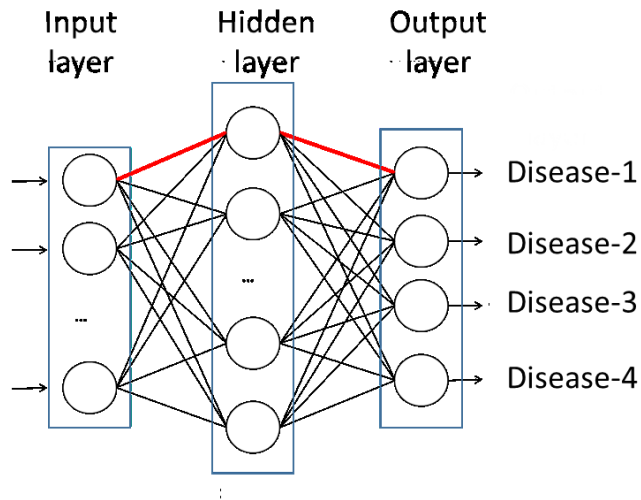
higher accuracy), lower computational cost, and better model interpretability. According to whether the training set is labelled or not, feature selection algorithms can be categorized into supervised, unsupervised and semi-supervised feature selection [29]. Supervised feature selection methods can further be broadly categorized into filter models, wrapper models and embedded models [29]. The filter model separates feature selection from classifier learning so that the bias of a learning algorithm does not interact with the bias of a feature selection algorithm [29]. It relies on measures of the general characteristics of the training data such as distance, consistency, and Information Gain based methods [29]. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features, which is time-consuming to run for data with a large number of features [29]. Due to these shortcomings in each model, the embedded model, was proposed to bridge the gap between the filter and wrapper models [30].

**In this dissertation, I applied feature selection method combining filter and wrapper method when applying TML techniques, which is described in detail in Chapter 3.**

## 2.3. Deep Learning Techniques to Develop Medical AIs

In this section, the neural network, that is a kind of TML method is introduced in detail, as its basic principles are same with deep learning, and deep learning techniques, which are well used in creating high accurate medical AIs recently are introduced.

Neural networks [27] are a TML method inspired from how the brain is structured, with hidden layers representing interneurons. Neural networks have several layers and units per layer to include in its architecture, designed initially (Fig.2.4). Each unit stores a numeric value (black circle in Fig.2.4), and each connection between units represents a weight (red line in Fig.2.4). Weights connect the units in different layers and represent the strength of connections between the units. The suitable value of these weights is estimated and tuned through the training process, that is necessary to obtain a correct classification result. A fully connected layer in which all units in one layer are connected to all neurons in the next can be interpreted and implemented by multiplication. The final layer encodes the desired outcomes or labelled categories, such as disease categories shown in Fig.2.4.



**Fig. 2.4. An example of a neural network with a single hidden layer**

A deep learning (DL) network is a neural network with multiple layers between the input and output layers. It has dramatically improved the state-of-the-art in image recognition [9]. When applied to image classification, a key difference between DL and TML algorithms is how they select and process image features. Features of input data are automatically learned in an unsupervised way by DL algorithms (Fig.2.5) [9]. Generally, TML is used when there is more limited, structured data available, while DL typically requires a large quantity of training data to ensure that the network that has millions of parameters does not overfit, that is not able to classify data other than the sample data for training the model. In other words, it is more difficult to build machine learning models with limited data via DL than via TML.



**Fig. 2.5. Medical image analysis with deep learning techniques**

A convolutional neural network (CNN) is a special case of the neural network that the most used DL method in the medical image recognition field is CNN. As Fig.2.6 shows, CNN consists of multiple convolutional layers that detect local conjunctions of features from the previous layer and mapping their appearance to a feature map and transform input images into hierarchical feature maps from simple features, such as edges and lines, to complicated features, such as shapes and colors [9]. In this phase, the activation function of units is used defining the output of that unit given an input or set of inputs should be used [9], and recently ReLU function, which returns the element-wise maximum of 0 and the input data, is most common used. The pooling layer is responsible for reducing the spatial size of the activation maps [9]. In general, they are used after multiple stages of other layers (i.e. convolutional layers) in order to reduce the computational requirements progressively through the network as well as minimizing the likelihood of overfitting [9]. The convolutional layer, pooling layer, activation layer form the automatic feature extraction part of a CNN model. Regarding the classification, fully connect layers that can combine these features and output a final probability value for the class [9]. Gulshan et al. [31] used deep learning to create an algorithm for the automated detection of two ocular diseases in retinal fundus photographs, using a dataset of 128,175 retinal images.

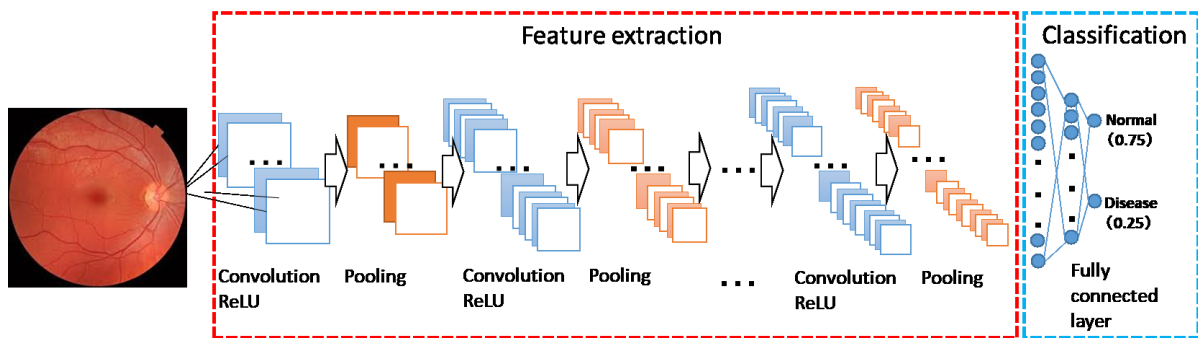
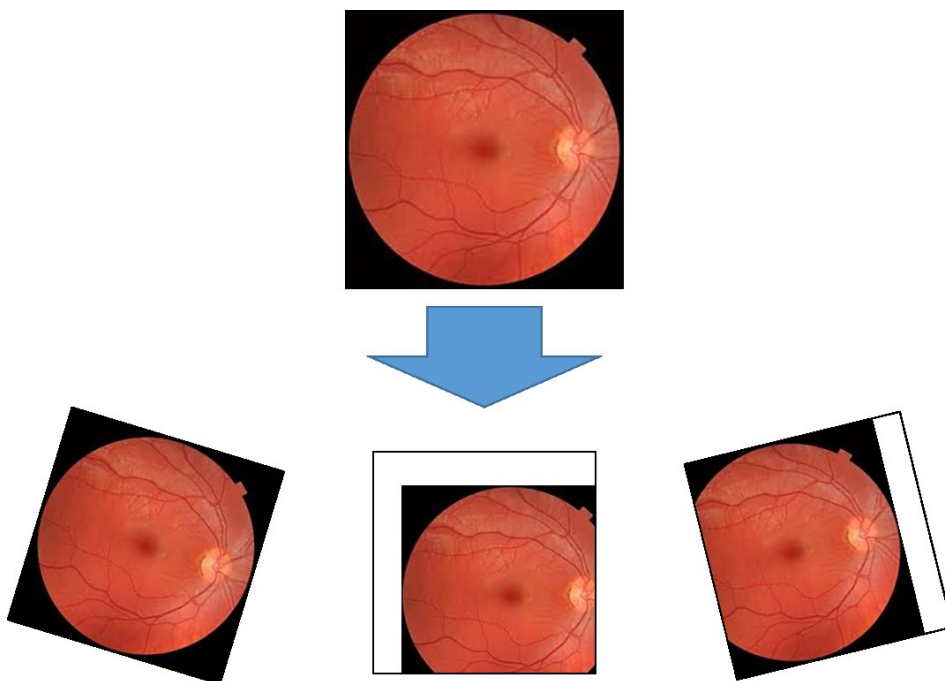


Fig. 2.6. An example of convolutional neural network

However, deep learning requires a large amount of supervised data, and the more input data they are provided, the better a deep learning-based model typically performs [6]. There are mainly two aspects to consider for machine learning achieving good performance, data and optimization, according to its definition: to solve an optimization task using collected sample data. Since deep learning is also a machine learning, there is no doubt to consider these

elements to overcome the obstacle of insufficient labelled data in the medical field using deep learning. Some approaches are focusing on more data for training by synthetically increasing the number of available samples, through data augmentation, via the geometrical transformation of medical images. There are three examples of data augmentation shown in Fig.2.7 with random rotation and shift. More recently, generative adversarial networks (GANs) are currently receiving tremendous attention in the computer vision community for their ability to mimic the distributions from which images are sampled [7], [32].



**Fig. 2.7. An example of data augmentation method**

The other approaches are focusing on how to optimize, such as designing better optimization method of machine learning classifiers, or decreasing difficulty of optimization for specific tasks. Recently, ‘not-so-supervised’ learning methods, which include semi-supervised, multi-instance, and transfer learning, among which transfer learning inspired by human thought processes has become the most popular [10], [11]. Transfer learning, inspired by human thought processes, is a method in which model parameter is effectively transferred across partially related or unrelated tasks [33]. Humans have an inherent ability to transfer knowledge across tasks. As Fig.2.8 shows, a pretrained model on a large visual dataset



(ImageNet) with more than 14 million natural images for visual object recognition can be reused regarding the parameters for disease detection (classifying healthy and not healthy subjects) via transfer learning. A study by Kermany, et al. [34] demonstrated the competitive performance of deep learning models built with transfer learning in classifying normal eyes and eyes with three macular diseases, using 4,000 optical coherence tomography images. In my previous work, a deep learning system using transfer learning technique in it can accurately differentiate between healthy and glaucomatous subjects based on their from retina image datasets [35].

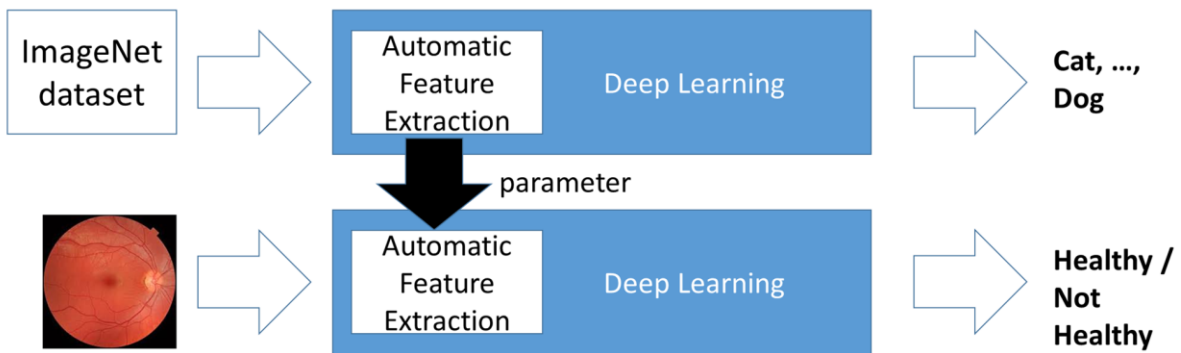


Fig. 2.8. Transfer learning from ImageNet pre-trained deep learning model in medical field

In this dissertation, I applied data augmentation method and transfer learning method for training deep learning models shown in Chapter 4 and Chapter 5.

## Chapter 3.

# Hierarchical Machine Learning Models Using Feature Selection Based on Quantified Parameters

### 3.1. Overview

In this chapter, to break the ice of building machine learning models of ophthalmologic disease detection and classification using proposed framework based on the analyzed doctors' diagnostic processes, a hierarchical classification method using feature selection was discussed in the status of using quantified parameters on the effectiveness of building high accurate machine learning models and the applicability of using small sized dataset.

As mentioned in the previous chapter, existing traditional machine leaning (TML) algorithms applied in ophthalmologic field include decision trees [23], random forests (RF) [24], support vector machines (SVM) [25], k-nearest neighbors [26], and neural networks (NN) [27] have achieved good accuracy [6]. Moreover, in some cases, the feature selection is a useful method to achieve a higher accuracy for the machine learning models with quantified parameters [6]. However, those researches simply applied straight forward approach of one-step directly to detect and classify diseases (flat classification). Although the obvious advantage of flat classification is its simplicity, it obviously loses some important information. The natural hierarchy of the data while the task of disease detection and classification would have highly valuable classification, thus ignoring those hierarchical class relationships would reduce performance to some extent. In this research, I try to improve the classification performance of machine learning models for ophthalmologic disease and classification, by building a hierarchical classification model using feature selection based on quantified parameters.

The rest of this chapter is organized as follows. Section 3.2 presents a two-step hierarchical classification model handling quantified parameters using feature selection in each step of the hierarchical classification model. Finally, experimental evaluations and concluding remarks are respectively presented in Section 3.3 and Section 3.4.

The contents of this work are based on the journal papers of Guangzhou An, et al. (2018) “Comparison of machine-learning classification models for glaucoma management” Journal of healthcare engineering, Kazuko Omodaka, et al. (2017) “Classification of optic disc shape in glaucoma using machine learning based on quantified ocular parameters” PLoS One, and Guangzhou An, et al. (2019) “Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images,” Journal of healthcare engineering.

### 3.2. Hierarchical Machine Learning Models Using Feature Selection

In this section, a training framework of building machine learning models based on quantified parameters for disease detection and classification is proposed and implemented by a two-step hierarchical classification method and a feature selection method.

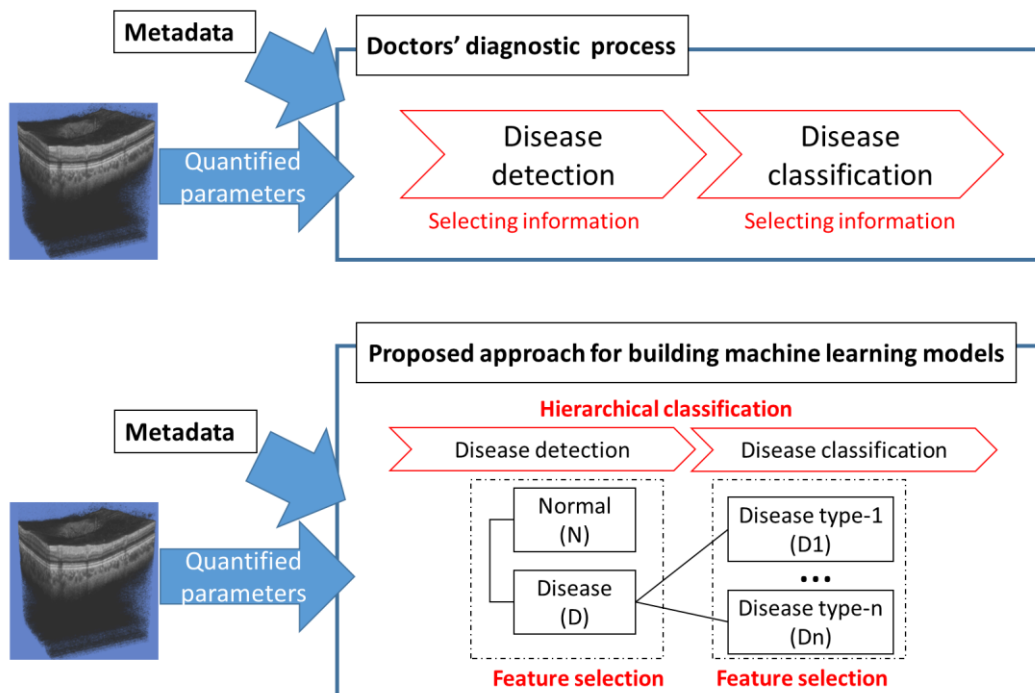


Fig. 3.1. Overview of building machine learning models based on quantified parameters

Sketched in upper part of Fig.3.1, first I analyzed the doctors' diagnostic processes as follows. Doctors first use extracted information (i.e. quantified ocular parameters) from image data together with the metadata to diagnose diseases. The determination of a treatment plan is difficult, as it requires disease classification based on complex symptoms, and it should be performed after disease detection. In other words, doctors classify healthy and diseased cases and to classify diseased cases into subcategories after excluding the healthy cases. Second, in each step, doctors select valid information to classify healthy and diseased cases, or to classify diseased cases into subcategories.

A novel training framework for machine learning models is proposed and created based on the analysis of the diagnostic process to achieve high accuracy in bottom part of Fig.3.1. The key components of the methods are hierarchical classification, feature selection, and they are used to build machine learning classification models based on quantified parameters. A hierarchical classification method is an efficient solution for building classification models with hierarchically structured local classification models according to a predefined hierarchy [36]. Thus, a two-step hierarchical classification method is created and used to separate the training steps of the model classifying healthy versus diseased cases and the model for disease classification, and to build the disease classification model after building the disease detection model. Moreover, feature selection is performed separately in the step of building models for classifying healthy versus diseased cases and the step of building models for classifying diseases into subcategories.

### **3.3. Experiments Using a Dataset of Quantified Parameters**

In this section, the description of the experimental setup and input data features are introduced in detail. The performances for the machine learning models built with proposed method are validated via quantitative evaluation. Two experiments are designed and performed to verify the proposed methods. One experiment compares the classification performance of models built with different training approaches, whereas the other experiment evaluates the applicability of the proposed approaches in training high accurate machine learning models using small size training dataset.

#### 3.3.1. Datasets

Glaucoma is a disease that causes progressive damage of the optic nerves, and it is the leading cause of blindness globally. The neurodegeneration is irreversible and patients may not be aware of it until its later stages; thus, early diagnosis and treatment are essential to prevent blindness. The optic disc is the point of exit for all retinal nerve fibers to the brain, and thus, it is important to observe the optic disc in glaucoma management. Besides intraocular pressure, which is an evidenced and treatable influencing factor, glaucoma is considered to be a multi-factorial disease: some of these factors are myopia, ocular blood flow, and oxidative stress [37]. However, there are no clear guidelines for the treatments. Nicolela proposed a guideline for identifying a glaucomatous optic disc based on its shape [38]. Nicolela's classification contains four types of glaucoma: local ischemic type (focal ischemic, FI), age-related hardening type (senile sclerotic, SS), myopic type (myopic, MY), and generalized enlargement (GE) [38]. Many studies have shown that this classification is helpful for understanding glaucoma pathogenesis [39]–[41]. Clinically, doctors always diagnose glaucoma by reading color fundus photos and subjectively identifying the specific optic disc type for glaucoma management. Unfortunately, some doctors have reported unmatched cases that make it difficult to decide further glaucoma treatment. Thus, accurate and objective methods are required for classifying optic discs. Relevant with the application of AI in ophthalmology field, the studies have used AI for classifying glaucoma and healthy eyes [42]–[44]. However, relevant studies for glaucoma management have not been conducted yet and more research efforts are required.

In some previous literature, the average intergrader or interobserver difference has been researched between ophthalmologists on the task of classifying healthy and diseased cases, and classifying diseases into their subcategories. For disease classification, according to the work by Abramoff, et al. [45] evaluated the sensitivity and specificity for detecting referable ophthalmologic disease (diabetic retinopathy) of the 3 masked independent retina specialists were 0.80 / 0.98, 0.71 / 1.00, and 0.91 / 0.95, and the average intergrader or interobserver difference (Cohen's kappa) was 0.822, which can be described as very good coincidence according to the guideline of Cohen's kappa [46]. On the other hand, for disease classification, it is more difficult for the doctors even the experienced ophthalmologists, for example, in the task of classification of age-related macular degeneration severity from color fundus photographs according to research by Peng et al. [47], the specialists' ability is Cohen's kappa: 0.517 to recognize large drusen and Cohen's kappa: 0.535 in recognizing pigmentary abnormalities. This means the specialists' agreement is just moderate level,

according to the guideline of Cohen’s kappa.

Hence, in this study to demonstrate the effectiveness of the training framework using a two-step hierarchical classification method, eyes reviewed and classified into four categories according to Nicoleta’s definition and normal by three glaucoma specialists, labels of which are consistent. In other words, only the labels of the same opinion by three different doctors were included in this research. Finally, totally 273 eyes were collected for evaluation of the machine learning models built with different proposed approaches (Table 3.1).

**Table 3.1. Collected cases for evaluation of training methods**

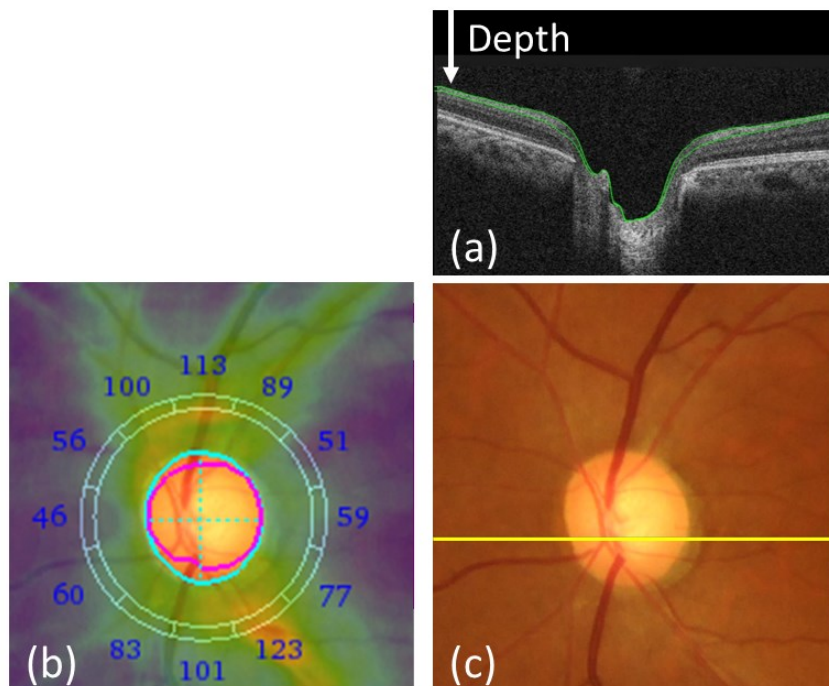
|                            | Normal | Glaucoma |    |    |    |
|----------------------------|--------|----------|----|----|----|
|                            |        | FI       | GE | MY | SS |
| Total data number<br>n=273 | 55     | 27       | 56 | 89 | 46 |

With the development of measuring techniques, many methodologies are available for observing the optic disc by shape, which has close relevance with the glaucomatous characteristics. Compared with color fundus photography, optical coherence tomography (OCT) based on low-coherence interferometry can image the tissue morphology with micrometer resolution, and therefore, it is being used widely in the ophthalmological field (Fig. 3.2). All the eyes in Table 3.1, All participants were additionally examined with swept source domain OCT (DRI OCT Atlantis, Topcon Corp., Tokyo, Japan), using a horizontal disc volumetric scan (6mm×6mm, 512 A-scans×256 frames). Cross-sectional OCT image at yellow line in a color fundus photography of the optic disc area (Fig. 3.2.c) is shown in Fig.3.2.a. The green lines in that OCT image show the detected layer information for calculating the retinal nerve fiber layer (RNFL) thickness. Using this RNFL thickness, a RNFL thickness map is created, where the number indicates the thickness in micrometers in 12 sectors around the optic disc and cyan and magenta circles show automatically detected disc and cup boundaries.

In recent years, the power of OCT in observing the optic disc more detailly is demonstrated to reveal the relationship of optic disc shape and glaucoma risk factors. For example, in my previous paper, OCT was used to develop a method to quantify the 3D

structure of the laminar pores; providing a useful tool to assess lamina cribrosa-associated risk factors for glaucoma [48].

There have been lots of commercial product to quantify the shapes of optic discs. For instance, by using integrated layer analysis software (DRI OCT Atlantis FastMap ver.9.30), 48 ocular parameters relevant to the circumpapillary retinal nerve fiber layer thickness (cpRNFLT) and optic disc morphology were quantified [49]–[51]. The evaluation result against OCT segmentation has been published online as a whitepaper (available at <http://www.topcon.co.jp/eyecare/handout>) [52].



**Fig. 3.2. Quantified parameters from volumetric optical coherence tomography images.** Cross-sectional OCT image at yellow line in (c) where green lines in (a) show the detected layer information for calculating the retinal nerve fiber layer (RNFL) thickness, (b) RNFL thickness map, where the number indicates the thickness in micrometers in 12 sectors around the optic disc and cyan and magenta circles show automatically detected disc and cup boundaries, and (c) a color fundus photo of the optic disc area.

**Table 3.2. Extracted quantified parameters as the input data for machine learning**

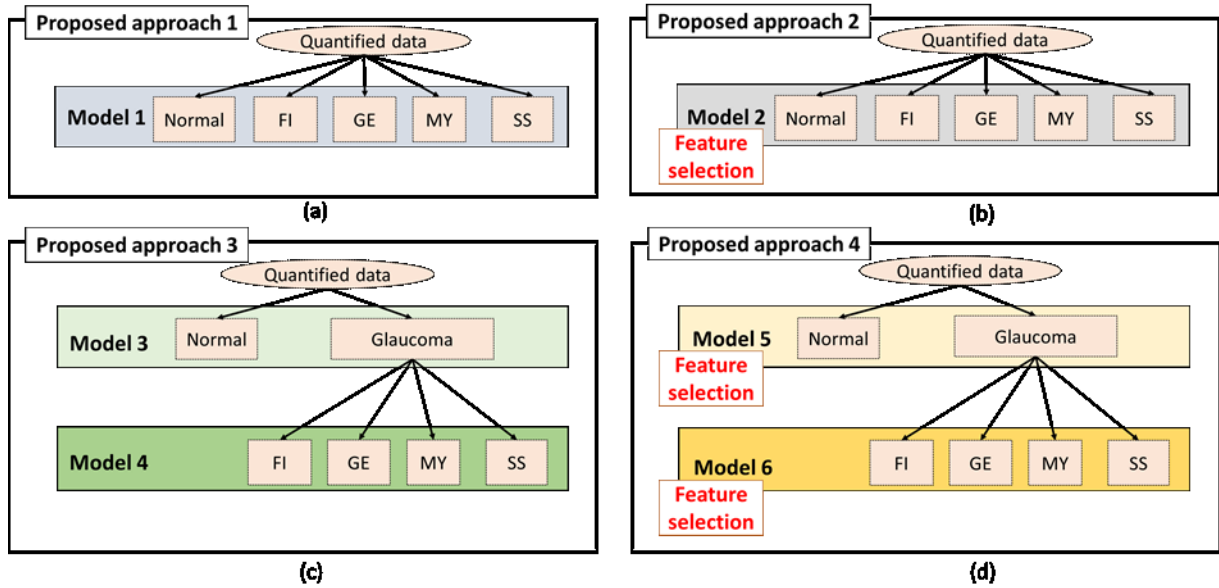
| No.   | Quantification                                   | Features  |
|-------|--|---|
| 1     |  | gender  |
| 2     |  | age   |
| 3     | patient's<br>metadata                            | spherical equivalent  |
| 4     |  | mean deviation  |
| 5     |  | pattern standard deviation                                    |
| 6     |  | internal ocular pressure                                      |
| 7     |  | central corneal thickness                                     |
| 8     |  | disc area   |
| 9     |  | cup area  |
| 10    |  | rim area  |
| 11    |  | vertical disc diameter  |
| 12    |  | horizontal disc diameter                                      |
| 13    |  | vertical cup/disc diameter ratio                              |
| 14    | optic disc                                       | horizontal cup/disc diameter ratio                            |
| 15    | shape  | cup/disc area ratio   |
| 16    | parameters                                       | rim/disc area ratio   |
| 17    | obtained from                                    | maximum cup depth   |
| 18    | OCT  | average cup depth   |
| 19–24 |  | average rim/disc area ratio (six sectors)                     |
| 25    |  | rim decentering area ratio                                    |
| 26    |  | horizontal disc angle   |
| 27    |  | disc height difference  |
| 28    |  | retinal pigment epithelium (RPE) height difference            |
| 29    |  | disc tilt angle   |
| 30    |  | average cpRNFLT   |
| 31–34 | cpRNFLT<br>average<br>thickness<br>obtained from | cpRNFLT (quadrants)   |
| 35    |  | difference in cpRNFLT (superior and inferior in four sectors) |
| 36–42 |  | cpRNFLT (six sectors)   |
| 42    |  | rim decentering cpRNFLT ratio                                 |
| 43    |  | OCT   |
| 44–55 |  | cpRNFLT (clockwise sectors)                                   |



Seven demographic parameters, such as gender, age, and spherical equivalent, were also extracted among the 48 quantified ocular parameters for each eye after rudimentary judgement of various ocular parameters, as shown in Table 3.2. OCT parameters, including 22 parameters related to disc topography and 26 parameters related to cpRNFLT, were measured with SS-OCT software. The cpRNFLT was calculated in the quadrants, 6 radial sectors, and the clockwise sectors. Some parameters were demonstrated powerful to build a high accurate (partial area under receiver operating characteristic curve = 0.864) machine learning model to classify healthy and glaucomatous eyes, in my previous work [35].

#### **3.3.2. Built Machine Learning Models and Training Details**

In this sub-section, the built machine learning model trained with different proposed approaches are presented, and training details are also shown in detail.



**Fig. 3.3. Built machine learning models with proposed approaches.** (a) Proposed approach 1: flat classification method is used to directly classify normal, FI, GE, MY, SS by training a traditional machine learning (TML) classifier to create Model 1. (b) Proposed approach 2: flat classification method is used to directly classify normal, FI, GE, MY, SS by training a TML classifier with feature selection applied to create Model 2. (c) Proposed approach 3: hierarchical classification method is used to create a low-level model (Model 3) for classifying normal versus glaucoma and a high-level model (Model 4) for classifying glaucoma classification. (d) Proposed approach 4: hierarchical classification method is used to create a low-level model (Model 5) for classifying normal versus glaucoma and a high-level model (Model 6) for classifying glaucoma classification, with feature selection method applied in each level.

The first experiment was performed as described below, using the entire set of training data to compare classification performances among models trained using four proposed approaches (Fig. 3.3). Machine learning models for glaucoma detection and classification were trained separately with traditional machine learning techniques. A neural network (NN), with a fixed unit number 8 of the just one hidden layer, was used as the classifier. The flat classification method was used to directly classify normal, FI, GE, MY, SS by training a NN to create Model 1 (Fig. 3.3.a) The flat classification method was used to directly classify normal, FI, GE, MY, SS by training the NN with feature selection applied to create Model 2 (Fig. 3.3.b).

The hierarchical classification method was used to create a low-level model (Model 3) for classifying normal versus glaucoma and a high-level model (Model 4) for classifying glaucoma classification (Fig. 3.3.c). The hierarchical classification method was used to create a low-level model (Model 5) for classifying normal versus glaucoma and a high-level model (Model 6) for classifying glaucoma classification, with feature selection method applied in each level (Fig. 3.3.d).

The second experiment was performed to evaluate the proposed approaches' applicability by small amounts of training data in building with just half training data, which was created using a stratified random sampling strategy. With the size-decreased training datasets, machine learning models were newly built with the proposed approaches (Fig. 3.3).

#### **Feature selection details**

In this chapter, regarding the feature selection method, a combination of a filter method of Minimum redundancy maximum relevance (mRMR) and a wrapper method of genetic-algorithm-based feature selection (GAFS) is applied (Fig.3.4).

The mRMR algorithm tends to select a subset of features having the most correlation with the labels and the least correlation among themselves [53]. It ranks features according to the minimal-redundancy-maximal-relevance criterion which is based on mutual information. In other words, it has been widely used recently because it assesses the tradeoff of maximizing the relevance between each feature and label and minimizing the feature redundancy [53]. In wrappers, a heuristic search shows higher performance but is too time-consuming, especially for a large number of features. Thus, instead of brute-force selection, more efficient strategies have been developed, such as genetic-algorithm-based feature selection (GAFS) using randomness that mimics natural evolution [54]. Filters are often used in combination with heuristic wrappers for principal selection [55].

The feature selection steps are as below. First standardization that is a scaling technique let the mean of the feature becomes zero and the resultant distribution has a unit standard deviation, is applied on the input data, then with mRMR candidate features (15 features) are found, and then GAFS using the neural network with one hidden layer (number of units: 8) as the classifier is applied to find the most valid features and classifiers (Fig. 3.4). The details of the GAFS and 5-fold cross validation will be presented in the next sub sections.

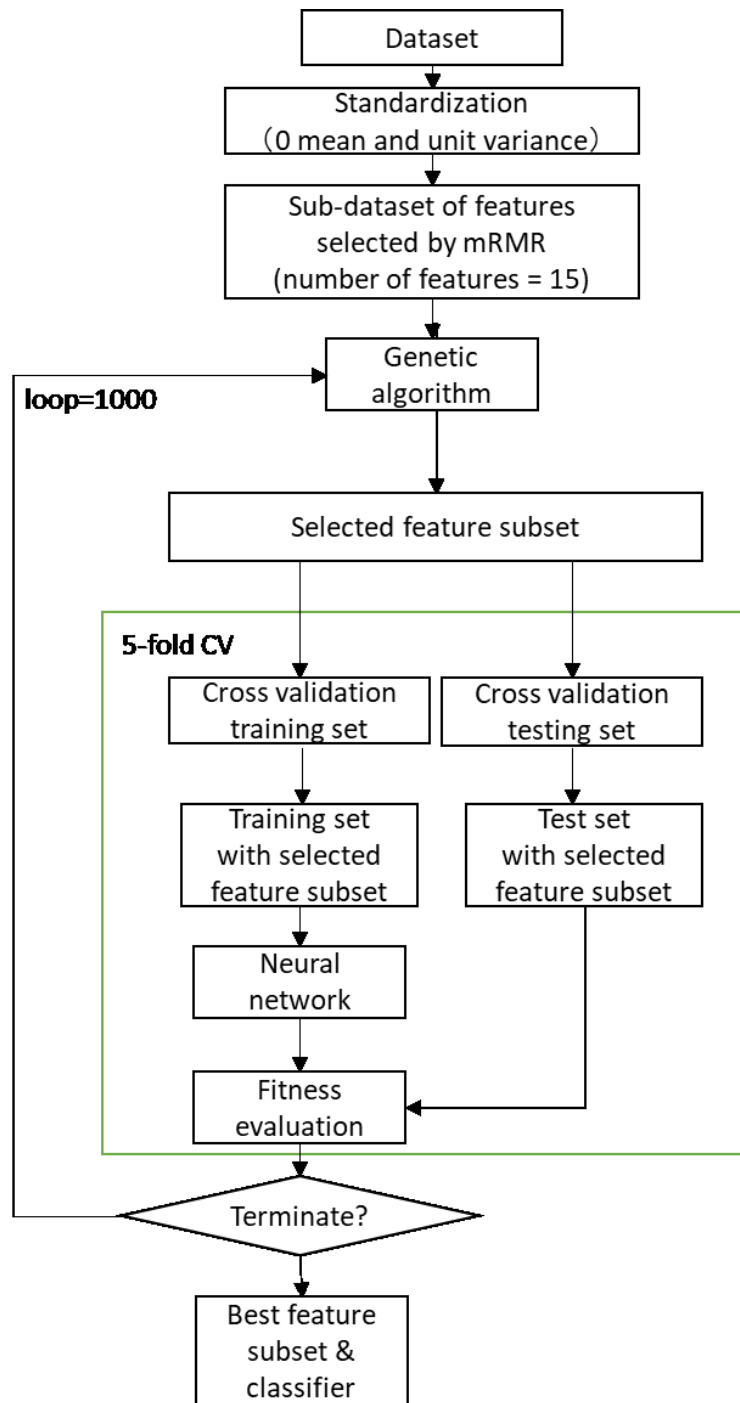


Fig. 3.4. Flowchart of the proposed approach of feature selection applied

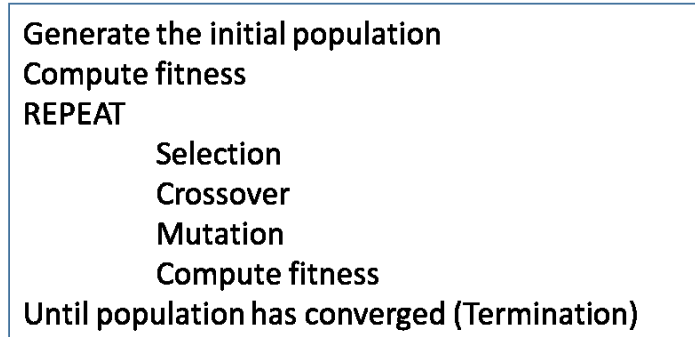
#### **Genetic Algorithm Based Feature Selection**

A genetic algorithm (GA) is a method for solving optimization tasks based on a natural selection process that mimics biological evolution [56]. This algorithm mimics the process of natural selection where the fittest individuals which are well adapted to the environment. This algorithm can be adopted in the feature selection field [54].

Following phases should be considered in usage of a genetic algorithm [56].

- 1) Initial population. The process begins with a set of individuals which is called a Population. Each individual is a solution to the problem to be solved, for example a subset of features in feature selection task. An individual is characterized by a set of features known as genes. In a genetic algorithm, the set of genes of an individual is represented usually, binary values are used: one means the feature selected while zero means the feature not selected.
- 2) Fitness function. The ability of an individual to compete with other individuals is calculated with fitness function. It gives a fitness score to each individual such as classification performance in feature selection.
- 3) Selection. One pair of individuals called parents, are selected based on their fitness scores. Individuals with high fitness score have more chance to be selected to next generation.
- 4) Crossover. For each pair of parents to be mated, a crossover point is chosen randomly from within the genes.
- 5) Mutation. Some of the genes of certain new offspring formed are subjected to a mutation with a low random probability.
- 6) Termination. The process terminates if the population has converged. Then it is recognized as a set of solutions for the task by genetic algorithm.

Based on the elements of genetic algorithm, the pseudocode can be described as Fig.3.5.



**Fig. 3.5. Pseudocode of generic algorithm**

Table 3.3 lists the parameters used in GAFS of this research.

**Table 3.3. Parameters used in genetic algorithm based feature selection**

| <b>GAFS Parameter</b> | <b>Value</b> |
|-----------------------|--------------|
| Population size       | 20           |
| Crossover probability | 0.7          |
| Mutation probability  | 0.2          |
| Number of generations | 1000         |
| Early stopping        | Used         |

#### **3.3.3. Evaluation Metrics**

In this work, to compare the performance of different training approaches of whether using a hierarchical classification method or feature selection method or not, weighted accuracy and Cohen's kappa of 5-fold cross-validation (CV) are used as the evaluation criteria.

##### **Weighted accuracy**

To evaluate the machine learning classification performance, there are many performance

metrics commonly used. Most of the metrics have been calculated from a confusion matrix that comprises false negatives (FN) true negatives (TN), true positives (TP), and false positives (FP). The importance of these four measures may shift depending on the application. The fraction of all correctly predicted overall number of test set samples is the accuracy as the Eq. (3.1), where the confusion matrix is as Table.3.4.

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) * 100\% \quad (3.1)$$

**Table 3.4. An example of confusion matrix**

|               |          | Predicted class |          |
|---------------|----------|-----------------|----------|
|               |          | Positive        | Negative |
| Correct class | Positive | TP              | FP       |
|               | Negative | FN              | TN       |

The weighted accuracy adjusted with the number of data in each class [57] was applied as the evaluation metrics, to evaluate the machine learning models built with imbalanced dataset.

#### Cohen's kappa

Cohen's kappa is an interesting alternative measure to the accuracy, since it compensates for random hits [46]. It was first introduced as a measure of agreement between observers of psychological behavior. The original intent of Cohen's kappa is to measure the degree of agreement, or disagreement, between two people observing the same phenomenon. The range of Cohen's kappa values extends from positive to negative one, with positive one indicating strong agreement, negative one indicating strong disagreement, and zero indicating chance level agreement. The Cohen's kappa can be calculated according to the Eq. (3.2).

$$\text{Cohen's kappa} = \frac{N \sum_{i=1}^m CM_{ii} - \sum_{i=1}^m C_{i\text{corr}} C_{i\text{pred}}}{N^2 - \sum_{i=1}^m C_{i\text{corr}} C_{i\text{pred}}} \quad (3.2)$$

where  $CM_{ii}$  represents the diagonal elements of the confusion matrix,  $C_{i\text{corr}}$  represents the number of data labelled as  $C_i$ , and  $C_{i\text{pred}}$  represents the number of data predicted by the machine learning model as  $C_i$ .  $N$  is the total number of cases.

**Table 3.5. An example of confusion matrix for a 3-class problem to calculate Cohen’s kappa**

|               |       | Predicted class   |                   |                   |                   |
|---------------|-------|-------------------|-------------------|-------------------|-------------------|
|               |       | C1                | C2                | C3                | Total             |
| Correct class | C1    | a                 | b                 | c                 | $C1_{corr}=a+b+c$ |
|               | C2    | d                 | e                 | f                 | $C2_{corr}=d+e+f$ |
|               | C3    | g                 | h                 | i                 | $C3_{corr}=g+h+i$ |
|               | Total | $C1_{pred}=a+d+g$ | $C2_{pred}=b+e+f$ | $C3_{pred}=c+f+i$ | N                 |

Note a 3-class problem is taken, for which confusion matrix including marginal values is shown in Table 3.5. Being N the total number of patterns, C1, C2 and C3 the label related with class 1, 2 or 3, respectively. Their Cohen’s kappa is given by Eq. (3.3).

$$\text{Cohen's kappa} = \frac{N*(a+e+i)-(C1_{corr}*C1_{pred}+C2_{corr}*C2_{pred}+C3_{corr}*C3_{pred})}{N^2-(C1_{corr}*C1_{pred}+C2_{corr}*C2_{pred}+C3_{corr}*C3_{pred})} \quad (3.3)$$

There is a guideline to interpret the Cohen’s kappa value (Table.3.6).

**Table 3.6. The interpretation of the Cohen’s kappa**

| Value of Cohen’s kappa | Strength of agreement |
|------------------------|-----------------------|
| <0.2                   | Poor                  |
| 0.21-0.40              | Fair                  |
| 0.41-0.60              | Moderate              |
| 0.61-0.80              | Good                  |
| >0.80                  | Very Good             |



#### Cross validation

Cross validation is a statistical method used to estimate the performance of machine learning models. The procedure of K-fold cross validation is shown below, and 5-fold cross validation is shown as an example in Fig.3.6.

- 1) splitting the full dataset into k equal length partitions
- 2) selecting k-1 partitions as the training set and
- 3) selecting the remaining partition as the test set
- 4) training the model on the training set
- 5) using the trained model to predict labels on the test set
- 6) computing an evaluation metric (e.g. accuracy) and setting aside the value for later, repeating all of the above steps k-1 times, until each partition has been used as test set for an iteration
- 7) calculating the mean of the k evaluation metric values



**Fig. 3.6. An example of cross validation (5-fold cross validation)**

**3.3.4. Results and Discussion**

The Cohen’s kappa and weighted accuracy of 5-fold CV for the machine learning models trained by the proposed training framework whether using hierarchical classification strategy or feature selection method or not, were shown in Table. 3.7 and Table. 3.8.

**Table 3.7. Classification performance of machine learning models based on quantified parameters (Cohen’s kappa of 5-fold CV)**

| Number of training data | Ratio of training data | Cohen’s kappa of 5-fold CV |         |       |              |
|-------------------------|------------------------|----------------------------|---------|-------|--------------|
|                         |                        | FC                         | FC & FS | HC    | HC & FS      |
| 273                     | 100%                   | 0.697                      | 0.786   | 0.712 | <b>0.846</b> |
| 187                     | 50%                    | 0.496                      | 0.566   | 0.537 | <b>0.614</b> |

**Table 3.8. Classification performance of machine learning models based on quantified parameters (weighted accuracy of 5-fold CV)**

| Number of training data | Ratio of training data | Weighted accuracy of 5-fold CV |         |       |              |
|-------------------------|------------------------|--------------------------------|---------|-------|--------------|
|                         |                        | FC                             | FC & FS | HC    | HC & FS      |
| 273                     | 100%                   | 83.8%                          | 89.7%   | 85.3% | <b>91.2%</b> |
| 187                     | 50%                    | 65.3%                          | 72.4%   | 71.5% | <b>75.5%</b> |

With the entire 100% of dataset (n=273), all the proposed approaches achieved good Cohen’s kappa (Cohen’s kappa>0.6), and the proposed approach-4 in Fig.3.3 (hierarchical classification and feature selection (HC & FS)) achieved the highest Cohen’s kappa 0.846, very good performance, followed by flat classification (FC) & FS, HC, and FC. The Cohen’s kappa for HC&FS is higher than FC&FS, and Cohen’s kappa for HC was higher than FC. This means HC was better to achieve higher kappa, compared with FC with or without using feature selection. Moreover, FS let both HC and FC achieved higher Cohen’s kappa. The performance change of HC via FS ( $0.846 / 0.712 * 100\% = 118\%$ ) was bigger than FC via FS ( $0.786 / 0.697 * 100\% = 112\%$ ).

### 3.3 Experiments Using a Dataset of Quantified Parameters

With the half 50% of entire dataset (n=187), only the HC&FS achieved good Cohen's kappa (Cohen's kappa>0.6). The Cohen's kappa for HC&FS is higher than FC&FS, and Cohen's kappa for HC was higher than FC. This means HC was better to achieve higher kappa, compared with FC with or without using feature selection. Moreover, FS let both HC and FC achieved higher Cohen's kappa. The performance change of HC via FS ( $0.614 / 0.537 * 100\% = 114\%$ ) was bigger than FC via FS ( $0.566 / 0.496 * 100\% = 114\%$ ).

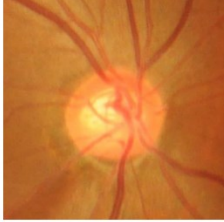
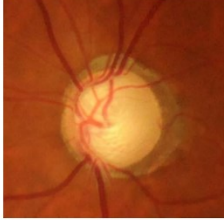
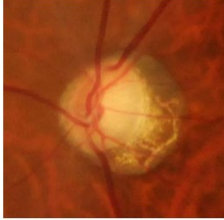
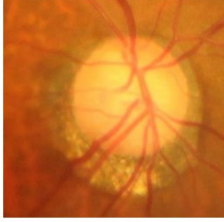
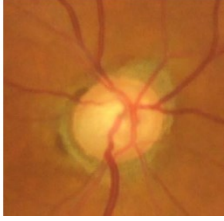
For the other evaluation metrics of weighted accuracy of 5-fold CV, the trends of the classification performance were similar with the trends of Cohen's kappa of 5-fold CV, for the built machine learning models with different training frameworks. With both the entire and half dataset, HC & FS achieved the highest weighted accuracy, followed by flat classification (FC) & FS, HC, and FC.

| No. | Features  | Contribution |
|-----|---|--------------|
| 1   | horizontal disc angle                             | 1.00         |
| 2   | spherical equivalent                              | 0.82         |
| 3   | cup area  | 0.50         |
| 4   | age   | 0.48         |
| 5   | cpRNFLT (temporal superior sector in six sectors) | 0.42         |
| 6   | average cup depth                                 | 0.40         |
| 7   | nasal rim/disc area ratio                         | 0.38         |
| 8   | maximum cup depth                                 | 0.33         |
| 9   | cpRNFLT (superior sector in four sectors)         | 0.31         |

**Fig. 3.7. Selected best feature subset for the glaucoma classification using entire dataset**

Using the NN with just a single hidden layer (number of units: 8) as the classifier for this problem, the nine most valuable ocular parameters were chosen by hybrid FS (Fig.3.7). Seven parameters (horizontal disc angle, cup area, cpRNFLT (six sectors: temporal superior), average cup depth, nasal rim/disc ratio, maximum cup depth, and cpRNFLT (four sectors: superior)) were extracted from OCT, and two parameters (spherical equivalent and age) pertained to patients' demographic data. The contribution of each selected parameter was also calculated by using the weights for each unit in the trained NN [56]. Doctors can classify and

check optic disc types using these features along with their contribution values, and not just by reading the color fundus images. The most contributed features for the glaucoma detection were cup area, cpRNFLT (six sectors: temporal superior), rim/disc ratio (area), nasal rim/disc ratio, and cpRNFLT (total), cpRNFLT(four sectors: superior superior)) are extracted from OCT, and parameters (spherical equivalent), similar with the ones for glaucoma classification, but some are new for the glaucoma detection model.

|                         |                         |                        |                   |   |   |
|-------------------------|-------------------------|------------------------|-------------------|---|---|
| (a)                     | <b>Label by Doctors</b> | <b>Predicted Label</b> | <b>Type</b>       | <b>Confidence</b>   |    |
|                         | <b>FI</b>               |                        | <b>FI</b>         | 95.99%  |   |
|                         |                         |                        | <b>SS</b>         | 2.14%   |   |
|                         |                         |                        | <b>GE</b>         | 1.82%   |   |
|                         |                         |                        | <b>MY</b>         | 0.05%   |   |
| <b>Label by Doctors</b> | <b>Predicted Label</b>  | <b>Type</b>            | <b>Confidence</b> |  |   |
| <b>GE</b>               |                         | <b>GE</b>              | 98.48%            |   |   |
|                         |                         | <b>SS</b>              | 1.38%             |   |   |
|                         |                         | <b>FI</b>              | 0.13%             |   |   |
|                         |                         | <b>MY</b>              | 0.01%             |   |   |
| (c)                     | <b>Label by Doctors</b> | <b>Predicted Label</b> | <b>Type</b>       | <b>Confidence</b>   |    |
|                         | <b>MY</b>               |                        | <b>MY</b>         | 99.92%  |   |
|                         |                         |                        | <b>SS</b>         | 0.06%   |   |
|                         |                         |                        | <b>GE</b>         | 0.01%   |   |
|                         |                         |                        | <b>FI</b>         | 0.01%   |   |
| (d)                     | <b>Label by Doctors</b> | <b>Predicted Label</b> | <b>Type</b>       | <b>Confidence</b>   |   |
|                         | <b>SS</b>               |                        | <b>SS</b>         | 98.58%  |   |
|                         |                         |                        | <b>MY</b>         | 0.71%   |   |
|                         |                         |                        | <b>FI</b>         | 0.36%   |   |
|                         |                         |                        | <b>GE</b>         | 0.35%   |   |
| (e)                     | <b>Label by Doctors</b> | <b>Predicted Label</b> | <b>Type</b>       | <b>Confidence</b>   |  |
|                         | <b>SS</b>               |                        | <b>GE</b>         | 56.06%  |   |
|                         |                         |                        | <b>SS</b>         | 43.92%  |   |
|                         |                         |                        | <b>FI</b>         | 0.02%   |   |
|                         |                         |                        | <b>MY</b>         | 0.00%   |   |

**Fig. 3.8. Predicted result by the high-level model of the hierarchical model for glaucoma classification.** (a) Successful example of prediction for FI and color fundus photo, (b) successful example of prediction for GE and color fundus photo, (c) successful example of prediction for MY and color fundus photo, (d) successful example of prediction for SS and color fundus photo, and (e) failure example of prediction and color fundus photo.

The high-level model of the proposed model can calculate the confidence of the prediction of glaucoma classification (Fig.3.8). When validating the prediction with the test data by using the highest one as the prediction, the overall accuracy was 87.8%. With regard to failure prediction examples, it was found that the developed classification model classified the correct answer as the second choice in most cases (Fig. 3.8.e). If the second choice is also considered to be correct, the accuracy was 95.9%. In some cases, specialists also narrow down the answer to two or more, such as FI and MY optic discs (Fig. 3.8.e), because FI optic discs clinically always have myopic characteristics as do the MY type. Thus, the machine-learning classification model might well reflect the actual clinical problem, and the prediction calculated by this approach can assist doctors in understanding the glaucomatous optic disc shape among glaucomatous subjects.

### 3.4. Conclusions

In this chapter, I proposed the first approach of building machine learning models, that tries to build models for classifying diseases after building models for disease detection based on the analysis result of doctors' diagnostic processes. It is implemented by the hierarchical classification method of two steps, and in each step, feature selection that is one field of feature engineering is used to find the optimized feature subset for the classification.

In experimentations, I evaluated this novel method by building an ophthalmologic disease (glaucoma) detection and classification machine learning models with a clinical data extracted from a hospital database. The dataset consists of extracted quantified parameters from medical images and demographic data, together with the labelled annotation data by experienced ophthalmologists. The classification performance of the models built with the proposed approach were compared to the flat models built in one step to detect and classify diseases, trained with different size datasets random sampled from the dataset above. The result of the experiments demonstrates that the classification performance of the machine learning models trained with hierarchical classification method combining with feature selection can be elevated on the disease detection and classification task compared with the flat models, and the applicability of the proposed framework in small size datasets is also demonstrated in achieving high accuracy for the built machine learning models.

In this work, two major contributions are made: 1) to build machine learning models for disease detection and classification hierarchically can boost the classification performance;

### 3.4 Conclusions

---

first classifying normal and diseased cases, then classifying diseased cases into sub-categories, 2) a feature selection method to select the most effective quantified parameters help improve the performances.

In conclusion, the proposed two-step hierarchical framework has a high potential of deploying high accurate machine learning models for ophthalmologic disease detection and classification based on quantified parameters, and applicability in using small size dataset.

## Chapter 4.

# Hierarchical Deep Learning Models Using Hierarchy Transfer Learning Based on Single Input Image

### 4.1. Overview

In the foregoing attempt, the approach of building machine learning models for ophthalmologic disease detection and classification hierarchically in two steps, building disease classification model after building disease detection model, was suggested. To implement this approach, a hierarchical classification technique of two steps with feature selection method in each step for building the machine learning models separately was also proposed. However, there are some limitations existed in that approach. For example, selecting specific features manually prior to classification, and image processing technique is needed to perform on the medical images to prepare the features (quantified parameters) for a good classification performance. Image processing technique is needed to be developed on specific disease, separately. Moreover, in some cases it is difficult to extract effective quantified parameters from the medical images, because the variation of diseases results in variant medical images according to the symptoms. Hence, it needs a better training approach for machine learning models those handle medical images directly as the input.

Deep learning, that a subtype of machine learning, can omit the cumbersome step of the extracting specific features manually prior to classification [6], while the difficulty of building machine learning models increases just using the raw images of large features with limited data. In other words, deep learning requires more supervised data to achieve high accuracy, compared to the traditional machine learning. To overcome the obstacle of insufficient labelled data in the medical field using deep learning. there are two kinds of approaches. One kind of approaches is focusing on using more good data for training by synthetically increasing the number of available samples, through data augmentation, via the geometrical transformation of medical images. More recently, generative adversarial networks (GANs) are currently receiving tremendous attention in the computer vision community for their ability to mimic



the distributions from which images are sampled [7], [32]. Another kind of approaches is to decrease the difficulty of optimization. Recently, ‘not-so-supervised’ learning case, which includes semi-supervised, multi-instance, and transfer learning, among which transfer learning has recently become the most popular [11]. Transfer learning, inspired by human thought processes, is a method in which model parameter is effectively transferred across partially related or unrelated tasks [33]. Humans have an inherent ability to transfer knowledge across tasks. Similarly, in recent studies, pretrained models on a large visual dataset (ImageNet) with more than 14 million natural images for visual object recognition can be reused for disease detection via transfer learning. A study by Kermany, et al. [34] demonstrated the competitive performance of deep learning models built with transfer learning in classifying normal eyes and eyes with three macular diseases, using 4,000 optical coherence tomography images. In my previous work, a deep learning system using transfer learning technique in it can accurately differentiate between healthy and glaucomatous subjects with retina image datasets [35]. Meanwhile, regarding the deep learning model using hierarchical classification, in the previous literature, few efforts have been made to leverage in medical field. Nevertheless, hierarchical models have shown better performance compared to flat models in image classification across multiple domains [58], [59]. Sali et al. employed a hierarchical classification model for the classification of gastrointestinal disorders on histopathological images [60].

The objective of this research in Chapter 4 is to develop a training method for building high accurate deep learning models of ophthalmologic disease detection and classification with small size dataset.

The remaining chapter is organized as follows. Section 4.2 presents a method that is based on the characteristics of analyzed diagnostic processes of doctors to build a hierarchical deep learning model with using transfer learning between different level model of the hierarchical models named hierarchy transfer learning. In this chapter, two image datasets are used to demonstrate the proposed approach of hierarchical classification and hierarchy transfer learning to build deep learning models. One is a natural image dataset that is easy to evaluate, which is used for introducing the concept of the approach to build a deep learning model. The other one is a clinical image dataset relevant with an ophthalmologic disease (age-related macular degeneration) to evaluate the training method for building deep learning models by the applicability to medical field. The detailed experimental setups and evaluation results are shown in Section 4.3 and Section 4.4. Finally, concluding remarks are presented in Section 4.5.

This work is based on my previous work of a published journal paper: Naohiro Motozawa, Guangzhou An, et al. (2019) “Optical Coherence Tomography-Based Deep-Learning Models for Classifying Normal and Age-Related Macular Degeneration and Exudative and Non-Exudative Age-Related Macular Degeneration Changes”.

### **4.2. Hierarchical Deep Learning Models Using Hierarchy Transfer Learning**

In this section, a framework of building deep learning models for disease detection and classification is proposed based on single input image, and implemented using hierarchical classification method and hierarchy transfer learning method.

Sketched in upper part of Fig.4.1, first I analyzed doctors' diagnostic processes as follows.

- 1) Doctors perform disease classification after classifying healthy and diseased cases (disease detection), because the disease classification is difficult even for ophthalmologists, which is based on complex symptoms, and it should be performed in the status of normal cases excluded thoroughly.
- 2) Doctors classify diseases into subcategories by reusing the knowledge of classifying normal and diseased cases.

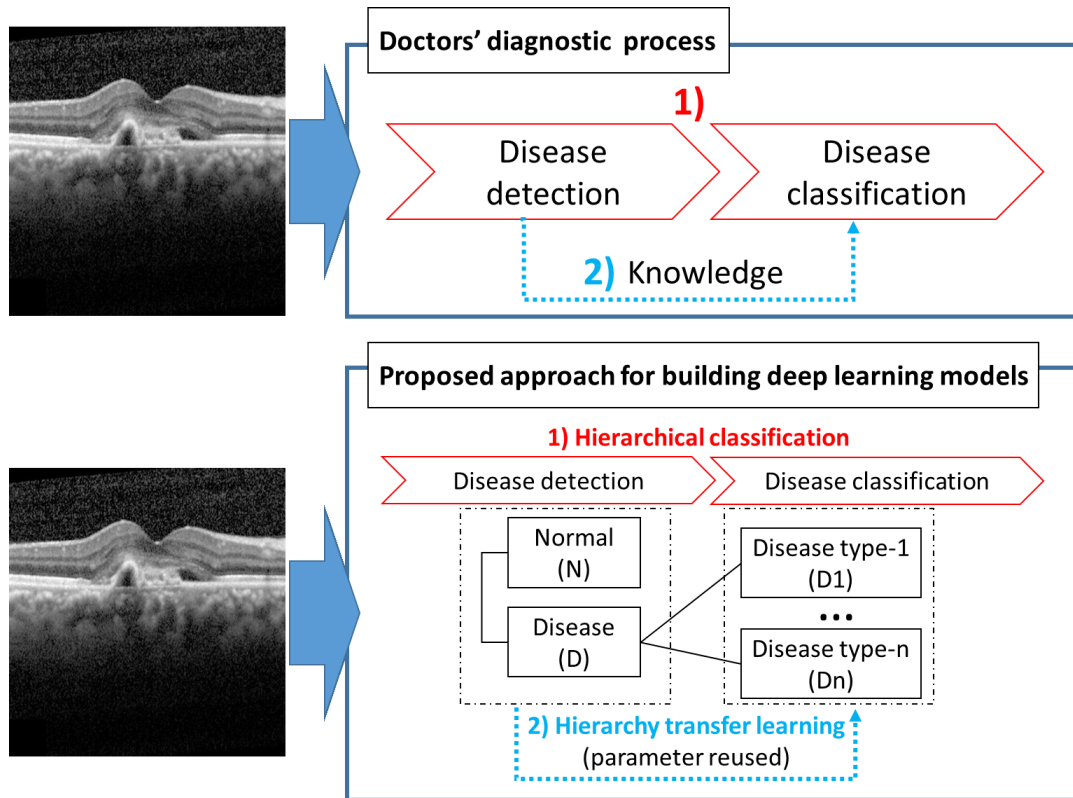


Fig. 4.1. Overview of building deep learning models based on single input image

According to the characteristic of doctors' diagnostic processes above, I proposed a two-step framework to build deep learning models for ophthalmologic disease detection and classification as bottom part of Fig.4.1. In the first step, the deep learning model for classifying healthy and disease (disease detection) is built, and in the second step the model for disease classification is built by reusing parameters of the model for disease detection. To implement this framework in building deep learning models, two deep learning methods are created and applied: one is two-step hierarchical classification method, presented in Chapter 3, and the other one is hierarchy transfer learning method.

#### 1) Hierarchical classification method

As mentioned in Chapter 3, a hierarchical classification method was applied, which is an efficient solution for building classification models with hierarchically structured local classification models according to a predefined hierarchy [36], to separate the training steps of the model classifying normal versus disease cases and the model for

disease classification, as the doctors perform disease classification after excluding normal cases. As the deep learning can extract useful features automatically, the feature selection can be omitted in this hierarchical machine learning models via deep learning.

#### 2) Hierarchy transfer learning method

The second method of hierarchy transfer learning is performed to build disease classification model, reusing the parameters (knowledge) obtained from the model of classifying normal and diseased cases in the previous step of disease detection.

## 4.3. Conceptual Experiments Using Natural Image Dataset

In this section, the description of the experimental setup and data features of the first experiment that is designed to prove the concept of the proposed approach using natural image dataset is shown in detail. Compared with the disease detection, it is difficult to do disease classification into subcategories, thus in this conceptual experiments, two public natural image datasets were found to create a new image dataset to design an image classification task, in which the images have hierarchy relationship to introduce the idea of building deep learning models using hierarchical classification and hierarchical transfer learning method. Two sub-experiments are designed to evaluate the training method, the first one compares the performance of machine learning models trained with different training approaches, whereas the other experiment tries to demonstrate the applicability of the proposed approaches in building high accurate deep learning models using small size training data.

### 4.3.1. Datasets

In this sub-section, the classification task and the details of the natural image dataset are presented.



(a)



(b)

**Fig. 4.2. Two public image datasets to create a new dataset for evaluation.** (a) cats and dogs image dataset on Kaggle site, (b) Stanford dog breed image dataset.

Two public datasets were found and used to create a new image dataset to evaluate the proposed method of building deep learning models. One is cats and dogs image dataset from Kaggle site, that contains 25,000 images of dogs and cats, 12,500 image for each class (<https://www.kaggle.com/c/dogs-vs-cats/data>) (Fig.4.2.a). From this dataset, 1,000 images randomly from the images labelled cats were extracted. The other dataset was the Stanford dog breed image dataset that contains images of 120 breeds of dogs from around the world (<http://vision.stanford.edu/aditya86/ImageNetDogs/main.html>) (Fig.4.2.b). Contents of this dataset is as following: Number of categories: 120, Number of images: 20,580, Annotations: Class labels (dog breed names). From the entire number of images, only the most 4 breeds of

### 4.3 Conceptual Experiments Using Natural Image Dataset

---

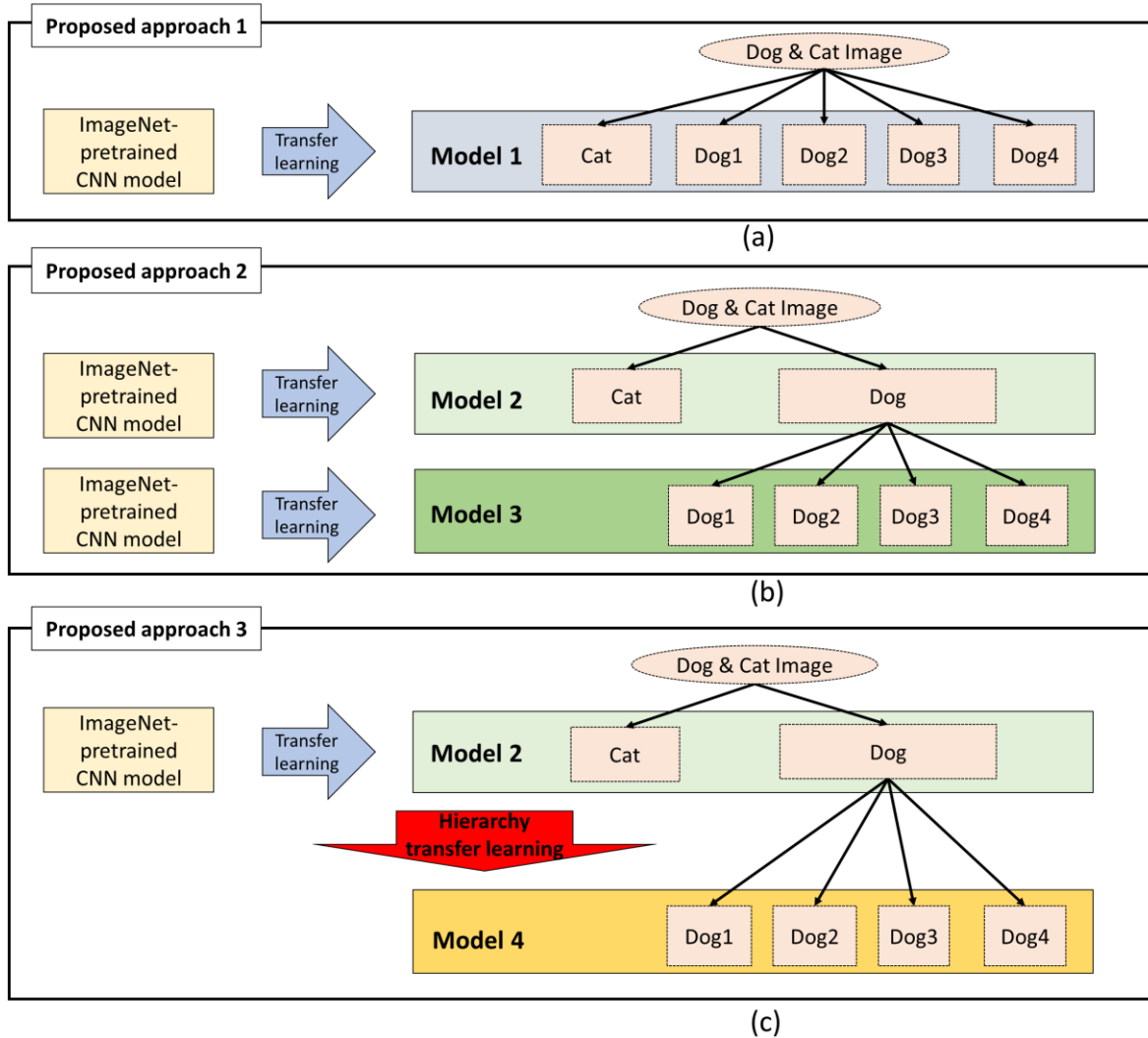
dogs in number were extracted for evaluation, and relabeled as Dog1, Dog2, Dog3, Dog4. The number of each class for the newly created image dataset as shown in Table.4.1.

**Table 4.1. A newly created natural image dataset for conceptual experiments**

| Label            | Cat   | Dog1 | Dog2 | Dog3 | Dog4 |
|------------------|-------|------|------|------|------|
| Number of images | 1,000 | 252  | 239  | 232  | 219  |

#### 4.3.2. Built Deep Learning Models and Training Details

In this sub-section, the built deep learning models trained with different proposed approaches are presented, and training details are shown.



**Fig. 4.3. Proposed approaches to build deep learning models of classifying cats and four breeds of dogs.** (a) Proposed approach 1: flat classification method is used to create Model 1 directly classifying cats and four breeds of dogs with transfer learning from an ImageNet-pretrained CNN model. (b) Proposed approach 2: hierarchical classification method is used to create a low-level model (Model 2) for classifying cats versus dogs and a high-level model (Model 3) for classifying dog breeds; both models apply transfer learning from the ImageNet-pretrained CNN models. (c) Proposed approach 3: the method of hierarchical classification using hierarchy transfer learning between different-level models in the hierarchical classification model is used. Compared with proposed approach 2, the high-level model (Model 4) for classifying dog breeds, transfer learning from the low-level model (Model 2) is used instead of from the ImageNet-pretrained CNN model.

The first experiment is performed as described below, using the entire set of training data to compare classification performances among models trained using three proposed approaches (Fig. 4.3). First, as mentioned in Chapter 2, I applied data augmentation method and transfer learning from the ImageNet pretrained model for training all the deep learning models. Deep learning models are trained separately for natural images of dogs or cats. A convolutional neural network (CNN), which can automatically create efficient image features for the classification, is used as the classifier. Flat classification models are trained for classifying Cat, Dog1, Dog2, Dog3 and Dog4 images directly with the method of transfer learning from a deep learning model (CNN) pretrained on the ImageNet dataset (ImageNet-pretrained CNN model) to build Model 1 (Fig. 4.3.a). A hierarchical classification model (Fig. 4.3.b) is built by applying transfer learning method from the ImageNet-pretrained CNN model separately for a low-level model (Model 2 in Fig. 4.3.b) of classifying dogs and cats and a high-level model (Model 3 in Fig. 4.3.b) for dog breeds classification. As described in the Section 4.2, a hierarchical classification model applies transfer learning from the ImageNet-pretrained CNN model to create a low-level model for classifying dogs versus cats images; then it is hierarchically transferred to build a dog breeds classification model from the low-level model (Fig. 4.3.c).

The second experiment is designed and performed to evaluate the proposed training framework of applicability to small amounts of training data in building high accurate deep learning models by using partial training data. Partial training datasets are created using a stratified random sampling strategy using different percentages of the entire training data (20.0 %, 40.0 %, 60.0%, 80.0% and 100.0 %). With the different training datasets, deep learning models are built with the proposed approaches (Fig. 4.3).

In this study, the VGG16 architecture is adopted for the CNN classifier, which is widely used to solve image classification tasks [61], for all the deep learning models and customized it. Regarding the last two fully connected layers, the units of each layer were changed to 256 with a batch normalization layer and ReLU activation function. The framework of building deep learning models is shown in Fig.4.4, the weights of the feature extractor (green rectangle in Fig.4.4) in VGG16 CNN model for classifying images of 1,000 classes are reused by transfer learning to build deep learning models to classify images of dogs and cats, then the weights of the layers except the softmax layer (red rectangle in Fig.4.4.) were transfer learned (hierarchy transfer learning method) to build deep learning models for the dog breed classification.



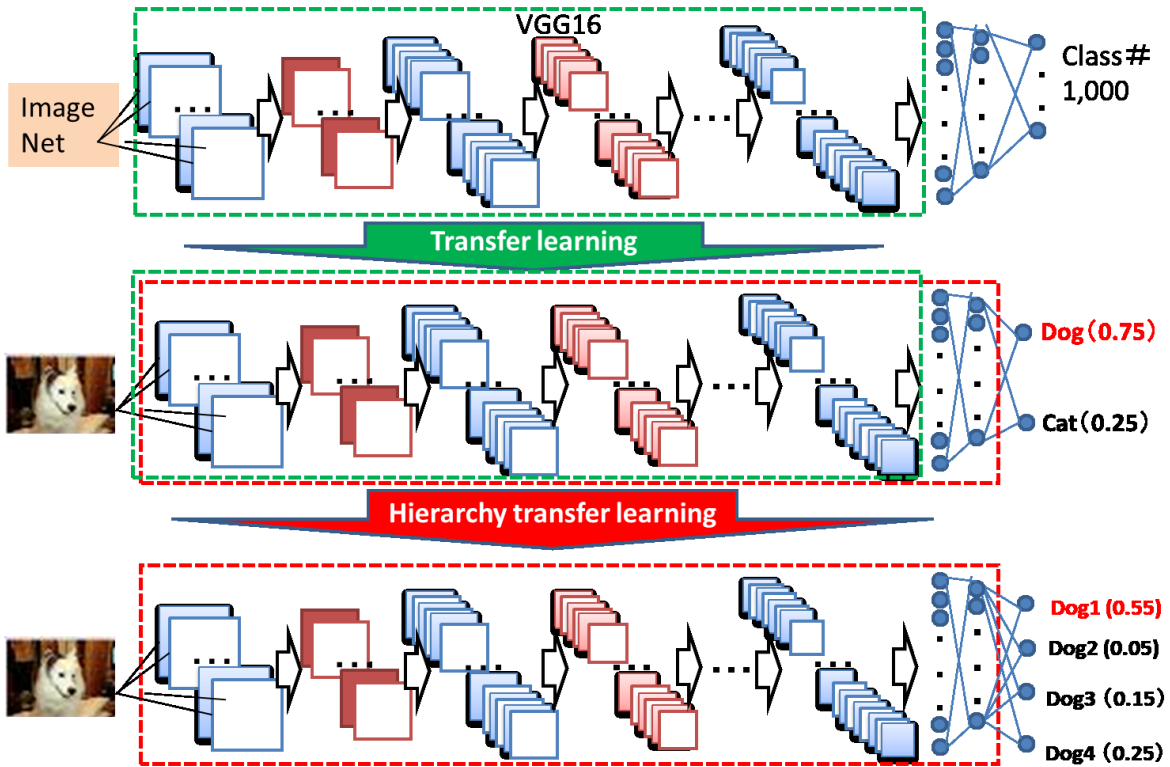


Fig. 4.4. Hierarchy transfer learning in hierarchical deep learning models

For all experiments, the same training setups are applied. Data augmentation techniques are used to improve the classification performance for limited training data, including horizontal flip, random rotation, and random shift. An epoch of 200 was used with a batch size of 32, the optimization method of stochastic gradient descent (SGD) with a learning rate of  $10^{-4}$ , and the weighted categorical cross entropy by the data size of each class as loss function. Finally, the model with the minimum validation loss from 200 deep learning models was selected with early stopping. The experiments are performed using Python 3.6 on an Intel Xeon Gold 6130 @ 2.10 GHz of 32 GB of RAM with a Quadro GV100 (32 GB), using Keras 2.2.4 with TensorFlow 1.13.1.

### 4.3.3. Evaluation Metrics

As an evaluation index of classification, the weighted accuracy was selected to evaluate the

classification performance and applicability of the training framework in small size dataset in building high accurate deep learning models.

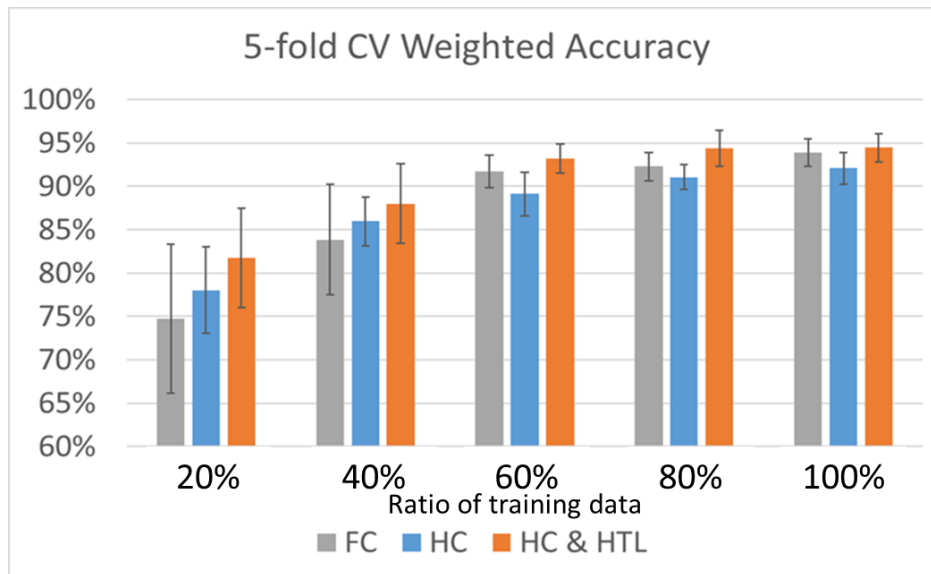
The fraction of all correctly predicted overall number of test set samples was the overall accuracy as Eq. (4.1), where a confusion matrix that comprises false negatives (FN) true negatives (TN), true positives (TP), and false positives (FP).

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) * 100\% \quad (4.1)$$

The weighted accuracy adjusted with the number of data in each class [57] was applied as the evaluation metrics. 5-fold cross validation is performed in this experiment.

#### 4.3.4. Results and Discussion

The weighted accuracy of 5-fold CV for the deep learning models trained with different size dataset by the proposed training framework whether using hierarchical classification strategy or hierarchy transfer learning or not, were shown in Fig. 4.5.



**Fig. 4.5. Classification performance of deep learning models based on natural images with different training methods and different size dataset.** FC: flat classification, HC: hierarchical classification, HC & HTL: hierarchical classification & hierarchy transfer learning

The hierarchical classification strategy with or without hierarchy transfer learning between low-level and high-level models showed improved classification performance compared to flat classification. Proposed approach-3 (hierarchical classification and hierarchy transfer learning) in Fig.4.3 achieved the highest weighted accuracy, 93.9 %. The performance change provided by the deep learning classification models were compared. All the models trained with flat classification (FC; Fig. 4.5.a), hierarchical classification without hierarchy transfer learning (HC; Fig. 4.5.b), and hierarchical classification with hierarchy transfer learning (HC & HTL; Fig. 4.5.c) using a larger training dataset achieved an increased performance. Although there was no significant difference between FC, HC and HC & HTL in the case of a small training dataset. the decrease of classification performance was smaller than FC and HC.

### 4.4. Experiments Using Clinical Image Dataset

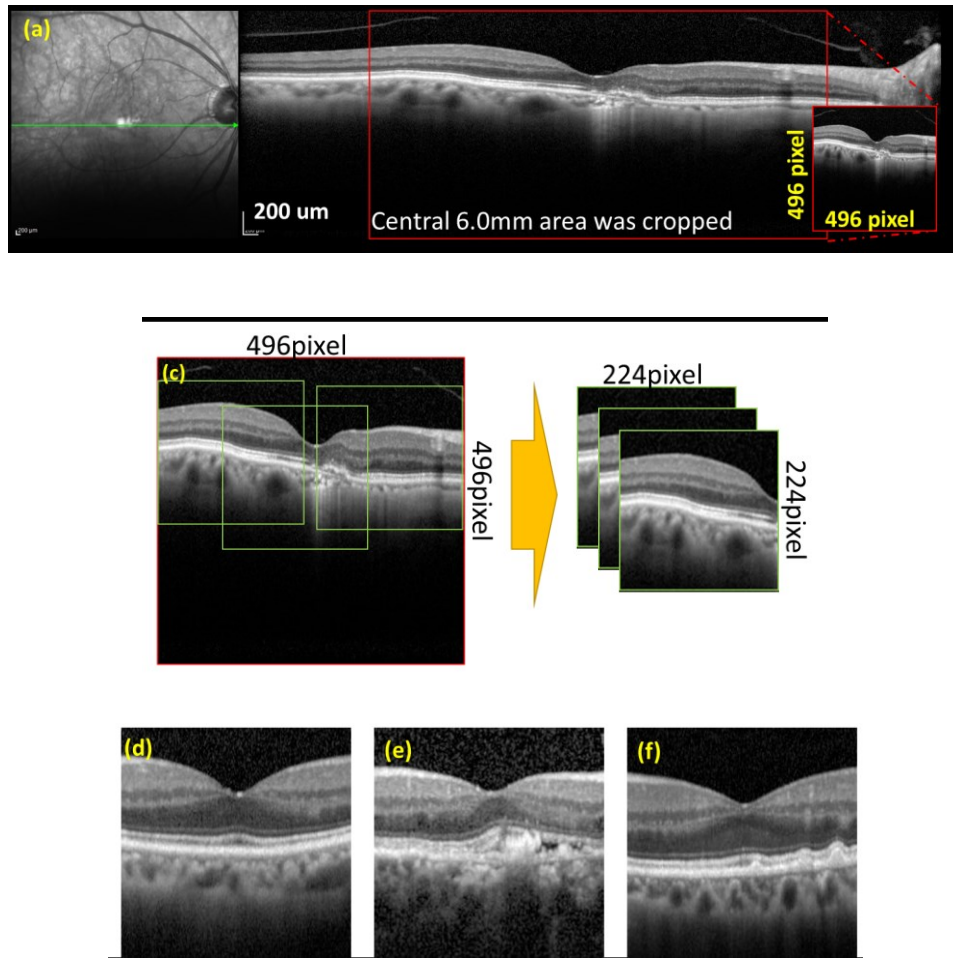
In this section, a clinical dataset is used to demonstrate the proposed approach of hierarchical classification and hierarchy transfer learning to build deep learning models. Two sub-experiments are designed to verify the proposed training methods. One experiment compares the performance of different training approaches, whereas the other evaluates the applicability of the proposed approaches in small size training data.

#### 4.4.1. Datasets

As mentioned in Chapter 3, optical coherence tomography (OCT) is commonly used ophthalmologic instrument which can also be used in clearly describe particular pathology of age-related macular degeneration (AMD) such as drusen, intra-retinal fluid (IRF) sub-retinal fluid (SRF), sub-retinal hyper-reflective material and retinal pigment epithelium detachment [62]. Among them, absence of IRF and/or SRF is the very important interpretation point for most doctors as the therapeutic initiation of anti-VEGF therapy and evaluation of its effect [63]. However, these increased requirements of interpretation on huge amount of OCT data are a big burden for doctors [64]. It is required by the doctors to develop automatic analyzing methods to screen and provide an indication for applying and evaluating the anti-VEGF therapy effect, with classifying normal, AMD with fluid, and AMD without fluid. Recently, machine learning technology especially deep learning has seen dramatic progress, and has

enabled the development of new algorithms for automate diagnosis of eye disease including AMD, glaucoma, diabetic retinopathy with OCT images or fundus photography as input [19], [61], [65]–[69]. In previous papers, there was no classification model reported for classifying normal, AMD with fluid, and AMD without fluid. It is difficult to judge absence of fluid with not obvious differentiation between them.

In this study, I present a notable method of building deep learning classification models with OCT images to classify normal and AMD, and distinguish image of AMD with fluid from AMD without any fluid, together. This study enrolled 120 eyes of 120 AMD patients, and 49 eyes from 49 normal subjects, as training data group, and enrolled another group as test data group, including 77 eyes of 77 AMD patients, and 25 eyes from 25 normal subjects. OCT images of subjects in both groups were captured with Heidelberg Spectralis OCT device, which is spectral domain OCT, in protocol of either radial-scan with 6.0 mm scan length or cross-scan with 9.0 mm scan length. In cross scan images, the central 6.0 mm area were cropped to have the same scan length as radial-scans and resized them into 496\*496 pixels. As a result, there were 185 normal OCT images, 535 OCT images of AMD with fluid, and 514 OCT mages of AMD without fluid in training data, while in test data, there were 49 normal images, 188 AMD OCT images with fluid and 154 AMD images without any fluid. To increase the number of training data, 3 images from each OCT image were cropped from left, middle, right side, with the size of 224\*224pixels, while the vertical position of cropping center is on the RPE line, which is detected automatically. All these cropped images were reviewed by three ophthalmologists independently, labelled as normal, AMD with fluid, and AMD without fluid (Fig. 4.6). Only the images with the same results by the graders were selected for training or validating the classification models, as discussed in Chapter 3. Finally, as training data, 476 normal images, 1,145 images of AMD with fluid and 1,026 mages of AMD without fluid were included (Table 4.2), while in test data, 134 normal images, 402 AMD OCT images with fluid and 347 AMD without any fluid were included (Table 4.3).



**Fig. 4.6. Preprocessing medical image dataset for evaluation of deep learning models built with proposed approaches.** (a) Preprocessing for cross scan OCT images. (b) Preprocessing for radial cross scan OCT images. (c) Cropping from preprocessed OCT images, (d) A sample of cropped OCT image labelled as normal. (e) A sample of cropped OCT image labelled as AMD with fluid. (f) A sample of cropped OCT image labelled as AMD without fluid.

**Table 4.2. Training dataset to build deep learning models**

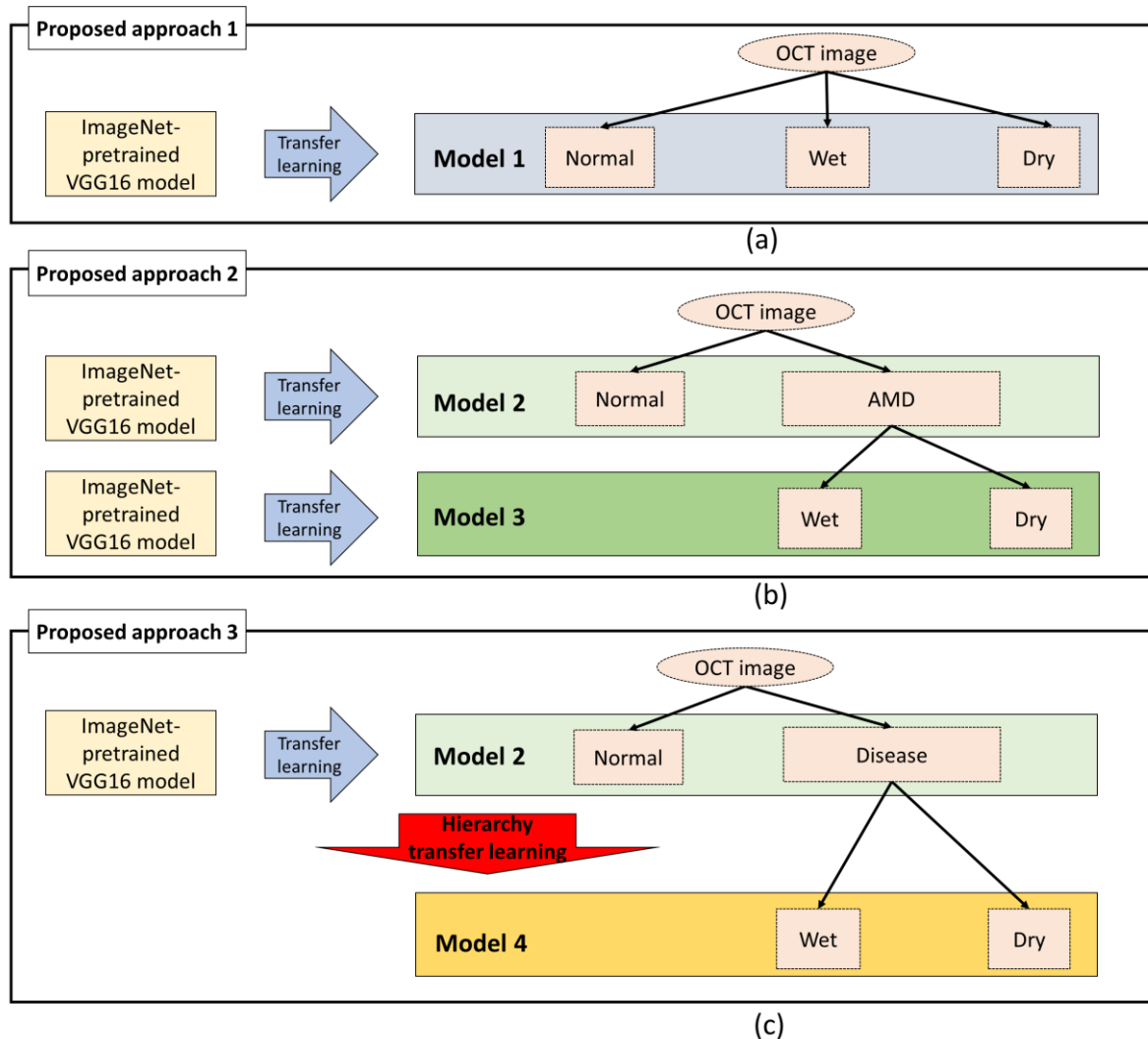
| Training dataset | Normal | AMD  |      |
|------------------|--------|------|------|
|                  |        | wet  | dry  |
| Original images  | 185    | 535  | 514  |
| Cropped images   | 476    | 1145 | 1026 |

**Table 4.3. Test dataset to evaluate deep learning models**

| Test dataset    | Normal | AMD |     |
|-----------------|--------|-----|-----|
|                 |        | wet | dry |
| Original images | 49     | 188 | 154 |
| Cropped images  | 134    | 402 | 347 |

#### **4.4.2. Built Deep Learning Models and Training Details**

In this sub-section, the built deep learning models trained with different proposed approaches are presented, and training details are shown.



**Fig. 4.7. Proposed approaches to build deep learning models of classifying normal, AMD with fluid, (wet), and AMD without fluid (dry).** (a) Proposed approach 1: flat classification method is used to create Model 1 directly classifying normal and AMD eyes with two subtypes with transfer learning from an ImageNet-pretrained CNN model. (b) Proposed approach 2: hierarchical classification method is used to create a low-level model (Model 2) for classifying normal versus AMD cases and a high-level model (Model 3) for classifying wet and dry of AMD; both models in high and low level apply transfer learning from the ImageNet-pretrained CNN model. (c) Proposed approach 3: hierarchical classification method using hierarchy transfer learning method between different-level models is used in the hierarchical classification model. In contrast with proposed approach 2, the high-level model (Model 4) for classifying AMD subtypes, transfer learning method is applied from the low-level model (Model 2) instead of from the ImageNet-pretrained CNN model.

The first experiment is performed as described below, using the entire set of training data to compare classification performances among models trained using three proposed approaches (Fig. 4.7). Deep learning models for discriminating normal images from AMD images are separately trained. A convolutional neural network (CNN), which can automatically create efficient image features for the classification, is used as the classifier. Flat classification methods are used to classify normal eyes and wet AMD and dry AMD directly with transfer learning from a deep learning model (CNN) pretrained on the ImageNet dataset (ImageNet-pretrained CNN model) to create Model 1 (Fig. 4.7.a). A hierarchical classification model (Fig. 4.7.b) is built by applying transfer learning from the ImageNet-pretrained CNN model separately for a low-level model (Model 2 in Fig. 4.7.b) of classifying normal and AMD and for a high-level model (Model 3 in Fig. 4.7.b) of classifying wet AMD and dry AMD. As described in the Section 4.2, a hierarchical classification model applies transfer learning from the ImageNet-pretrained CNN model to create a low-level model for classifying AMD versus normal images; then it is hierarchically transferred to build an AMD classification model from the low-level model (Fig. 4.7.c).

The second experiment is designed to evaluate the training framework's applicability in using small amounts of training data to build deep learning models with the proposed approaches by using partial training data. Partial training datasets are created using a stratified random sampling strategy using different percentages of the entire training data (25.0 %, 50.0 %, 75.0% and 100.0 %). With the different training datasets, deep learning models are built using the proposed approaches (Fig. 4.7) based on OCT image.

In this study, the VGG16 architecture for the CNN classifier is adopted, for all the deep learning models and customized it. Regarding the last two fully connected layers, the units of each layer were changed to 256 with a batch normalization layer and ReLU activation function.

For all experiments, the same training setups are applied. Data augmentation techniques are used to improve the classification performance for limited training data, including horizontal flip, random rotation, and random shift. An epoch of 200 is used with a batch size of 32, the optimization method of stochastic gradient descent (SGD) with a learning rate of  $10^{-4}$ , and the weighted categorical cross entropy by the data size of each class as loss function. Finally, the model with the minimum validation loss from 200 deep learning models is selected with early stopping. The experiments are performed using Python 3.6 on an Intel Xeon Gold 6130 @ 2.10 GHz of 32 GB of RAM with a Quadro GV100 (32 GB), using Keras 2.2.4 with TensorFlow 1.13.1.



### 4.4.3. Evaluation Metrics

For the first experiment, as an evaluation index of classification, the area under receiver operating characteristic curve (AUC) for the ability of VGG16 models was used. A ROC curve (receiver operating characteristic curve) was used to evaluating the classification performance. A Roc curve is graph showing the performance TPR vs. FPR of a classification model at all classification thresholds. This curve plots two parameters, calculated by Eq. (4.2, 4.3, 4.4):

$$\text{True Positive Rate (TPR) = Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \quad (4.2)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (4.3)$$

$$\text{False Positive Rate (FPR)} = 1 - \text{Specificity} = \text{FP} / (\text{TN} + \text{FP}) \quad (4.4)$$

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. AUC is desirable for It measures how well predictions are ranked, rather than their absolute values. AUC is classification threshold invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

Besides AUC, accuracy of judgment for the original image was also examined, judgment of each cropped image was combined. In the first model, if there was an image of AMD of more than one, it was judged as AMD, otherwise it was judged as normal. The judgment of the AMD with fluid is done similarly, if there was more than one image of fluid of 3, it was judged as fluid image, or it was AMD without fluid. To compare the classification performance of classifying AMD and normal, in the third model, considering AMD with fluid and AMD without fluid as AMD category, the test data was input into the model, and calculated on the AUC and accuracy. Furthermore, to compare the ability of classifying AMD with fluid, and AMD without fluid OCT images were input into the third model to calculate the AUC and accuracy.

For the second experiment, weighted accuracy was selected to evaluate the applicability of proposed approach of building deep learning models. The fraction of all correctly predicted overall number of test set samples is the overall accuracy as Eq. (4.5), where a confusion matrix that comprises false negatives (FN) true negatives (TN), true positives (TP), and false positives (FP).

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) * 100\% \quad (4.5)$$

The weighted accuracy adjusted with the number of data in each class was applied as the evaluation metrics [57]. For the second experiments, three randomly sampled datasets, and the mean weighted accuracy on the three test datasets were calculated as the final evaluation metrics.

### 4.4.4. Results and Discussion

In this report, deep learning models were trained with proposed approach (proposed approach-3 in Fig. 4.7.c) based on OCT images to classify normal and AMD, and to distinguish AMD with fluid from AMD without any fluid. For classifying normal versus AMD images, AUC was 0.999, and accuracy was 99.2%. For the classification task of AMD with fluid and AMD without fluid, the classification performances were 0.992 AUC, and 95.1% accuracy. Compared with the model created by the comparison approach (proposed approach-1 in Fig. 4.7.a), in both cases, the proposed approach achieved higher performance (Table 4.4).

**Table 4.4. Classification performance of built deep learning models using entire dataset**

| Classification                                   | Proposed approach |          | Comparison approach |          |
|--|-------------------|----------|---------------------|----------|
|  | AUC               | Accuracy | AUC                 | Accuracy |
| Normal versus AMD (with fluid and without fluid) | 0.999             | 99.2%    | 0.994               | 98.9%    |
| AMD with fluid versus AMD without fluid          | 0.992             | 95.1%    | 0.988               | 94.2%    |

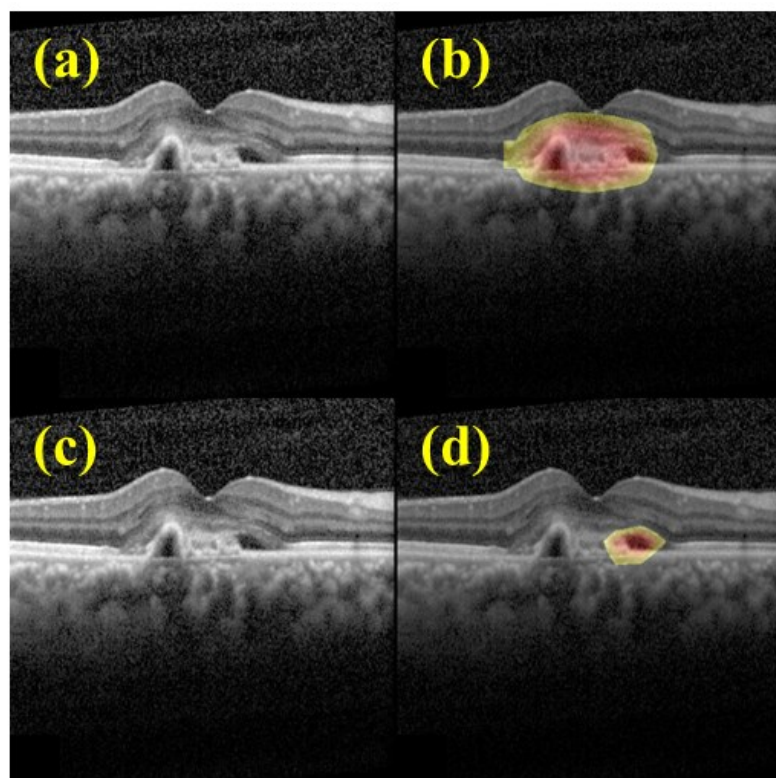
In this study, it was found that a transfer learning of VGG16 was suitable deep learning technique for automate screening of AMD, and judging AMD with fluid or not. In the first model, the performance of classification for normal and AMD OCT images were 0.999 AUC, and 99.2% accuracy. In recent studies' report, deep learning method could achieve high accuracy in screening AMD from normal with OCT images. With OCT images as input, a CNN model pre-trained with ImageNet dataset, was transfer learned with 1,012 B-Scan OCT images was able to distinguish normal from AMD images with 96% accuracy. The result was consistent with recent reports [70]. Data shuffling was applied twice on sum of training and

test data group, to create two different sets of training and test data, to evaluate performance stability of the models created by the proposed approach. In the additional experiments, the AUCs of the models to classify normal versus AMD (with fluid and without fluid) using the proposed approach, were both over 0.999, and were better than the AUCs of the models created with the comparison approach.

In the second model, AMD with fluid from AMD without any fluid could be distinguished, with 0.992 AUC, and 95.1% accuracy. Detecting the presence of fluid at macular is the most important point for many ophthalmologists in therapeutic indication. Treatment of anti-VEGF injection therapy and redosing criteria depend on the whether there is IRF and/or SRF or not. Recent reports have shown that the amount of fluid can be quantified accurately and clearly recognizing the differentiation using deep learning method, and classify indicators of fluid in OCT images, which are key points for initial and anti-VEGF therapy decisions in AMD with 92% sensitivity, 91% specificity and 93% accuracy. The result was consistent with this report. In the second model, the transfer learning was applied. In transfer learning, high accuracy result with fewer dataset could be obtained. It is considered to be effective when image data is limited, especially in clinic data. However, in most cases, deep learning requires huge amount of data. Recently, effectiveness of building deep learning models with transfer learning from pretrained ones has been reported [70]. Similar with the first model, data shuffling was applied twice on sum of training and test data group, to create two different sets of training and test data, to evaluate performance stability of the models created by the proposed approach. In the additional experiments, the AUCs of the models to classify AMD with fluid and AMD without fluid using the proposed approach, were both higher than 0.980, and were better than the AUCs of the models created with the comparison approach.

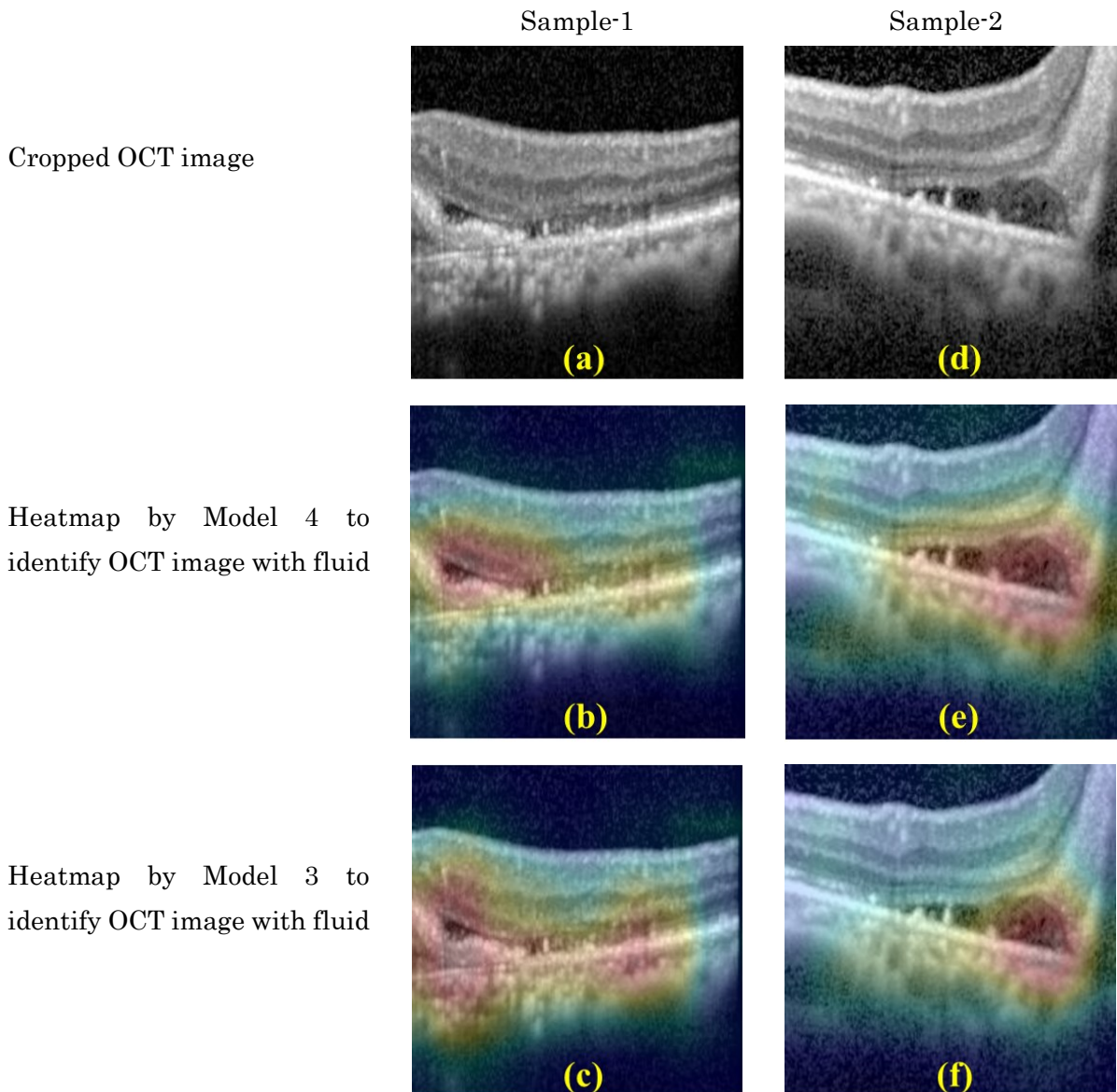
In practice, deep learning models are treated as a black-box method, recently there are several method including the one proposed in the paper of Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, to help debug deep learning models [71]. This method can be used to visually debug where a CNN model is looking in an image. Grad-CAM works by (1) finding the final convolutional layer in the network and then (2) examining the gradient information flowing into that layer. The output of Grad-CAM is a heatmap visualization for a given class label. In this dissertation, the Grad-CAM method was applied to find the models observe which area is important for VGG16 model to judge to perform qualitative evaluation. It was possible for the doctors to confirm the important area of CNN classification models on each image. In detail, a heatmap for a classic AMD category indicating the effective region for the model to identify AMD was generated, while a heatmap was created

with the discriminative region used by the second model to identifying wet AMD. Against a same OCT image, compared with heatmap created by the first model, the one created by the second model, the important area is just on the fluid, which is a quite important point to classify whether there is fluid or not in an OCT image (Fig. 4.8). It is considerable that with hierarchy transfer learning the weights of the CNN model tuned to the new classification task properly.



**Fig. 4.8. Important areas for the deep learning models trained with the proposed approach.** A classic AMD with SRF was selected to confirm the heatmap for the VGG16 models. (a) Cropped AMD OCT image (b) Heatmap created by the first model (Model 2) to classify AMD and normal OCT images. (c) Cropped AMD OCT image, same as (a). (d) Heatmap created by the second model (Model 4) to classify AMD with fluid and AMD without fluid OCT images. Red regions correspond to high score for the classification.

Furthermore, the heatmaps by Model 4 (proposed approach) and model 3 (comparison approach) to identify AMD with fluid or not was compared. The heatmaps by Model 4 is more sensitive to find the effective area for the true classification. In the cropped OCT image of sample-1, Model 4 was succeed to narrow the important area to the actual fluid area in OCT image, while in the cropped OCT image of sample-2, Model 4 expanded the important area to the fluid area in OCT image, which might help to get a higher classification performance to classify the AMD with fluid and the AMD without fluid (Fig.4.9). It is considerable that with hierarchy transfer learning from a model pretrained with a similar domain decreased the difficulty of optimization, thus it succeed in finding more powerful features to achieve high accuracy.

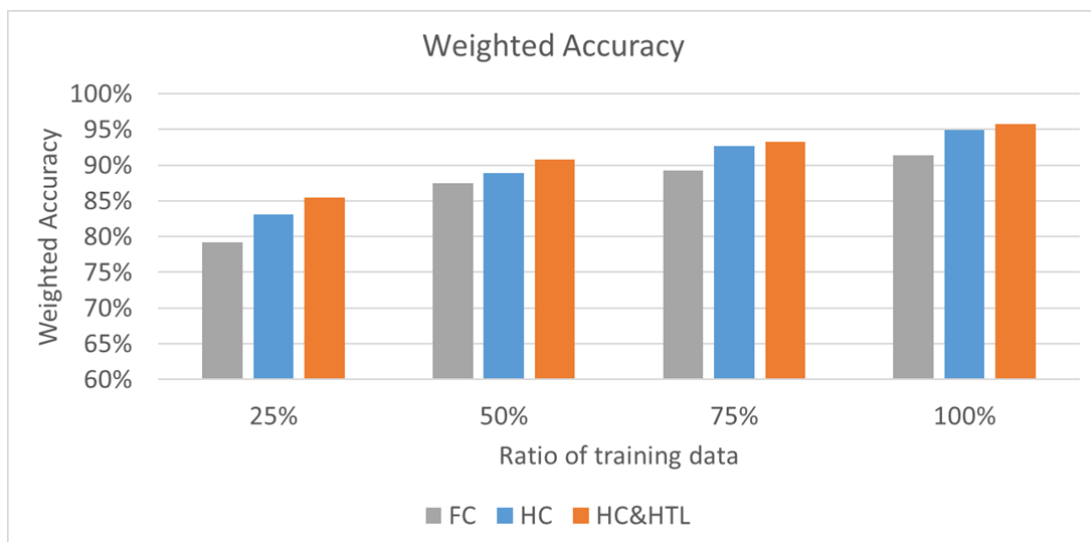


**Fig. 4.9. Differentiation of heatmaps created by built deep learning models to identify AMD OCT image with fluid.** Two cropped OCT images were randomly selected to show the differentiation of heatmaps created by the VGG models to classify AMD with fluid and AMD without fluid, compared with the built model by the comparison approach. Red regions correspond to high score for the classification. (a) Cropped AMD OCT image with fluid from sample-1, (b) Heatmap created by Model 4 (proposed approach) for (a) to identify AMD with fluid, (c) Heatmap created by Model 3 (comparison approach) for (a) to identify AMD with fluid, (d) Cropped AMD OCT image with fluid from sample-2, (e) Heatmap created by Model 4 (proposed approach) for (d) to identify AMD with fluid, (f) Heatmap created by Model 3 (comparison approach) for (d) to identify AMD with fluid.

#### 4.4 Experiments Using Clinical Image Dataset

---

In the second experiments, the hierarchical classification strategy with or without hierarchy transfer learning between low-level and high-level models showed improved classification performance compared to flat classification. Proposed approach-3 (hierarchical classification and hierarchy transfer learning) achieved the highest weighted accuracy, 93.9 %. The performance change provided by the deep learning classification models were compared. All the models trained with flat classification (FC; Fig. 4.7.a), hierarchical classification without hierarchy transfer learning (HC; Fig. 4.7.b), and hierarchical classification with hierarchy transfer learning (HC & HTL; Fig. 4.7.c) using a larger training dataset achieved an increase. Although there was no significant difference between FC, HC and HC & HTL in the case of a small training dataset, the decrease of classification performance was smaller than FC and HC (Fig.4.10).



**Fig. 4.10. Classification performance for deep learning models classifying normal, AMD with fluid, and AMD without fluid, with different training methods and different size dataset**

The training time for FC, HC and HC & HTL, and the average prediction time per image was measured, as shown in Table.4.5. The training time for HC & HTL and HC are longer than FC, but with an affordable time. In other words, just spending more time at the initial training, the proposed approach can achieve higher accuracy, with the prediction time difference is not obvious. It is found that, with HTL the training time for hierarchical classification decreased, it might due to start training from pre-trained model for disease detection, some features useful for disease classification have already been learnt.

**Table 4.5. Training and prediction time for each deep learning model**

| Training method | Training time<br>(seconds) | Prediction time<br>(milli-seconds per image) |
|-----------------|----------------------------|--|
| FC              | 1,789                      | 168.6  |
| HC              | 2,363                      | 183.6  |
| HC & HTL        | 2,072                      | 183.3  |

## 4.5. Conclusions

In this chapter, deep learning techniques are applied for handling medical images directly without feature engineering to build machine learning models for ophthalmologic disease detection and classification. A two-step framework, first step of which builds the model for disease detection, and the second step of which builds the model for disease classification by reusing parameters of the model for disease detection, is proposed to build deep learning models for ophthalmologic disease detection and classification. In detail, besides the two-step hierarchical classification method, hierarchy transfer learning method was created and used to build deep learning models.

In experimentations, firstly, I introduced and evaluated the concept of the proposed method of building deep learning models with a labelled natural image dataset that is easy to evaluate, in which the labels of the images have hierarchy relationship. The classification performance of the models built with the proposed approach were compared to the flat models built in one-step to detect and classify diseases, trained with different size datasets random sampled from the dataset above. As a result, compared with flat models, the effectiveness of



improved classification performance for the model built with the method of combining hierarchical classification and hierarchy transfer learning method was demonstrated even for the smaller size datasets. This approach also succeeded in building a high accurate deep learning model with labelled clinical medical image dataset for an ophthalmologic disease (age-related macular degeneration) detection and classification. Similar with the result of experiments using the natural image dataset above, the proposed approach enabled machine learning models to achieve high accuracies and improved classification performance in comparison with flat models, the effect and efficiency have been demonstrated even in smaller size datasets. As mentioned in Chapter 2, generally speaking, training a deep learning model using image data for high accuracy is more difficult than training a traditional machine learning model based on quantified parameters. With hierarchy transfer learning, the classification performance was improved from the one only used hierarchical classification method. It is believed that the combination method of hierarchical classification and hierarchy transfer learning can be applied in solving much easier problem of building the machine learning models based on quantified parameters for high accuracy with limited data, and this validation should be done in the future work.

In this work, two major contributions were made: (1) The effectiveness of building machine learning models for disease detection and classification hierarchically can improve the classification performances when using limited size dataset was demonstrated: first classifying normal and diseased cases, then classifying diseased cases into sub-categories was demonstrated. (2) The effectiveness of building high level models in hierarchical classification models for disease classification by reusing low-level models in the hierarchical classification models for disease detection was demonstrated.

Conclusively, the proposed two-step hierarchical framework has a high potential of deploying high accurate machine learning models for ophthalmologic disease detection and classification with small size labelled single kind of input image dataset.

## **Chapter 5.**

# **Stacked Hierarchical Deep Learning Models Using Hierarchy Transfer Learning Based on Multiple Input Images**

### **5.1. Overview**

In the foregoing attempt, usage of hierarchical classification and hierarchy transfer learning to build deep learning models for disease detection and classification was suggested. In this chapter, I present an extension method for the proposed approach of hierarchical deep learning classification and hierarchy transfer learning shown in Chapter 4, to handle multiple input images for higher accuracies. In detail, the extension was implemented with a stacking ensemble method combining classification models built with each type of input images.

The remaining chapter is organized as follows. Section 5.2 briefly gives an overview of the proposed approach of building deep learning models for disease detection and classification using hierarchical classification, hierarchy transfer learning and stacking method. Finally, experimental evaluations and concluding remarks are respectively presented in Section 5.3 and Section 5.4.

This work is based on my previous work of the journal paper: Guangzhou An, et al. (2019) “Deep Learning Classification Models Built with Two-step Transfer Learning for Age Related Macular Degeneration Diagnosis”, 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

## 5.2. Stacked Hierarchical Deep Learning Models Using Hierarchy Transfer Learning

In this section, a two-step framework using hierarchical classification and hierarchy transfer learning was extended by stacking ensemble method to build deep learning models based on multiple input images for disease detection and classification.

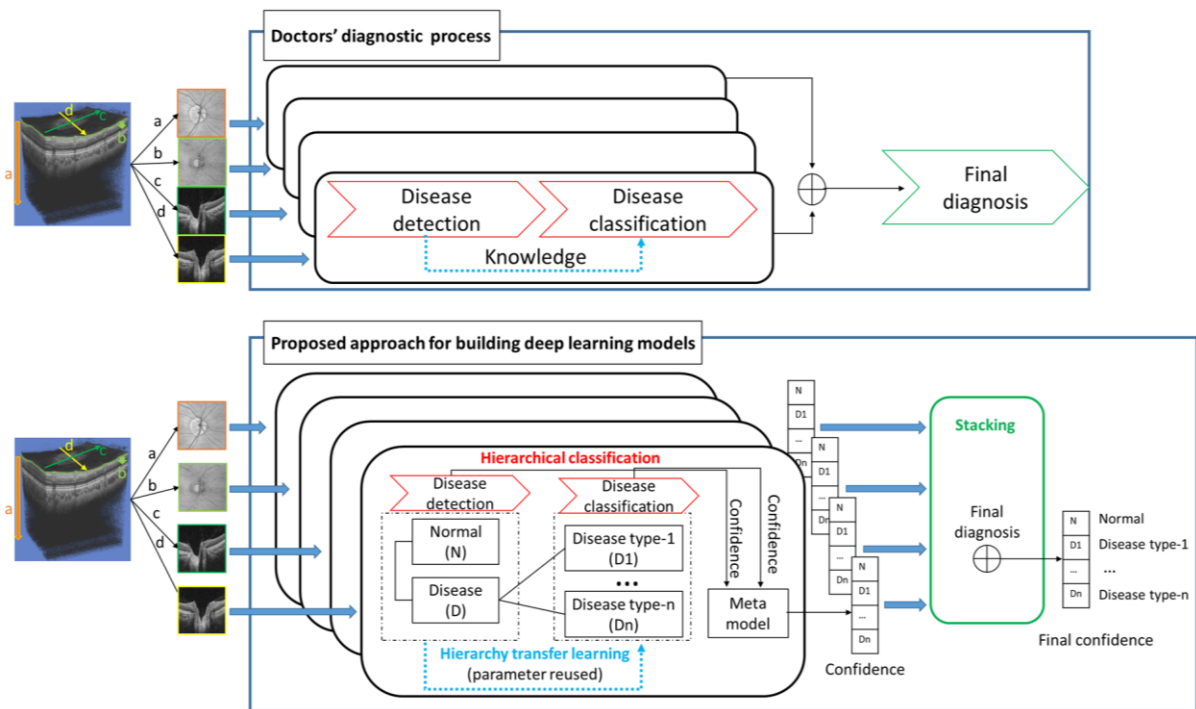


Fig. 5.1. Overview of building deep learning models based on multiple input images

Sketched in upper part of Fig.5.1, I analyzed doctors' diagnostic processes as follows. Doctors first use limited information (e.g., a single type of medical image) to detect disease. The determination of a treatment plan is difficult, as it requires disease classification based on complex symptoms, and it must be performed after disease detection (excluding normal images). Second, doctors reuse the knowledge of classifying normal and diseased cases to classify diseased cases into subcategories. Finally, for accurate diagnosis, doctors use multiple information sources (e.g., multiple types of medical images) to make optimal diagnostic decisions.

According to the characteristic of doctors' diagnostic processes above that I analyzed, I proposed a two-step framework to build machine learning models for ophthalmologic disease detection and classification. In the first step, the machine learning model for classifying healthy and disease (disease detection) was built, and in the second step the model for disease classification was built by reusing parameters of the model for disease detection. Moreover, this two-step framework was extended to handle multiple input data. To implement this framework in building machine learning models, three machine learning methods were created with separate roles as below (Fig.5.1).

1) Hierarchical classification method, to build machine learning models for disease classification, after building the models for classifying healthy and disease (disease detection)

2) Hierarchy transfer learning method, to build machine learning models for disease classification by reusing parameters of the model for disease detection

3) Stacking ensemble method, to build machine learning models handling multiple input data by combining the machine learning models trained separately on single input image with the method combining hierarchical classification and hierarchy transfer learning

For disease detection, transfer learning from a pretrained model on a large visual database (ImageNet) with more than 14 million natural images for visual object recognition was used. A metamodel was then used to combine the results of the model classifying normal versus disease cases and model for disease classification using a single type of input images. Finally, a stacking ensemble method was used to combine the separate deep learning models to obtain the overall result (Fig. 5.1).

### 5.3. Experiments Using Clinical Image Dataset

In this section, a clinical dataset of medical images was used to demonstrate the proposed approach to build deep learning models, implemented with hierarchical classification method, hierarchy transfer learning method and stacking ensemble method.

#### 5.3.1. Datasets

As mentioned in Chapter 3, in this chapter I tried to build deep learning models relevant with

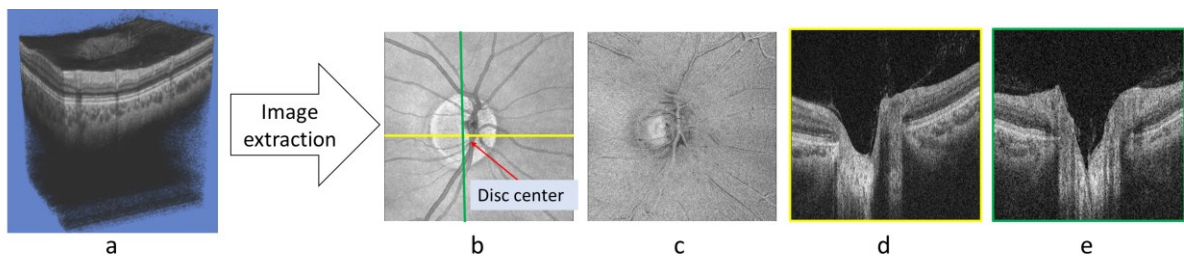
### 5.3 Experiments Using Clinical Image Dataset

glaucoma diagnosis. Glaucoma is a leading cause of blindness worldwide, and glaucoma-related blindness is irreversible if not detected early and treated appropriately [71]. Glaucoma is regarded as a multifactorial disease, and some ophthalmologists suggest that treatment ought to be categorized by its cause [37]. A published guideline that defines four types of retinal optic discs based on their morphology [38] is useful in understanding the pathology of glaucoma. Follow-up investigations also showed that this guideline is a useful addition to the determination of a proper treatment plan in glaucoma management [39], [40], [72]. Classification is difficult as it is based on subjective assessments of medical images [51].

Images of 156 normal and 798 glaucomatous eyes were obtained. These images were reviewed and labelled by two glaucoma specialists with more than 10 years of experience. The glaucomatous eyes were further classified into four sub-categories based on the optic disc morphology according to the definitions used in Nicolela classification as focally ischemic (FI), myopic glaucomatous (MY), generalised enlargement (GE), and senile sclerotic (SS) discs. Images with discordant classifications between the two glaucoma specialists were excluded, as discussed in Chapter 3. A total of 156 normal, 118 FI, 266 GE, 307 MY, and 107 SS eyes were used (Table 5.1).

**Table 5.1. Number of collected data for evaluation**

|                             | Normal | Glaucoma |     |     |     |
|-----------------------------|--------|----------|-----|-----|-----|
|                             |        | FI       | GE  | MY  | SS  |
| Total data number<br>n= 954 | 156    | 118      | 266 | 307 | 107 |



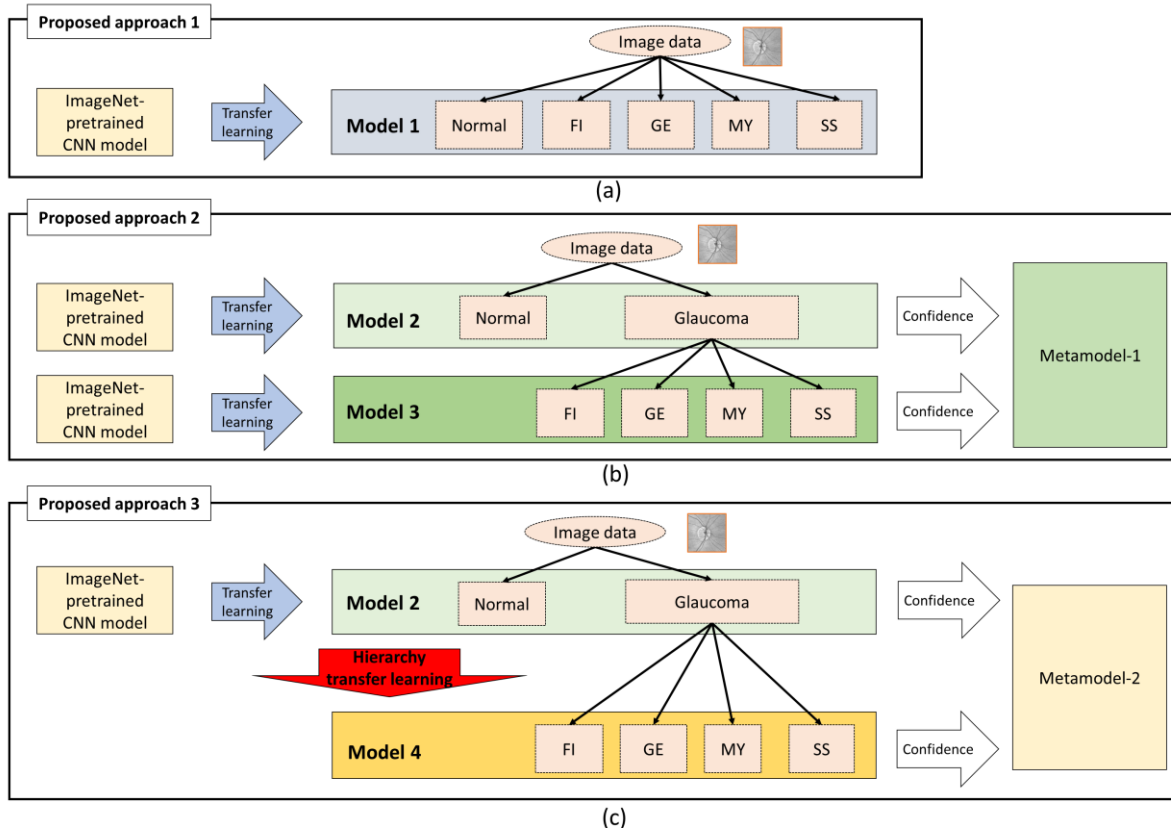
**Fig. 5.2. Image extraction from the volumetric OCT data.**

These eye images were captured using swept source OCT (DRI OCT Triton), and a three-dimensional (3D) scan of the disc region (6.0 mm x 6.0 mm) was scanned horizontally consisting of 256 B-scans with 512 axial depth scans (A-scans) each. The disc centers were detected, and vitreous/inner limiting membrane boundaries were segmented automatically with Topcon's commercial analysis software (FastMap V.10.13). The verification was performed by glaucoma specialists. Based on the position of the automatically detected disc centers and segmented vitreous/inner limiting membrane boundaries, four types of images were extracted from the 3D disc scan OCT data and used in the machine learning system. The four images, which the doctors used to analyse the 3D data for glaucoma diagnosis, are as follows: 1) projection images, integration performed across the entire image depth of 2.6 mm (image 'a' in Fig. 5.2); 2) en face images, integration performed across a fixed thickness of 52  $\mu\text{m}$  (20 voxels) below the vitreous/inner limiting membrane boundary (image 'b' in Fig. 5.2); 3) horizontal B-scan OCT images crossing the disc center (image 'c' in Fig. 5.2; disc H B-scan); 4) vertical B-scan OCT images crossing the disc center (image 'd' in Fig. 5.2; disc V B-scan).. The region of a vertical length of 512 pixels was cropped from the disc H B-scan and disc V B-scan based on the average vitreous/inner limiting membrane's axial position of the entire OCT data. The region of a vertical length of 512 pixels was cropped from the disc H B-scan and disc V B-scan based on the average vitreous/inner limiting membrane's axial position of the entire OCT data. Finally, all the four images types were resized into  $256 \times 256$  pixels and normalized to the range of 0 to 1.

#### 5.3.2. Built Deep Learning Models and Training Details

Two experiments were designed to verify the methods. One experiment compared the performance of deep learning models trained with different proposed approaches, whereas the other evaluated the applicability of the proposed approaches in small size training data.

### 5.3 Experiments Using Clinical Image Dataset



**Fig. 5.3. Proposed approaches for building deep learning models with single input images separately before applying stacking for the overall result.** (a) Proposed approach 1: flat classification method is used to directly classify normal eyes and those with four subtypes of disease with transfer learning from an ImageNet-pretrained CNN model to create Model 1. (b) Proposed approach 2: hierarchical classification method is used to create a low-level model (Model 2) for classifying normal versus disease cases and a high-level model (Model 3) for classifying subtypes of disease; both models apply transfer learning from the ImageNet-pretrained CNN model. The confidence in a “normal” result from Model 2 and the confidence in disease subtypes from Model 3 are concatenated to calculate the overall result from training Metamodel 1. (c) Proposed approach 3: hierarchical classification method using hierarchy transfer learning method between different-level models is used in the hierarchical classification model is. In contrast with proposed approach 2, the high-level model (Model 4) for classifying disease subtypes, transfer learning is from the low-level model (Model 2) instead of from the ImageNet-pretrained CNN model. The normal confidence from Model 2 and the disease subtype confidence from Model 4 are concatenated to train Metamodel 2 to calculate the overall result.

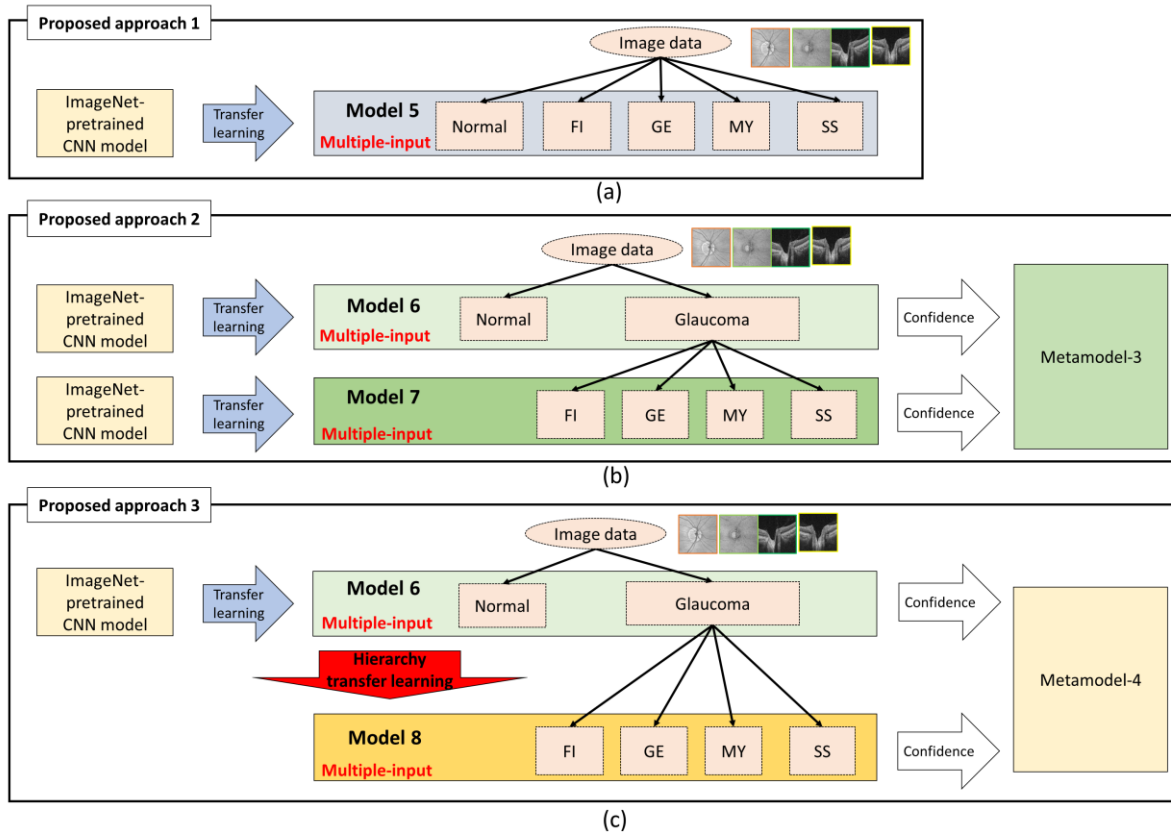
The first experiment was performed as described below, using the entire set of training data to compare classification performances among models trained using three proposed approaches (Fig. 5.3). Deep learning models were trained separately for each type of extracted image from the 3D OCT data. A convolutional neural network (CNN), which can automatically create efficient image features for the classification, was used as the classifier. Flat classification models were used to classify eyes as normal, FI, GE, MY, or SS directly with transfer learning from a deep learning model (a CNN) pretrained on the ImageNet dataset (ImageNet-pretrained CNN model) to create Model 1 (Fig. 5.3.a). A hierarchical classification model (Fig. 5.3.b) was built by applying transfer learning from the ImageNet-pretrained CNN model separately with a low-level model (Model 2 in Fig. 5.3.b) for classifying normal versus glaucoma and a high-level model (Model 3 in Fig. 5.3.b) for glaucoma classification. Furthermore, the normal confidence from Model 2 and the confidence of FI, GE, MY, and SS from Model 3 were concatenated into a confidence vector length of 5. Then, a metamodel of the linear support vector machine (SVM) was trained using the confidence vector data with the supervised labels to combine the models in a cascaded manner [73]. As described in the Section 5.2, a hierarchical classification model applies transfer learning from the ImageNet-pretrained CNN model to create a low-level model for classifying normal versus glaucoma cases; then it is hierarchically transferred to build a glaucoma classification model from the low-level model. Finally, a metamodel of the linear SVM was used to calculate the overall result (Fig. 5.3.c).

A stacking ensemble method was used to combine the separately trained single-input image models. The confidence vector calculated by models trained with different single-input images was extracted and concatenated to train the superior metamodel via a linear SVM to combine the single-input models. For comparison with the stacking method, a multiple-input CNN with a direct four-image input was selected to handle multiple images, and trained with the three proposed approaches (Fig. 5.4). A metamodel was also used to combine the results of different-level models to calculate the final classification result.

In this study, the VGG16 CNN architecture was adopted, which is widely used to solve image classification tasks [61], for all the deep learning models and customized it. For single-input CNN models, a CNN architecture VGG16 was customized by adding batch normalization after each convolutional layer to accelerate training as the model classifier. Regarding the last two fully connected layers, the units of each layer were changed to 256 with a batch normalization layer and ReLU activation function.

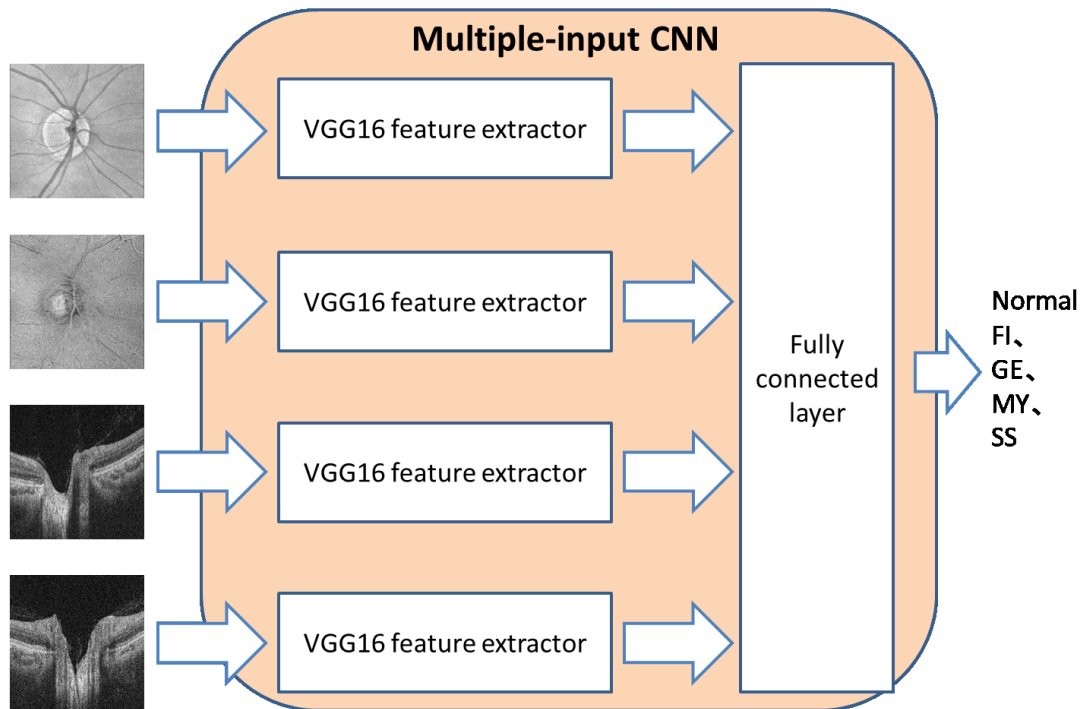


### 5.3 Experiments Using Clinical Image Dataset



**Fig. 5.4. Proposed approaches for building deep learning models with multiple input images directly.** (a) Proposed approach 1: flat classification method is used to directly classify normal eyes and those with four subtypes of disease with transfer learning from an ImageNet-pretrained CNN model to create Model 5. (b) Proposed approach 2: hierarchical classification method is used to create a low-level model (Model 6) for classifying normal versus disease cases and a high-level model (Model 7) for classifying subtypes of disease; both models apply transfer learning from the ImageNet-pretrained CNN model. The confidence in a “normal” result from Model 6 and the confidence in disease subtypes from Model 7 are concatenated to calculate the overall result from training Metamodel 3. (c) Proposed approach 3: hierarchical classification using hierarchy transfer learning between different-level models is used in the hierarchical classification model. In contrast with proposed approach 2, the high-level model (Model 8) for classifying disease subtypes, transfer learning is from the low-level model (Model 6) instead of from the ImageNet-pretrained CNN model. The normal confidence from Model 6 and the disease subtype confidence from Model 8 are concatenated to train Metamodel 4 to calculate the overall result.

For comparison, a deep learning models handling multiple ( $n=4$ ) input images directly (multiple input CNNs) was applied as the classifier. Flat classification models were used to classify eyes as normal, FI, GE, MY, or SS directly with transfer learning from a deep learning model (a multiple input CNN) pretrained on the ImageNet dataset (ImageNet-pretrained CNN model) to create Model 5 (Fig. 5.4.a). A hierarchical classification model (Fig. 5.4.b) applying transfer learning from the ImageNet-pretrained CNN model separately with a low-level model (Model 6 in Fig. 5.4.b) was built for classifying normal versus glaucoma and a high-level model (Model 7 in Fig. 5.4.b) for glaucoma classification. Furthermore, the normal confidence from Model 6 and the confidence of FI, GE, MY, and SS from Model 7 were concatenated into a confidence vector length of 5. Then, a metamodel of the linear support vector machine (SVM) was trained using the confidence vector data with the supervised labels to combine the models in a cascaded manner [73]. As described in the Section 5.2, a hierarchical classification model applies transfer learning from the ImageNet-pretrained CNN model to create a low-level model for classifying normal versus glaucoma cases; then it is hierarchically transferred to build a glaucoma classification model from the low-level model. Finally, a metamodel of the linear SVM was used to calculate the overall result (Fig. 5.4.c). For multiple-input CNN models, each input had the same feature extractor with the single-input CNN in the proposed approach of the weights pre-trained on ImageNet to extract features from the layer before the first fully connected layer (Fig.5.5). These were then concatenated and fed to newly created, two fully connected layers (both were 256 units with batch normalization layer and ReLU activation function). In all classification models, only the unit number of the softmax layer was changed according to the class number of each classification task.



**Fig. 5.5. Multiple input CNN models for glaucoma detection and classification**

The second experiment was designed to evaluate the deep learning models' applicability to small amounts of training data built using the proposed approach by using partial training data. Partial training datasets were created using a stratified random sampling strategy using different percentages of the entire training data (25.0 %, 37.5 %, 50.0 %, 62.5 %, 75.0 %, 87.5 %, and 100.0 %). With the different training datasets, deep learning models were built with the proposed approaches (Fig. 5.3) using one type of input image (projection image) and a combination model via a stacking ensemble method for different single-input models.

For all experiments, the same training setups were applied. Data augmentation techniques were used to improve the classification performance for limited training data, including horizontal flip, random rotation, and random shift. An epoch of 100 was used with a batch size of 32, the optimization method of stochastic gradient descent (SGD) with a learning rate of  $10^{-3}$ , and the weighted categorical cross entropy by the data size of each class as loss function. Finally, the model with the minimum validation loss from 100 deep learning models was selected with early stopping. The experiments were performed using Python 3.6 on an Intel Xeon Gold 6130 @ 2.10 GHz of 32 GB of RAM with a Quadro GV100 (32 GB), using

Keras 2.2.4 with TensorFlow 1.13.1.

### 5.3.3. Evaluation Metrics

The entire dataset was shuffled to create three different training (80%) and test (20%) datasets with a stratified sampling strategy. The training dataset was used to build the classification models, and the test dataset was used to evaluate the models. The average of the classification indexes for the three test datasets was used to evaluate the proposed approaches. The dataset was not balanced in class distribution; thus, weighted accuracy and Cohen's kappa were used to evaluate the deep learning models.

The fraction of all correctly predicted overall number of test set samples is the overall accuracy as Eq. (5.1), where a confusion matrix that comprises false negatives (FN) true negatives (TN), true positives (TP), and false positives (FP).

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) * 100\% \quad (5.1)$$

The weighted accuracy corresponds to the correctly detected samples divided by the total number of samples [57].

The Cohen's kappa [46] can be calculated according to the Eq. (5.2).

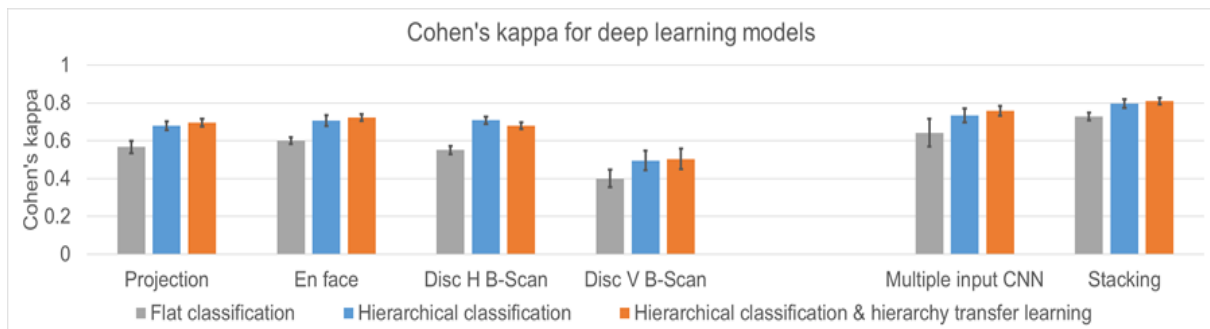
$$\text{Cohen's kappa} = \frac{N \sum_{i=1}^m CM_{ii} - \sum_{i=1}^m C_{i\text{corr}} C_{i\text{pred}}}{N^2 - \sum_{i=1}^m C_{i\text{corr}} C_{i\text{pred}}} \quad (5.2)$$

where  $CM_{ii}$  represents the diagonal elements of the confusion matrix,  $C_{i\text{corr}}$  represents the number of data labelled as  $C_i$ , and  $C_{i\text{pred}}$  represents the number of data predicted by the machine learning model as  $C_i$ .  $N$  is the total number of cases. Details of the calculation method of Cohen's kappa was shown in Section 3.3.3.

The agreement thresholds used in this study were based on the first guideline: good: 0.61–0.80 and almost perfect: 0.81–1 [46]. A paired t-test P-value (two-tailed) of 0.05 as the significance level was used to determine whether the mean difference between two sets of Cohen's kappa was zero.

### 5.3.4. Results and Discussion

The deep learning models of single-input images built with the flat classification strategy and transfer learning from the ImageNet-pretrained model (proposed approach-1 in Fig. 5.3.a) achieved good performance on en face images (Cohen's kappa: 0.6; weighted accuracy: 66.3 %). The deep learning models built with hierarchical classification with and without hierarchy transfer learning (proposed approach-2 in Fig. 5.3.b and proposed approach-3 in Fig. 5.3.c) achieved good performance using projection (Cohen's kappa: 0.678; weighted accuracy: 75.5 %), en face (Cohen's kappa: 0.707; weighted accuracy: 74.6 %), and disc H B-scan (Cohen's kappa: 0.708; weighted accuracy: 77.5 %) images. The deep learning models using hierarchical classification models provided significantly improved classification performance (P-values of the paired t-test: 0.027, 0.045, and 0.016, respectively) compared to models using the flat classification strategy for the same input images. There was no considerable difference between hierarchical classification with and without hierarchy transfer learning for all types of input images (left part of Fig. 5.6).



**Fig. 5.6. Cohen's kappa of deep learning models built with the proposed approaches.**

The stacked method of four models that were built with transfer learning and pretrained on ImageNet provided satisfactory performance, with a Cohen's kappa of 0.727 and weighted accuracy of 71.8 %. The multiple-input CNN models achieved a Cohen's kappa of 0.642 and weighted accuracy of 69.2 % with four input images (right part of Fig. 5.6). The standard deviation of Cohen's kappa for stacking was smaller than that of the multiple-input CNN trained with the same proposed approaches. The multiple-input CNN was not significantly higher than the single-input CNNs of projection images, en face images, and disc H B-scans.

### 5.3 Experiments Using Clinical Image Dataset

---

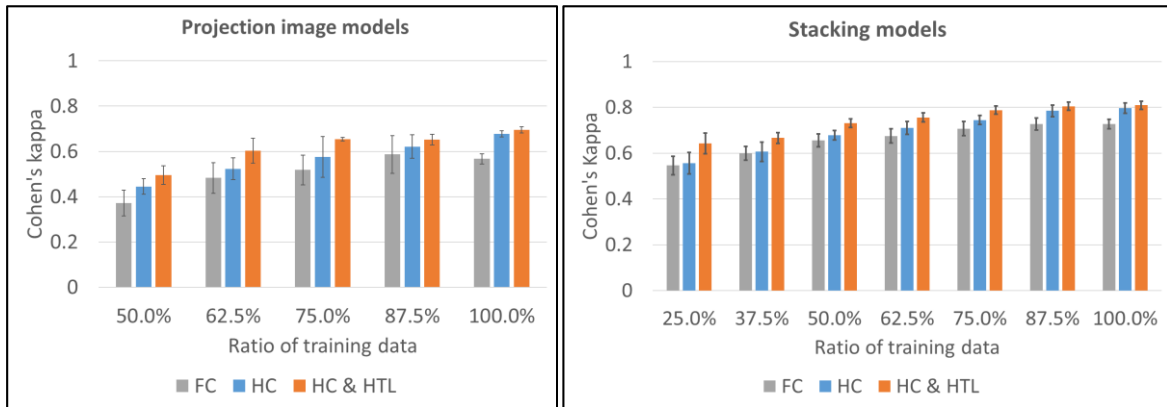
However, classification performance significantly improved with the stacking method for each single-input model (P-values of the paired t-test: 0.014, 0.022, 0.006, and 0.012 for projection, en face, disc H B-scan, and disc V B-scan images, respectively).

The hierarchical classification strategy with or without hierarchy transfer learning between low-level and high-level models showed significantly improved classification performance compared to flat classification (P-values of the paired t-test for the hierarchical classification strategy without hierarchy transfer learning were 0.001, 0.049, 0.008, and 0.013, and with hierarchy transfer learning were 0.047, 0.048, 0.030, and 0.031, for projection, en face, disc H B-scan, and disc V B-scan images, respectively, for both cases). There was no significant difference between the results of the hierarchical classification strategy with and without hierarchy transfer learning after applying the stacking ensemble method. Proposed approach-3 (hierarchical classification and hierarchy transfer learning with the stacking method) achieved the highest Cohen's kappa, 0.809, and weighted accuracy, 83.9 %.

**Table 5.2. Performance change of stacking models with different training dataset sizes**

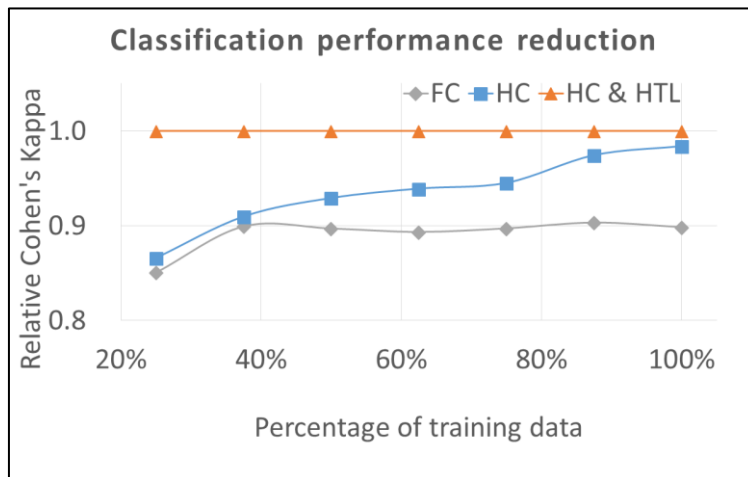
| Percent of total training data | Number of training data | Cohen's Kappa |       |          | Relative Cohen's Kappa |       |          |
|--------------------------------|-------------------------|---------------|-------|----------|------------------------|-------|----------|
|                                |                         | FC            | HC    | HC & HTL | FC                     | HC    | HC & HTL |
| <b>25.0 %</b>                  | <b>185</b>              | 0.546         | 0.556 | 0.642    | 0.851                  | 0.866 | 1.000    |
| <b>37.5 %</b>                  | <b>278</b>              | 0.600         | 0.606 | 0.666    | 0.899                  | 0.910 | 1.000    |
| <b>50.0 %</b>                  | <b>370</b>              | 0.655         | 0.679 | 0.731    | 0.897                  | 0.929 | 1.000    |
| <b>62.5 %</b>                  | <b>462</b>              | 0.675         | 0.710 | 0.756    | 0.893                  | 0.939 | 1.000    |
| <b>75.0 %</b>                  | <b>554</b>              | 0.707         | 0.744 | 0.788    | 0.897                  | 0.945 | 1.000    |
| <b>87.5 %</b>                  | <b>647</b>              | 0.727         | 0.784 | 0.805    | 0.903                  | 0.975 | 1.000    |
| <b>100.0 %</b>                 | <b>739</b>              | 0.727         | 0.796 | 0.809    | 0.899                  | 0.984 | 1.000    |

### 5.3 Experiments Using Clinical Image Dataset



(a)

(b)



(c)

**Fig. 5.7. Performance change with different training dataset sizes.** (a) Cohen's kappa with standard deviation error bars for deep learning models trained with different training methods based on single-input (projection) images. (b) Cohen's kappa with standard deviation error bars for deep learning models trained with different training methods based on all input images. (c) Calculated classification performance reduction for models using different training datasets and other training methods with stacking, with respect to Cohen's kappa for the CNN model built using HC & HTL with stacking.

The performance change (as measured using Cohen's kappa) provided by the deep learning classification models using projection input images were compared. All the models trained with flat classification (FC; Fig. 5.3.a), hierarchical classification without hierarchy transfer learning (HC; Fig. 5.3.b), and hierarchical classification with hierarchy transfer learning (HC & HTL; Fig. 5.3.c) using a larger training dataset achieved an increase in Cohen's kappa. Only HC and HC & HTL achieved satisfactory performance when using the entire training dataset. There was no significant difference between HC and FC in the case of a small training dataset. The HC & HTL strategy achieved a good Cohen's kappa when using 62.5 % of the entire training dataset (Fig. 5.7.a).

The performance change provided by the stacked deep learning classification models using all the types of input images were compared. All the stacked models trained with FC, HC, and HC & HTL using a larger training dataset achieved a higher Cohen's kappa (Table 5.2). Compared to the single-input (projection) model using the same number of training datasets, all the models achieved a higher Cohen's kappa. With sufficient training data, all the models achieved convergence classification performance. With the same number of training datasets, the classification performance of HC & HTL was better than that of FC (Table 5.2, Fig. 5.7.b; all  $P$ -values of paired t-tests  $< 0.05$ ). The performance of the models built with the HC strategy was better than that of the models built with the FC strategy for large training datasets (more than 75.0 % of the training dataset). However, there was no significant difference between HC and FC for a small training dataset (less than 75.0 % of training dataset). The HC & HTL strategy achieved a fairly good Cohen's kappa of 0.642 using only 25.0 % of the entire training dataset (Table 5.2, Fig. 5.7.b). Each model's performance reduction (relative Cohen's kappa) for each model using each training dataset was calculated with reference to the Cohen's kappa for the CNN model built with the HC & HTL strategy (Fig. 5.7.c). The performance reduction of flat classification was greater than  $(1.000-0.903) \times 100.0 \% = 9.7 \%$ , compared to using the proposed method and the same training dataset (Table 5.2, Fig. 5.7.c). The performance reduction of hierarchical classification was smaller than flat classification for each different size of training dataset, while the rate of performance reduction for hierarchical classification was larger in comparison with flat classification (Table 5.2, Fig. 5.7.c).

In this study, high-accuracy deep learning models were built for disease detection and classification. Experiments showed that the deep learning models trained with transfer learning from an ImageNet-pretrained CNN model with flat classification and data augmentation performed effectively in disease detection and classification on one type of input

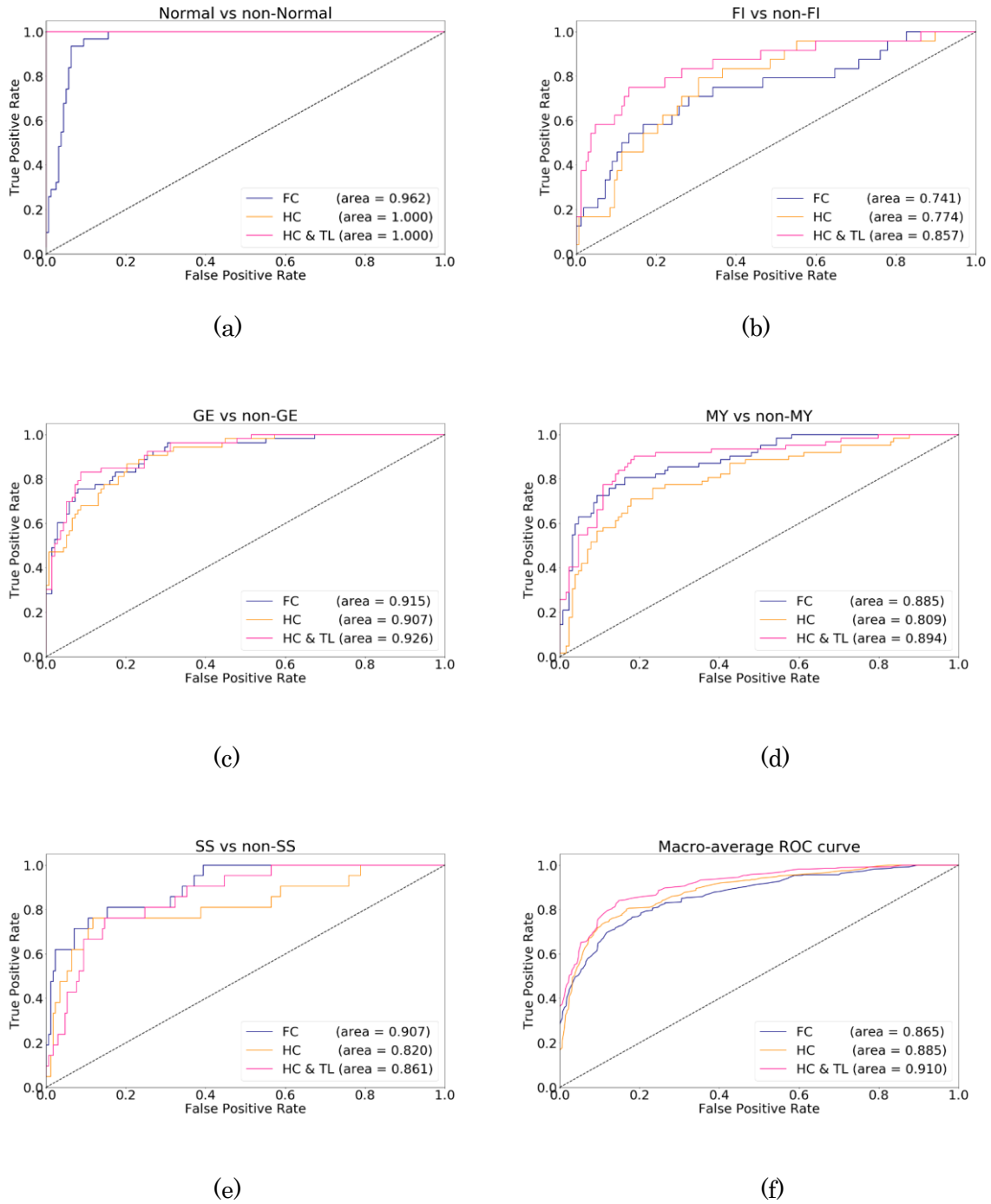


images (Fig. 5.3). The proposed approach by Kermany, et al. [34] is similar to proposed approach-1 (Fig. 5.3) as data augmentation for model training was adopted, which is useful in small datasets.

An approach for hierarchical classification was proposed. The deep learning models built with hierarchical classification performed well and producing substantially improved results compared to flat classification for three types of input images (Fig. 5.6). Similar to the proposed approach-2 (Fig. 5.3), a previous study used hierarchy knowledge in ophthalmologic disease classification and transfer learning from a pretrained model with a large natural image dataset and achieved a high accuracy [74]. The applicability of this approach in smaller datasets was evaluated and compared its performance with flat classification. Moreover, the same deep learning architecture for disease detection and classification was applied, which is a prerequisite for further use of hierarchy transfer learning. The combination of two methods on the features analyzed of doctors' diagnostic processes (proposed approach-3) produced higher classification performance using smaller datasets, in comparison with the flat and hierarchical classification.

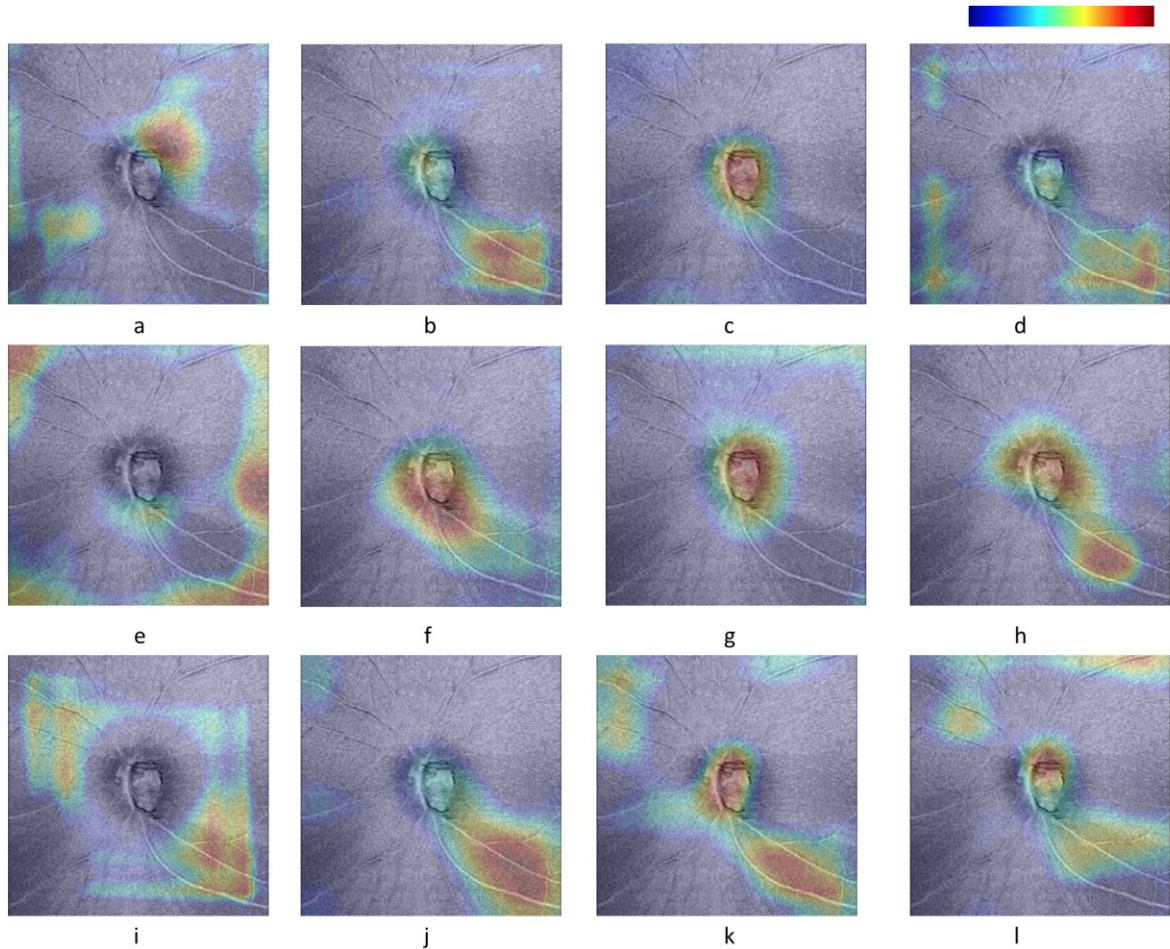
Two methods for handling multiple-input images were proposed, as doctors' requirement of having multiple images to make accurate diagnosis. As described in glaucoma guidelines, multiple sources of information, such as patient medical history, visual acuity, assessment of nerve fiber layer, and visual fields are required in order to accurately assess early-stage glaucoma [75]. Multiple useful images were extracted from one volumetric data to develop the deep learning models. Classification performance improved and was good for the multiple-input CNN and stacking methods. In comparison with the multiple-input CNN method, the stacking method considerably boosted the performance of four single-input CNNs with a small standard deviation error. Doctors' processes of making a diagnosis after asynchronously interpreting multiple images was implemented with the stacking method.

### 5.3 Experiments Using Clinical Image Dataset



**Fig. 5.8. ROC curves for stacked hierarchical classification models built using a small training dataset.** ROC curves for (a) normal vs others, (b) FI vs others, (c) GE vs others, (d) MY vs others, (e) SS vs others, (f) macro average ROC curve of all five classes.

The proposed approach of hierarchical classification and hierarchy transfer learning (approach-3; HC & HTL) showed good compatibility with the stacking method, and achieved higher classification performance than other proposed approaches. The classification performance was good even for smaller training datasets. The stacked deep learning models trained with three approaches (Fig. 5.8) achieved good classification performance using a small dataset (37.5 % of the training dataset). The area under the receiver operating characteristic (ROC) curve (AUC) for each class, with the confidence calculated by three different proposed methods, was used to confirm classification performance without a fixed threshold to predict. It is found that Normal, FI, GE, the proposed approach achieved the best AUC, followed by hierarchical classification without transfer learning (Fig. 5.8.a, 5.8.b, 5.8.c). To judge MY, and SS cases, hierarchical classification without transfer learning achieved the lowest AUC, and with the transfer learning effect, hierarchical training strategy let AUC got the highest AUC for distinguishing MY, but lower AUC than flat classification (Fig. 5.8.d, 5.8.e). The number of MY cases is the largest, and the SS is the smallest, it is considerable the imbalance problem seems to affect the performance even applying the weighted loss during training, and transfer learning from the low level model in the hierarchical ones has the effect of not being affected. Owing to the stacking method, eventually the macro AUC for hierarchical classification with transfer learning was the best (Fig. 5.8.f). In future studies, methods for overcoming the imbalance problem of the smallest class should be investigated.



**Fig. 5.9. Heatmaps for en face image of FI-type glaucoma (disease subtype) case.** (a) heatmap created with flat classification model trained with small dataset, (b) heatmap created with low level model trained with small dataset in hierarchical classification model of normal and glaucoma, (c) low level model trained with small dataset in hierarchical classification model without transfer learning, (d) low level model trained with small dataset in hierarchical classification model with transfer learning for glaucoma classification, (e) heatmap created with flat classification model trained with large dataset, (f) heatmap created with low level model trained with large dataset in hierarchical classification model of normal and glaucoma, (g) low level model trained with large dataset in hierarchical classification model without transfer learning, (h) low level model trained with large dataset in hierarchical classification model with transfer learning for glaucoma classification. The red area is the most important one for classification.

Furthermore, a commonly used method grad-CAM has been applied to observe the important areas for judgement by the built CNN models training with different strategies and training datasets [76]. In detail, grad-CAM makes the area characteristic as a heatmaps for each single input CNN model. A random sampled FI case, which was correctly identified with the proposed approach, but falsely judged with the other two classification models, was taken to show the heatmaps created with different models built with small (ratio: 37.5%), and medium (62.5%) training dataset, and correctly identified by all the models using the whole training dataset. One of the clinical characteristics of FI in en face image is retinal nerve fiber layer degeneration (NFLD), the black right bottom corner part from the center for this case. It was pitiful for the flat classification model could not find this feature in the en face image (Fig. 5.9.a, 5.9.e), but it succeeded to find the attractive region for low-level classification model of normal and glaucoma (Fig. 5.9.b, 5.9.f). Owing to transfer learning from low-level model, the high-level model for glaucoma classification continues to pay high attention on this area to achieve a better performance, while the high-level model without transfer learning just notes the center region (optic disc) (Fig.5.9.c, 5.9.d, 5.9.g, 5.9.h). The heatmaps created by the models built with medium training dataset seems changed gradually from the ones trained with the small dataset. It was very similar to find the NFLD area important in the models correctly identified the FI case.

A deep learning model for glaucoma detection and classification was built by hierarchical classification, hierarchy transfer learning and stacking ensemble method using different images, achieving a high Cohen's kappa of 0.809. Based on the previous guideline of Cohen's kappa [46], this value means the substantial performance of decision support system for glaucoma clinical care. For comparison, the other diagnosis test has also been conducted. Images of 50 cases were randomly sampled from the dataset used in this study, which were then classified into normal and glaucoma sub-categories by three medical ophthalmology interns. The average Cohen's kappa of these classification results with supervised data was 0.408, lower than that of the deep learning model. The model can be used in diagnosis support of glaucoma by providing the confidence level for normal cases and each disc type, which would help doctors to select a proper treatment plans according to the optic disc shape [37], [77].

## 5.4. Conclusions

This chapter presents an extension method proposed two-step framework introduced in Chapter 4 to handle multiple images. The machine learning models are trained separately

based on each type of input images with the two-step framework, first step of which builds the model for disease detection, and the second step of which builds the model for disease classification by reusing parameters of the model for disease detection. Then the models were combined in the stacked manner to handle multiple images. In detail, the models trained separately based on each type of input images with the framework using hierarchical classification and hierarchy transfer learning method were combined with a stacking ensemble method.

In experimentations, I evaluated this extended framework by building an ophthalmologic disease (glaucoma) detection and classification machine learning model via deep learning based on a retinal image dataset consisting of several types of medical images extracted from volumetric data. For the models using single input images, the classification performance of the models built with the proposed framework shown in Chapter 4 were compared to the flat models, while for the models using multiple input images, the classification performance of the models built with the extension approach of the proposed framework in Chapter 4, trained with random sampled smaller size datasets from the dataset above. As a result, the classification performance of the models based on single input image built with the proposed two-step framework shown in Chapter 4 was higher than the flat models. Furthermore, the stacked classification models built with the extended proposed framework achieved good performance, higher than the models using single input image. With a smaller size dataset, models with higher classification performance were built by proposed framework compared to flat models in both cases of handling single input and multiple input data. A high accuracy was achieved via by a combination of the proposed methods, and the effectiveness of the proposed approach in its applicability using smaller size dataset was demonstrated.

In this work, two major contributions were made: (1) The effectiveness of building machine learning models for disease detection and classification hierarchically can improve the classification performances when using limited size dataset was demonstrated: first classifying normal and diseased cases, then classifying diseased cases into sub-categories was demonstrated. (2) The effectiveness of building high level models in hierarchical classification models for disease classification by reusing low-level models in the hierarchical classification models for disease detection was demonstrated. (3) The applicability of the two-step framework in building machine learning models handling multiple input images was demonstrated.

Conclusively, the proposed two-step hierarchical framework has a high potential of

## 5.4 Conclusions

---

deploying high accurate machine learning models for ophthalmologic disease detection and classification with limited labelled data, in both cases of handling single input images and multiple input images.

## Chapter 6.

### Conclusions and Future Work

In this chapter, I summarize the contributions of this dissertation and discuss the necessary future work.

#### 6.1. Summary of This Dissertation

The major theme of this dissertation is focused on building high accurate machine learning models for ophthalmologic disease detection and classification with small-size dataset, which aims to assist the clinical decision making of selecting proper treatment for ophthalmologic disease.

To achieve this goal, first I analyze the doctors' diagnostic processes as follows:

- 1) Doctors perform disease classification after classifying healthy and diseased cases (disease detection), because the disease classification is difficult even for specialists, which is based on complex symptoms, and it should be performed in the status of normal cases excluded thoroughly.
- 2) Doctors classify diseases into subcategories by reusing the knowledge of classifying normal and diseased cases.
- 3) Doctors use multiple information (e.g., multiple types of medical images) to make optimal diagnosis.

According to the characteristic of doctors' diagnostic processes above, a two-step framework is proposed to build machine learning models for ophthalmologic disease detection and classification. In the first step, the machine learning model for classifying healthy and disease (disease detection) is built, and in the second step the model for disease classification is built by reusing parameters of the model for disease detection. Moreover, this two-step



framework is extended to handle multiple input data. To implement this framework in building machine learning models, three machine learning methods are created with separate roles as below.

- 1) Hierarchical classification method, to build machine learning models for disease classification, after building the models for classifying healthy and disease (disease detection).

- 2) Hierarchy transfer learning method, to build machine learning models for disease classification by reusing parameters of the model for disease detection

- 3) Stacking ensemble method, to build machine learning models handling multiple input data by combining the machine learning models trained separately on single input data with the training method of hierarchical classification and hierarchy transfer learning method.

A series of experiments are performed to evaluate the proposed framework for training machine learning models of ophthalmologic disease detection and classification.

In Chapter 3, I present the first approach of building machine learning models in two steps, that tries to classify diseases after disease detection. It is implemented by the hierarchical classification of two steps. In each step, feature selection that is one field of feature engineering is used. In experimentation, this method is evaluated by building an ophthalmologic disease (glaucoma) detection and classification machine learning models with a clinical data extracted from a hospital database. The dataset consists of extracted quantified parameters from medical images and demographic data, together with the labelled data by experienced ophthalmologists. The classification performance of the models built with the proposed approach are compared to the flat models built in one-step to detect and classify diseases, trained with different size datasets random sampled from the dataset above. The result of the experiments demonstrates that the classification performance of the machine learning models trained with hierarchical classification method combining with feature selection can be elevated on a disease detection and classification task compared with the flat models, and the applicability of the proposed framework to build high accurate machine learning models in limited size datasets is also demonstrated.

Chapter 4 attempts to apply deep learning technique for handling medical images directly without feature engineering to build machine learning models for ophthalmologic disease detection and classification. A two-step framework to build deep learning models for ophthalmologic disease detection and classification, first step of which builds the model for

disease detection, and the second step of which builds the model for disease classification by reusing parameters of the model for disease detection, is introduced. In detail, besides the hierarchical classification method, hierarchy transfer learning method is created and used to build deep learning models. In experimentation, firstly, the concept of the proposed method of building deep learning models is evaluated with a natural image dataset and labelled data that is easy to evaluate, in which the labels of the images have hierarchy relationship. The classification performance of the models built with the proposed approach are compared to the flat models built in one-step to detect and classify diseases, trained with different size datasets random sampled from the dataset above. As a result, compared with flat models, the effectiveness of improved classification performance for the model built with the method of combining hierarchical classification and hierarchy transfer learning method is demonstrated even for the smaller size datasets. This approach also succeeds in building a high accurate deep learning model with labelled clinical medical image dataset for an ophthalmologic disease (age-related macular degeneration) detection and classification. Similar with the result of experiments using the natural image dataset above, the proposed approach enables machine learning models to achieve high accuracies and improved classification performance in comparison with flat models, the effect and efficiency have been demonstrated even in smaller size datasets.

Chapter 5 presents an extension approach for the proposed two-step framework shown in Chapter 4 to handle multiple images. The machine learning models are trained separately based on each type of input images with the two-step framework, first step of which builds the model for disease detection, and the second step of which builds the model for disease classification by reusing parameters of the model for disease detection. Then the models are combined by the stacked manner to handle multiple images. In detail, the models trained separately based on each type of input images with the framework using hierarchical classification and hierarchy transfer learning method are combined with a stacking ensemble method. In experimentation, this extended framework is evaluated by building an ophthalmologic disease (glaucoma) detection and classification machine learning model via deep learning based on a retinal image dataset consisting of several types of medical images extracted from volumetric data. For the models using single input images, the classification performance of the models built with the proposed framework shown in Chapter 4 are compared to the flat models, while for the models using multiple input images, the classification performance of the models built with the extension approach of the proposed framework in Chapter 4, trained with smaller size datasets random sampled from the dataset above are compared with flat models. As a result, the classification performance of the models

based on single input image built with the proposed two-step framework shown in Chapter 4 was higher than the flat models. The stacked classification models built with the proposed framework achieve good performance, higher than the models using single type of input image. With a smaller size dataset, models with higher classification performance are built by proposed framework compared to flat models in both cases of handling single input and multiple input data. A high accuracy is achieved via by a combination of the proposed methods, and the effectiveness of the proposed approach in its applicability in using smaller size dataset is demonstrated.

In conclusion, the proposed two-step hierarchical framework has a high potential of deploying high accurate machine learning models for ophthalmologic disease detection and classification with limited labelled data, to assist the clinical decision making for selecting the proper treatments for the patients.

## 6.2. Discussion of the Future Work

In this section, the limitations of the approaches are described based on the summary in the previous sections, and the necessary future work is discussed.

The effect of my proposed framework to build high accurate machine learning models for ophthalmologic disease detection and classification is shown based on multiple images from just one single modality, thus using multiple images from different multi-modalities are promising to be utilized in future work. Moreover, a way of handling multiple images and metadata from patients together should be researched in the extension of proposed method.

The input image data is only two-dimensional (2D), thus the efficacy of training machine learning model by my proposed approach will be validated in handling three-dimensional (3D) image data, which is more difficult to build high accurate classification models. As shown in Chapter 5, handling multiple images at the same time by using a multiple input CNN does not achieve a higher accuracy than the stacking ensemble method using the proposed two-step training framework, thus the two-step framework might be further extended in future work in asynchronous processing manner for building a machine learning model handling 3D image data. For instance, with extracted information from adjacent frames should be firstly tried as the input to build the machine learning models, while recently long short-term memory (LSTM) [78] is shown powerful to process 3D medical image data combined with CNN,

when handling the images as a time-sequence data.

There are many new machine learning techniques being reported recently, such as deep learning architectures and data augmentation techniques, which can boost the performance of deep learning models. Hopefully, machine learning models built with these new techniques or new architectures and my approaches would achieve more excellent performance for ophthalmologic disease detection and classification, and this will be demonstrated in future work.

In this dissertation, simple data augmentation is used to build high accurate machine learning models. For many diseases, the distribution of disease sub-classes in collected datasets is heavily skewed by each class's prevalence among patients, and so detecting rare diseases in medical images with deep learning can be challenging. Furthermore, for some kinds of medical images, the image quality is not stable. With these problems, instead of simple data augmentation, GAN [79] that is capable of learning the distribution of the image data, will be considered to be applied for increasing the training samples for earning better classification performance in the future work.

As the disease classification into subcategories is difficult even for experienced doctors, the machine learning models in this dissertation were built with only the consistent labels by the several ophthalmologists to demonstrate the effectiveness and efficacy of the proposed framework for training machine learning models. Thus, in the future work, the images whose grading are difficult might be used in a semi-supervised classification manner to improve performances of the machine learning models built with the proposed framework.

In this dissertation, the medical data relevant with the ophthalmologic diseases, patient number of which are huge are used to demonstrate my approach, and in future work the effect of the proposed approach in real clinical medical data relevant with rare ophthalmologic diseases should be researched. Finally, since the proposed framework for building machine learning models is not limited to ophthalmologic data, the effectiveness of building high accurate machine learning models based on medical data from other medical departments will be demonstrated in future work.

## Acknowledgements

I am very grateful to many people for their help and support during my research in the whole doctoral course. Foremost, I would like to express my sincere gratitude to my supervisor Prof. Hideo Yokota for the continuous support of my doctoral study, for his patience, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing this dissertation, and he gave me critical advices for making my research better.

Moreover, I would like to thank the rest of my dissertation committee: Prof. Hisashi Tamaki, Prof. Hideyuki Usui and Prof. Takenao Ohkawa, for their encouragement, insightful comments, and valuable questions to improve my dissertation. Thanks go to Prof. Kuniaki Uehara in Kobe University and my boss in my compony Dr. Masahiro Akiba giving me continuous encouragement, and useful comments throughout my study. My sincere thanks also go to all staff including Dr. Shin Yoshizawa, Dr. Satoko Takemoto, and Dr. Takashi Michikawa in Image Processing Research Team of RIKEN Center for Advanced Photonics for their assistance, cooperation and of course friendship. I would also thank the doctors in the collaboration research teams of Tohoku University Hospital or Kobe Eye Center.

Last but not the least, I would like to thank my family members: my parents Jingzhe An and Jifu Fang, for giving birth to me at the first place, and my loving wife Dr. Liying Pei, my son Junxi An and my daughter Yuxi An giving continuous support spiritually throughout my life.

## Bibliography

- [1] P. D. United Nations, Department of Economic and Social Affairs, *World population prospects 2019: Highlights*. 2019.
- [2] R. R. A. Bourne *et al.*, “Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis,” *Lancet Glob. Heal.*, vol. 5, no. 9, pp. e888–e897, Sep. 2017, doi: 10.1016/S2214-109X(17)30293-0.
- [3] A. Gordois *et al.*, “An estimation of the worldwide economic and health burden of visual impairment,” *Glob. Public Health*, vol. 7, no. 5, pp. 465–481, May 2012, doi: 10.1080/17441692.2011.634815.
- [4] M. J. Burton *et al.*, “Announcing The Lancet Global Health Commission on Global Eye Health,” *The Lancet Global Health*, vol. 7, no. 12. Elsevier Ltd, pp. e1612–e1613, Dec. 01, 2019, doi: 10.1016/S2214-109X(19)30450-4.
- [5] S. Resnikoff *et al.*, “Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): Will we meet the needs?,” *Br. J. Ophthalmol.*, vol. 104, no. 4, pp. 588–592, Apr. 2020, doi: 10.1136/bjophthalmol-2019-314336.
- [6] W. Lu, Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen, “Applications of Artificial Intelligence in Ophthalmology: General Overview,” *J. Ophthalmol.*, vol. 2018, no. Cml, 2018, doi: 10.1155/2018/5278196.
- [7] J. Ker, L. Wang, J. Rao, and T. Lim, “Deep Learning Applications in Medical Image Analysis,” *IEEE Access*, vol. 6, pp. 9375–9379, Dec. 2017, doi: 10.1109/ACCESS.2017.2788044.
- [8] F. Altaf, S. M. S. Islam, N. Akhtar, and N. K. Janjua, “Going Deep in Medical Image Analysis: Concepts, Methods, Challenges and Future Directions,” *IEEE Access*, vol. 7, pp. 99540–99572, Feb. 2019, Accessed: Jul. 05, 2020. [Online]. Available: <http://arxiv.org/abs/1902.05655>.

- [9] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553. Nature Publishing Group, pp. 436–444, May 27, 2015, doi: 10.1038/nature14539.
- [10] S. H. Bach, B. He, A. Ratner, and C. Ré, “Learning the Structure of Generative Models without Labeled Data,” *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 1, pp. 434–449, Mar. 2017, Accessed: Jul. 06, 2020. [Online]. Available: <http://arxiv.org/abs/1703.00854>.
- [11] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, “Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Med. Image Anal.*, vol. 54, pp. 280–296, Apr. 2018, Accessed: Jul. 06, 2020. [Online]. Available: <http://arxiv.org/abs/1804.06353>.
- [12] M. Minsky, “Steps Toward Artificial Intelligence,” *Proc. IRE*, vol. 49, no. 1, pp. 8–30, 1961, doi: 10.1109/JRPROC.1961.287775.
- [13] J. Weng *et al.*, “Autonomous mental development by robots and animals,” *Science*, vol. 291, no. 5504, pp. 599–600, Jan. 26, 2001, doi: 10.1126/science.291.5504.599.
- [14] G. Huang, G. Bin Huang, S. Song, and K. You, “Trends in extreme learning machines: A review,” *Neural Networks*, vol. 61. Elsevier Ltd, pp. 32–48, Jan. 01, 2015, doi: 10.1016/j.neunet.2014.10.001.
- [15] A. Gudivada and N. Tabrizi, “A Literature Review on Machine Learning Based Medical Information Retrieval Systems,” in *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, Jan. 2019, pp. 250–257, doi: 10.1109/SSCI.2018.8628846.
- [16] B. E. Bejnordi *et al.*, “Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images,” *J. Med. Imaging*, vol. 4, no. 4, p. 1, Dec. 2017, doi: 10.1117/1.jmi.4.4.044504.
- [17] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.
- [18] B. van Ginneken, “Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning,” *Radiological Physics and Technology*, vol. 10, no. 1. Springer Tokyo, pp. 23–32, Mar. 01, 2017, doi: 10.1007/s12194-017-0394-5.
- [19] M. Caixinha and S. Nunes, “Machine Learning Techniques in Clinical Vision Sciences,”

- Current Eye Research*, vol. 42, no. 1. Taylor and Francis Ltd, pp. 1–15, Jan. 02, 2017, doi: 10.1080/02713683.2016.1175019.
- [20] D. S. W. Ting *et al.*, “Artificial intelligence and deep learning in ophthalmology,” *British Journal of Ophthalmology*, vol. 103, no. 2. BMJ Publishing Group, pp. 167–175, Feb. 01, 2019, doi: 10.1136/bjophthalmol-2018-313173.
- [21] J. Akkara and A. Kuriakose, “Role of artificial intelligence and machine learning in ophthalmology,” *Kerala J. Ophthalmol.*, vol. 31, no. 2, p. 150, 2019, doi: 10.4103/kjo.kjo\_54\_19.
- [22] Y. Yamamoto *et al.*, “Automated acquisition of explainable knowledge from unannotated histopathology images,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–9, Dec. 2019, doi: 10.1038/s41467-019-13647-8.
- [23] L. Rokach and O. Maimon, “Top-down induction of decision trees classifiers - A survey,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 35, no. 4, pp. 476–487, Nov. 2005, doi: 10.1109/TSMCC.2004.843247.
- [24] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
- [25] C. C. Chang and C. J. Lin, “LIBSVM: A Library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–39, 2011, doi: 10.1145/1961189.1961199.
- [26] J. M. Keller and M. R. Gray, “A Fuzzy K-Nearest Neighbor Algorithm,” *IEEE Trans. Syst. Man Cybern.*, vol. SMC-15, no. 4, pp. 580–585, 1985, doi: 10.1109/TSMC.1985.6313426.
- [27] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural networks design*. 1997.
- [28] D. E. Freund, N. Bressler, and P. Burlina, “Automated detection of drusen in the macula,” in *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, 2009, pp. 61–64, doi: 10.1109/ISBI.2009.5192983.
- [29] C. C. Aggarwal, X. Kong, Q. Gu, J. Han, and P. S. Yu, “Active learning: A survey,” *Data Classif. Algorithms Appl.*, pp. 571–605, 2014, doi: 10.1201/b17320.
- [30] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.



- [31] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA - J. Am. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016, doi: 10.1001/jama.2016.17216.
- [32] I. J. Goodfellow *et al.*, “Generative Adversarial Nets,” 2014. Accessed: Jul. 06, 2020. [Online]. Available: <http://www.github.com/goodfeli/adversarial>.
- [33] L. Torrey and S. Jude, “Transfer learning,” *Handb. Res. Mach. Learn. Appl. IGI Glob.*, vol. 3, pp. 17–35, 2009, doi: 10.1201/b17320.
- [34] D. S. Kermany *et al.*, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, p. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.
- [35] A. G *et al.*, “Glaucoma Diagnosis With Machine Learning Based on Optical Coherence Tomography and Color Fundus Images,” *J. Healthc. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/4061313.
- [36] C. N. Silla and A. A. Freitas, “A survey of hierarchical classification across different application domains,” *Data Mining and Knowledge Discovery*, vol. 22, no. 1–2. Springer, pp. 31–72, Jan. 07, 2011, doi: 10.1007/s10618-010-0175-9.
- [37] T. Nakazawa, “Ocular blood flow and influencing factors for glaucoma,” *Asia-Pacific J. Ophthalmol.*, vol. 5, no. 1, pp. 38–44, 2016, doi: 10.1097/APO.0000000000000183.
- [38] M. T. Nicolela and S. M. Drance, “Various glaucomatous optic nerve appearances: Clinical correlations,” *Ophthalmology*, vol. 103, no. 4, pp. 640–649, 1996, doi: 10.1016/S0161-6420(96)30640-4.
- [39] K. Omodaka, N. Takada, T. Yamaguchi, H. Takahashi, M. Araie, and T. Nakazawa, “Characteristic correlations of the structure-function relationship in different glaucomatous disc types,” *Jpn. J. Ophthalmol.*, vol. 59, no. 4, pp. 223–229, Jul. 2015, doi: 10.1007/s10384-015-0379-z.
- [40] T. Nakazawa *et al.*, “Progression of visual field defects in eyes with different optic disc appearances in patients with normal tension glaucoma,” *J. Glaucoma*, vol. 21, no. 6, pp. 426–430, Aug. 2012, doi: 10.1097/IJG.0b013e3182182897.
- [41] T. Nakazawa *et al.*, “Association between optic nerve blood flow and objective

- examinations in glaucoma patients with generalized enlargement disc type,” *Clin. Ophthalmol.*, vol. 5, p. 1549, Oct. 2011, doi: 10.2147/oph.s22097.
- [42] S. J. Kim, K. J. Cho, and S. Oh, “Development of machine learning models for diagnosis of glaucoma,” *PLoS One*, vol. 12, no. 5, May 2017, doi: 10.1371/journal.pone.0177726.
- [43] K. Chan, T. W. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, and T. J. Sejnowski, “Comparison of machine learning and traditional classifiers in glaucoma diagnosis,” *IEEE Trans. Biomed. Eng.*, vol. 49, no. 9, pp. 963–974, 2002, doi: 10.1109/TBME.2002.802012.
- [44] D. Bizios, A. Heijl, J. L. Hougaard, and B. Bengtsson, “Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fibre layer thickness parameters measured by Stratus OCT,” *Acta Ophthalmol.*, vol. 88, no. 1, pp. 44–52, Feb. 2010, doi: 10.1111/j.1755-3768.2009.01784.x.
- [45] M. D. Abràmoff *et al.*, “Automated analysis of retinal images for detection of referable diabetic retinopathy,” *JAMA Ophthalmol.*, vol. 131, no. 3, pp. 351–357, Mar. 2013, doi: 10.1001/jamaophthalmol.2013.1743.
- [46] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
- [47] Y. Peng *et al.*, “DeepSeeNet: A Deep Learning Model for Automated Classification of Patient-based Age-related Macular Degeneration Severity from Color Fundus Photographs,” *Ophthalmology*, vol. 126, no. 4, pp. 565–575, Apr. 2019, doi: 10.1016/j.ophtha.2018.11.015.
- [48] K. Omodaka *et al.*, “Pilot study for three-dimensional assessment of laminar pore structure in patients with glaucoma, as measured with swept source optical coherence tomography,” *PLoS One*, vol. 13, no. 11, pp. 1–12, 2018, doi: 10.1371/journal.pone.0207600.
- [49] C. K. S. Leung *et al.*, “Analysis of retinal nerve fiber layer and optic nerve head in glaucoma with different reference plane offsets, using optical coherence tomography,” *Investig. Ophthalmol. Vis. Sci.*, vol. 46, no. 3, pp. 891–899, Mar. 2005, doi: 10.1167/iovs.04-1107.
- [50] N. Takada *et al.*, “OCT-based quantification and classification of optic disc structure in glaucoma patients,” *PLoS One*, vol. 11, no. 8, p. 160226, Aug. 2016, doi:

- 10.1371/journal.pone.0160226.
- [51] K. Omodaka *et al.*, “Classification of optic disc shape in glaucoma using machine learning based on quantified ocular parameters,” *PLoS One*, vol. 12, no. 12, p. e0190012, Dec. 2017, doi: 10.1371/journal.pone.0190012.
- [52] Q. Yang *et al.*, “Automated layer segmentation of macular OCT images using dual-scale gradient information,” *Opt. Express*, vol. 18, no. 20, p. 21293, Sep. 2010, doi: 10.1364/oe.18.021293.
- [53] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.
- [54] J. Yang and V. Honavar, “Feature Subset Selection Using a Genetic Algorithm,” in *Feature Extraction, Construction and Selection*, Springer US, 1998, pp. 117–136.
- [55] A. El Akadi, A. Amine, A. El Ouardighi, and D. Aboutajdine, “A two-stage gene selection scheme utilizing MRMR filter and GA wrapper,” *Knowl. Inf. Syst.*, vol. 26, no. 3, pp. 487–500, Mar. 2011, doi: 10.1007/s10115-010-0288-x.
- [56] L. Milne, “Feature Selection Using Neural Networks with Contribution Measures 1 Introduction 2 Calculating Proportion Contribution of Input Features to Outputs,” no. November, pp. 1–8, 1995.
- [57] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” *Proc. - Int. Conf. Pattern Recognit.*, pp. 3121–3124, 2010, doi: 10.1109/ICPR.2010.764.
- [58] Z. Yan *et al.*, “HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 2740–2748, 2015, doi: 10.1109/ICCV.2015.314.
- [59] Y. Seo and K. shik Shin, “Hierarchical convolutional neural networks for fashion image classification,” *Expert Syst. Appl.*, vol. 116, pp. 328–339, Feb. 2019, doi: 10.1016/j.eswa.2018.09.022.
- [60] R. Sali *et al.*, “Hierarchical Deep Convolutional Neural Networks for Multi-category Diagnosis of Gastrointestinal Disorders on Histopathological Images,” May 2020,

- Accessed: Jul. 06, 2020. [Online]. Available: <http://arxiv.org/abs/2005.03868>.
- [61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.
- [62] C. V. Regatieri, L. Branchini, and J. S. Duker, "The role of spectral-domain OCT in the diagnosis and management of neovascular age-related macular degeneration.," *Ophthalmic surgery, lasers & imaging: the official journal of the International Society for Imaging in the Eye*, vol. 42 Suppl, no. 0. NIH Public Access, p. S56, 2011, doi: 10.3928/15428877-20110627-05.
- [63] A. E. Fung *et al.*, "An Optical Coherence Tomography-Guided, Variable Dosing Regimen with Intravitreal Ranibizumab (Lucentis) for Neovascular Age-related Macular Degeneration," *Am. J. Ophthalmol.*, vol. 143, no. 4, 2007, doi: 10.1016/j.ajo.2007.01.028.
- [64] M. R. Alexandru and N. M. Alexandra, "Wet age related macular degeneration management and follow-up.," *Rom. J. Ophthalmol.*, vol. 60, no. 1, pp. 9–13, 2016, Accessed: Jul. 06, 2020. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/27220225>.
- [65] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," *Progress in Retinal and Eye Research*, vol. 67. Elsevier Ltd, pp. 1–29, Nov. 01, 2018, doi: 10.1016/j.preteyeres.2018.07.004.
- [66] BECKER and S., "Improving the convergence of backpropagation learning with second order methods," *Proc. 1988 Connect. Model. Summer Sch.*, pp. 29–37, 1988, Accessed: Jul. 06, 2020. [Online]. Available: <https://ci.nii.ac.jp/naid/10008946219>.
- [67] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," *JAMA Ophthalmol.*, vol. 135, no. 11, pp. 1170–1176, Nov. 2017, doi: 10.1001/jamaophthalmol.2017.3782.
- [68] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images," *Kidney Int. Reports*, vol. 1, no. 4, pp. 322–327, Jul. 2017, doi: 10.1016/j.oret.2016.12.009.

- [69] J. De Fauw *et al.*, “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, Sep. 2018, doi: 10.1038/s41591-018-0107-6.
- [70] M. Treder, J. L. Lauermann, and N. Eter, “Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning,” *Graefe’s Arch. Clin. Exp. Ophthalmol.*, vol. 256, no. 2, pp. 259–265, Feb. 2018, doi: 10.1007/s00417-017-3850-3.
- [71] H. Quigley and A. T. Broman, “The number of people with glaucoma worldwide in 2010 and 2020,” *British Journal of Ophthalmology*, vol. 90, no. 3. Br J Ophthalmol, pp. 262–267, Mar. 2006, doi: 10.1136/bjo.2005.081224.
- [72] H. Y. Shin, H. Y. L. Park, Y. Jung, J. A. Choi, and C. K. Park, “Glaucoma diagnostic accuracy of optical coherence tomography parameters in early glaucoma with different types of optic disc damage,” *Ophthalmology*, vol. 121, no. 10, pp. 1990–1997, Oct. 2014, doi: 10.1016/j.ophtha.2014.04.030.
- [73] F. K. Nakano, S. Martiello Mastelini, S. Barbon, and R. Cerri, “Stacking methods for hierarchical classification,” in *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, 2017, vol. 2017–December, pp. 289–296, doi: 10.1109/ICMLA.2017.0-145.
- [74] R. Zhang, J. Zhao, G. Chen, T. Wang, G. Zhang, and B. Lei, “Aggressive Posterior Retinopathy of Prematurity Automated Diagnosis via a Deep Convolutional Network,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Oct. 2019, vol. 11855 LNCS, pp. 165–172, doi: 10.1007/978-3-030-32956-3\_20.
- [75] M. Fingeret, “Care of the Patient with Open-angle glaucoma,” *Am. Optom. Assoc.*, vol. 1, pp. 1–161, 2011, [Online]. Available: <http://www.nejm.org/doi/pdf/10.1056/NEJM199304153281507>.
- [76] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.
- [77] T. Nakazawa, N. Fuse, K. Omodaka, N. Aizawa, S. Kuwahara, and K. Nishida, “Different

- types of optic disc shape in patients with advanced open-angle glaucoma,” *Jpn. J. Ophthalmol.*, vol. 54, no. 4, pp. 291–295, Jul. 2010, doi: 10.1007/s10384-010-0816-y.
- [78] I. Shahzadi, F. Meriadeau, T. B. Tang, and A. Quyyum, “CNN-LSTM: Cascaded framework for brain tumour classification,” *2018 IEEE EMBS Conf. Biomed. Eng. Sci. IECBES 2018 - Proc.*, no. December, pp. 633–637, 2019, doi: 10.1109/IECBES.2018.8626704.
- [79] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Med. Image Anal.*, vol. 58, 2019, doi: 10.1016/j.media.2019.101552.

## Publications

### Journal Papers

1. Kazuko Omodaka, [Guangzhou An](#), Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Hidetoshi Takahashi, Hideo Yokota, Masahiro Akiba, Toru Nakazawa, “Classification of optic disc shape in glaucoma using machine learning based on quantified ocular parameters,” PLoS ONE, 12(12), e0190012, 2017, doi: 10.1371/journal.pone.0190012.
2. [Guangzhou An](#), Kazuko Omodaka, Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Toru Nakazawa, Hideo Yokota, Masahiro Akiba, “Comparison of machine-learning classification models for glaucoma management,” Journal of healthcare engineering, Article ID 6874765, 2018, doi: 10.1155/2018/6874765.
3. Kazuko Omodaka, Shigeto Maekawa, [Guangzhou An](#), Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Hidetoshi Takahashi, Hideo Yokota, Masahiro Akiba, Toru Nakazawa, “Pilot study for three-dimensional assessment of laminar pore structure in patients with glaucoma, as measured with swept source optical coherence tomography,” PLoS ONE, 13(11), e0207600, 2018, doi: 10.1371/journal.pone.0207600.
4. [Guangzhou An](#), Kazuko Omodaka, Kazuki Hashimoto, Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Hideo Yokota, Masahiro Akiba, Toru Nakazawa, “Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images,” Journal of healthcare engineering, Article ID 4061313, 2019, doi: 10.1155/2019/4061313.
5. Naohiro Motozawa, [Guangzhou An](#), Seiji Takagi, Shohei Kitahata, Michiko Mandai, Yasuhiko Hiramami, Hideo Yokota, Masahiro Akiba, Akitaka Tsujikawa, Masayo Takahashi, Yasuo Kurimoto, “Optical Coherence Tomography-Based Deep-Learning Models for Classifying Normal and Age-Related Macular Degeneration and Exudative and Non-Exudative Age-Related Macular Degeneration Changes,” Ophthalmology and Therapy, 8, 527–539, 2019, doi: 10.1007/s40123-019-00207-y.
6. Kazuko Omodaka, Shunsuke Fujioka, [Guangzhou An](#), Takuma Udagawa, Satoru Tsuda,

Yukihiro Shiga, Soichiro Morishita, Tsutomu Kikawa, Kyongsun Pak, Masahiro Akiba, Hideo Yokota, Toru Nakazawa, "Structural characterization of glaucoma patients with low ocular blood flow," *Current Eye Research*, pp.1-7, 2020, doi: 10.1080/02713683.2020.1736306.

7. Guangzhou An, Masahiro Akiba, Kazuko Omodaka, Toru Nakazawa, Hideo Yokota, "Hierarchical deep learning models using transfer learning for disease detection and classification based on biomedical images," *Scientific Reports* (Under Review).

### Conference Papers

1. Masahiro Akiba, Guangzhou An, Hideo Yokota, Kazuko Omodaka, Satoru Tsuda, Tsutomu Kikawa, Hidetoshi Takahashi, Toru Nakazawa, "Most contributing quantified ocular parameters for classification of glaucomatous optic disc shape," *Investigative Ophthalmology & Visual Science*, 59(9), 1718-1718, 2018.
2. Maiko Abe, Kazuko Omodaka, Guangzhou An, Tsutomu Kikawa, Masahiro Akiba, Hideo Yokota, Toru Nakazawa, "Assisting glaucoma diagnosis with optical coherence tomography and color fundus images using machine learning approach," *Investigative Ophthalmology & Visual Science*, 59 (9), 2079-2079, 2018.
3. Masahiro Akiba, Guangzhou An, Kazuko Omodaka, Kazuki Hashimoto, Satoru Tsuda, Yukihiro Shiga, Toru Nakazawa, "Evaluation of glaucoma diagnosis machine learning models based on color optical coherence tomography and color fundus images" *Investigative Ophthalmology & Visual Science*, 60 (9), 1298-1298, 2019.
4. Guangzhou An, Hideo Yokota, Naohiro Motozawa, Seiji Takagi, Michiko Mandai, Shohei Kitahata, Yasukiko Hirami, Masayo Takahashi, Yasuo Kurimoto, Masahiro Akiba, "Deep Learning Classification Models Built with Two-step Transfer Learning for Age Related Macular Degeneration Diagnosis," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2049-2052, 2019, doi: 10.1109/EMBC.2019.8857468.
5. Takahashi Michikawa, Satoshi Wada, Hideo Yokota, Guangzhou An, Masahiro Akiba, Kazuko Omodaka, Toru Nakazawa, "Retinal Thickness Analysis in High Myopia based on Medial Axis Transforms," 2019 41st Annual International Conference of the IEEE



Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2805-2808, 2019, doi: 10.1109/EMBC.2019.8857091.

## Conference Presentations

1. Guangzhou An, Kazuko Omodaka, Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Toru Nakazawa, Hideo Yokota, Masahiro Akiba, Identification of the glaucomatous pathogenesis using machine learning classifiers, Poster Presentation, Riken RAP Symposium, 2016.10.31.
2. Guangzhou An, Kazuko Omodaka, Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Toru Nakazawa, Hideo Yokota, Masahiro Akiba, Evaluation of machine learning classifier for glaucomatous optic nerve head structure using optical coherence tomography images, Oral Presentation, Medical and Biological Imaging, 2017.3.25.
3. 安光州、面高宗子、津田聡、志賀由己浩、高田菜生子、木川勉、中澤徹、横田秀夫、秋葉正博、眼底検査装置からの出力データを用いた緑内障視神経乳頭形状分類の機械学習モデルの構築、口頭発表、医用画像研究会、2017.5.25.
4. 安光州、面高宗子、津田聡、志賀由己浩、高田菜生子、木川勉、中澤徹、横田秀夫、秋葉正博、視神経乳頭形状分類の機械学習モデル構築のための特徴量抽出、口頭発表、眼光学学会、2017.9.3.
5. Guangzhou An, Kazuko Omodaka, Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Toru Nakazawa, Hideo Yokota, Masahiro Akiba, Useful features for automatic classification of glaucomatous optic disc using machine learning, Poster Presentation, Riken RAP Symposium, 2017.11.29.
6. Guangzhou An, Seiji Takagi, Yasuhiko Hiramami, Michiko Mandai, Masayo Takahashi, Yasuo Kurimoto, Hideo Yokota, Masahiro Akiba, Convolutional neural network for classification of normal versus Age-related Macular Degeneration OCT images, Poster Presentation, Riken RAP Symposium, 2017.11.29.
7. 安光州、許沢尚弘、北畑将平、高木誠二、平見恭彦、万代道子、高橋政代、栗本康夫、横田秀夫、秋葉正博、加齢黄斑変性と健常眼の OCT 画像を用いた深層学習の構築と分

- 類、口頭発表、眼光学学会、2018.9.8.
8. 安光州、横田秀夫、秋葉正博、眼科医療画像情報を用いた機械学習による診断支援システム、口頭発表、理研 RAP シンポジウム、 2018.11.20.
  9. Guangzhou An, Kazuko Omodaka, Satoru Tsuda, Yukihiro Shiga, Naoko Takada, Tsutomu Kikawa, Toru Nakazawa, Hideo Yokota, Masahiro Akiba, Glaucoma screening models using machine learning based on optical coherence tomography and color fundus images, Poster Presentation, Riken RAP Symposium, 2019.12.9.

Doctor Thesis, Kobe University

“Hierarchical Machine Learning Models for Ophthalmologic Disease Detection and Classification”, 105 pages

Submitted on Jul, 17<sup>th</sup>, 2020

The date of publication is printed in cover of repository version published in Kobe University Repository Kernel.

© Guangzhou AN  
All Right Reserved, 2020