Applications of Deep Learning to Medical Image Analysis in Ophthalmology

by

Yusuke Kikuchi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering-Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Xin Guo, Chair
Professor Anil Aswani
Professor Zeyu Zheng
Professor Xiaohua Gong

Spring 2022

Applications of Deep Learning to Medical Image Analysis in Ophthalmology

Abstract

Applications of Deep Learning to Medical Image Analysis in Ophthalmology

by

Yusuke Kikuchi

Doctor of Philosophy in Engineering-Industrial Engineering and Operations Research

University of California, Berkeley

Professor Xin Guo, Chair

Medical image is an important information source to understand the patient's condition. As a result, interpreting the medical images is a critical part of the clinical procedure. However, physicians' visual evaluation of medical images in clinics has a few challenges such as the limited human resources, the unavailability of appropriate experts, the increasing number of medical images, and so on.

Deep learning is a subarea of machine learning that uses deep neural networks to learn the patterns behind the given data set. Deep learning showed excellent performances in image analysis problems including medical image analysis. In this dissertation, I propose and evaluate new methods for evaluating the images of three different vision problems in ophthalmology.

The first problem is the early detection of retinopathy of prematurity (ROP). ROP is a leading cause of childhood blindness globally, and early detection is key to preventing ROP to progress to severe conditions. We developed two convolutional neural networks with different depths. The deeper model showed an excellent performance including better metrics than an experienced human expert.

The second problem is about transfer learning in retinal vascular diseases. We propose a transfer learning method that uses the detection of a well-studied retinal vascular disease as a source problem and uses the knowledge to the detection of an under-studied retinal vascular disease. Our proposed method showed better performance with more robustness to the stochasticity in the training process and the reduction of sample size.

The final problem is to predict the treatment response to a drug from the baseline characteristics. Both symbolic features like clinical measurements and medical images are considered for the modeling. To merge the two types of input, we proposed two approaches. The results showed the potential of the proposed method to the problem.

Deep learning is successfully applied to three medical image analysis problems in ophthalmology in this dissertation. These results offer key evidence for further development of deep learning-based medical image analysis systems in the future.

To my mom.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **vi**

1 Outline **1**

2 Background **2**
2.1 Deep Learning 2
2.2 Medical Image Analysis and Deep Learning 14

3 Early Detection of Retinopathy of Prematurity (ROP) in Retinal Fundus Images Via Convolutional Neural Networks **16**
3.1 Introduction 17
3.2 Literature Review 17
3.3 Methodologies 18
3.4 Results 26

4 Transfer Learning for Retinal Vascular Disease Detection **30**
4.1 Introduction 30
4.2 Methodologies 35
4.3 Experiment 39
4.4 Results 40
4.5 Conclusion 41

5 Predicting Treatment Outcomes in Patients with Neovascular Age-Related Macular Degeneration **46**
5.1 Introduction 46
5.2 Methodologies 49
5.3 Results and discussion 57
5.4 Conclusion 58

Bibliography **61**

List of Figures

2.1 Illustration of a neuron 3
2.2 Example of layered neural network 5
2.3 Example of backpropagation algorithm 9
2.4 Example of cross-correlation 10
2.5 Examples of max pooling and average pooling 11
2.6 Example of convolutional neural network 12
2.7 Inception module 13
2.8 Residual connection 14

3.1 Positive image (top) and negative image (bottom) in Data_0; the difference between the vascular area and the non-vascular area is clear; note the apparent thickened ridge (indicated in the red box) in the positive image between the vascular and the non-vascular areas, with no such appearance in the negative image 20
3.2 Positive image (top) and negative image (bottom) in Data_1; note the thickened white line (indicated in the red box) in the positive image between the vascular and the non-vascular areas 21
3.3 The top image is an original color fundus photograph; the bottom image is the same image after preprocessing 22
3.4 Structure of the building block (residual block) of ResNet50. 25
3.5 Comparison between the most experienced ophthalmologist and our models . . . 27
3.6 Feature maps from ROPBaseCNN; the top left is the preprocessed 300×300 image fed into the ROPBaseCNN; the top right is the extracted feature that shows abnormal blood vessel growth; the bottom is the output from the second layer of ROPBaseCNN 28
3.7 Feature maps from ROPResCNN; the top left is the preprocessed 300×300 image fed into the ROPResCNN; the top middle and the top right are the extracted features showing the occurrence of the thickened ridge; the bottom is the output from the fifth layer of ROPResCNN 29

4.1 Examples from ImageNet data set; unlike medical images, the main object is located in the center and images do not have a common structure 33

4.2 Transfer learning; In each domain, samples are drawn from the feature distribution and a machine learning (ML) algorithm learns the relation between the feature and the label; Transfer learning aims to transfer the knowledge learned in the source domain to the target domain 36
4.3 ROP positive sample (left) and negative sample (right); the white line pointed by a red arrow in the positive image is a disease feature called thickened ridge . 37
4.4 DR positive sample (left) and negative sample (right); the pathological region is pointed by red arrows 38
4.5 Changes in four metrics over training sample reduction, with the dark curves averaged over 3 experiments and the area around the curves showing the minimum and the maximum values in 3 experiments 44
4.6 Changes in four metrics over different resolution, with the dark curves averaged over 3 experiments and the area around the curves showing the minimum and the maximum values in 3 experiments; the vertical axis is the size of one side of the input image; for example, the size of 300 means the input image size is 300×300 45

5.1 Structure of AVENUE trial; QXW means the patients received the treatment every X weeks; Patients in the top treatment arm is exclusively treated with ranibizumab and are excluded from the analysis 50
5.2 Administration schedule of each arm; Empty means sham;Arm A is ranibizumab 0.5 mg Q4W; Arm B is faricimab 1.5 mg Q4W; Arm C is faricimab 6.0 mg Q4W; Arm D is faricimab 6.0 mg Q4W for 4 months and faricimab 6.0 mg Q8W for 5 months; Arm E is ranibizumab 0.5 mg Q4W for 3 months and faricimab 6.0 mg Q4W for 6 months 51
5.3 Model stacking; DNN, deep neural network 55
5.4 Model averaging; DNN, deep neural network 56
5.5 General model stacking 57
5.6 Model stacking as a general model stacking; DNN, deep neural network 58
5.7 Model averaging as a general model stacking; DNN, deep neural network 59

List of Tables

3.1 Architecture of ROPBaseCNN 24
3.2 Architecture of ROPResCNN 24
3.3 Experimental results with Data_0 and Data_1 26

4.1 Mean percentage improvement from direct training 40
4.2 Mean percentage reduction of standard deviation from direct training: the standard deviation calculated over different runs and the mean calculated over different training sizes 41
4.3 Percentage changes in all metrics from 100% training size to 10% training size . 41

5.1 Mean metrics of 5-fold cross-validation; numbers in parenthesis are standard deviation; DNN, deep neural networks 59
5.2 Metrics evaluated on the test set; DNN, deep neural networks 60

Acknowledgments

I first would like to show my gratitude to my advisor, Professor Xin Guo. Her guidance with patience was an essential part of my Ph.D. study. Her diligent attitude toward research has always been my ideal model in research. Even outside of research, she taught me many important lessons. I would also like to thank Professor Anil Aswani, Professor Zeyu Zheng, and Professor Xiaohua Gong for serving in my qualifying exam and dissertation committee and giving me helpful feedback in research.

I am also grateful to my collaborators, Guan Wang at Tsinghua–UC Berkeley Shenzhen Institute, Jinglin Yi, Qiong Zou, and Rui Zhou from Affiliated Eye Hospital of Nanchang University, and Jian Dai, Carlos Quezada Ruiz, Michael Kawczynski, and Neha Annegondi from Genentech.

A special thanks to my classmates in the IEOR Ph.D. program: Caleb Bugg, Hansheng Jiang, Heyuan Liu, Igor Molybog, and Mengxin Wang, and research group members: Haoyang Cao, Anran Hu, Mahan Tajrobehkar, Haotian Gu, and Xinyu Li.

Finally, I would like to deeply thank my mom, Mayumi Kikuchi, for her warm support from afar and unconditional love for me.

Chapter 1

Outline

This dissertation is organized in the following way. First, the basics of deep learning and its application to medical image analysis are covered in Chapter 2. The core algorithms of deep learning in image analysis are reviewed, and the connection to medical image analysis is explained. The following chapters cover three different applications of deep learning and machine learning in ophthalmology. Chapter 3 and Chapter 4 are sequential, and Chapter 5 is independent of them. In Chapter 3, an application of deep learning to the early detection of retinopathy of prematurity (ROP) is covered. The motivation for the early detection of ROP is introduced, and a deep convolutional neural network is developed. In Chapter 4, the early detection of ROP is investigated from a perspective of transfer learning. The ROP is regarded as a disease in the category of retinal vascular disease, and a knowledge transfer technique is discussed. Finally, an application of machine learning to a treatment response prediction problem is discussed in Chapter 5.

Chapter 2

Background

In this chapter, we review the basic concepts of deep learning with an emphasis on computer vision and applications of deep learning in medical image analysis.

2.1 Deep Learning

Deep learning is an area of machine learning that uses deep neural networks to process data. A deep neural network is an artificial network with a certain depth. We first review the basic ideas of artificial neural networks.

Basics of Artificial Neural Network

An (artificial) neural network is a mathematical model of a human neural network. The basic building block of a neural network is called a neuron. A neuron has input connections to other neurons and output connections to other neurons as well. A neuron takes inputs, applies a mathematical transform, and outputs the number. To explain the process more precisely, we define

- $n \in \mathbb{N}$: The number of input connections
- $x \in \mathbb{R}^n$: The input vector
- $w \in \mathbb{R}^n$: The weight vector
- $b \in \mathbb{R}$: The bias
- $g \colon \mathbb{R} \to \mathbb{R}$: The activation function, usually nonlinear

Given an input x, a neuron outputs

$$y = g(w^T x + b).$$

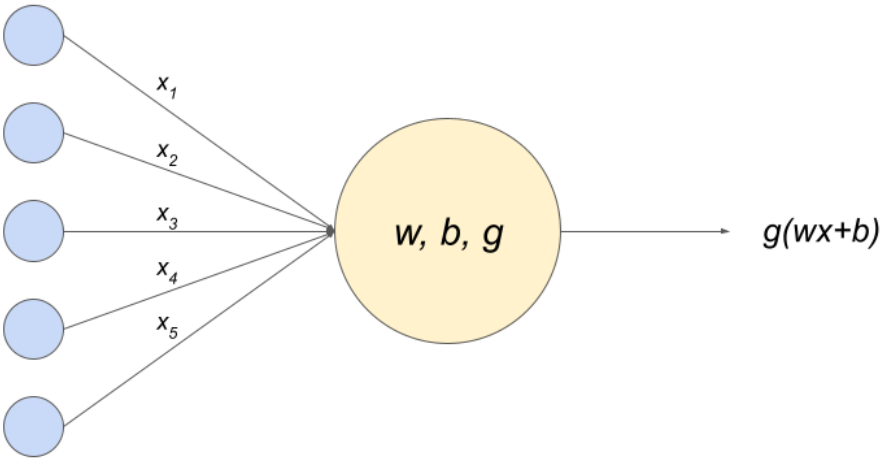


Figure 2.1: Illustration of a neuron

Namely, it applies a linear transform and a nonlinear transform in this order. Some common choices for the activation function g are

- Rectified Linear Unit (ReLU): $g(x) = \max(x, 0)$
- Sigmoid function: $g(x) = \dfrac{1}{1 + e^{-x}}$
- Hyperbolic tangent function: $g(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

Usually, neurons are aligned into layers and a neural network consists of a number of layers. The number of layers is called the depth of the neural network. Each neuron in a layer performs the transform above. Therefore, the transform applied by a layer to the input is

$$y = g(W^T x + b) \in \mathbb{R}^m,$$

where

- $m \in \mathbb{N}$: The number of neurons in the layer, also called the width of the layer
- $W \in \mathbb{R}^{n \times m}$: The weight matrix of the layer, each column corresponds to the weight vector of a neuron

- $b \in \mathbb{R}^m$: The bias vector, each element is the bias of a neuron
- $g \colon \mathbb{R}^n \to \mathbb{R}$: The activation function, usually, the same function is applied to each element

On the whole, a neural network, f, with depth $d \in \mathbb{N}$ is defined by

$$f(x) = o({W^{(d)}}^T y^{(d-1)} + b^{(d)}),$$

where the superscript indicates the number of layer, and $y^{(i)}$ is recursively defined by

$$\begin{aligned} y^{(0)} &= x \\ y^{(i)} &= g({W^{(i)}}^T y^{(i-1)} + b^{(i)}),\ i = 1, \ldots, d-1, \end{aligned}$$

and o is a function called output activation function. The output activation function is chosen based on the nature of the problem. For example,

- Regression problem: Identity (linear) function $o(x) = x$
- Binary classification problem: Sigmoid function $o(x) = \dfrac{1}{1+e^{-x}}$
- Multiclass-classification problem: Softmax function $o(x) = \left[\dfrac{e^{x_1}}{\sum_i e^{x_i}}, \ldots, \dfrac{e^{x_n}}{\sum_i e^{x_i}}\right]^T$

Universal Approximation Theorem

One of the most important theoretical backbones of neural networks is the universal approximation theorem. It states that a large enough neural network can approximate most of the given functions well. There are several versions of the theorem with a different class of functions for the activation function and the target functional space. Here, we cite a version of them.

Theorem 1 *[45] Assume $X \subset \mathbb{R}^n$ is compact. If the activation function is continuous, bounded, and non-constant, then the class of single-layer neural network with linear output activation is dense in $C(X)$, where $C(X)$ is the space of all continuous functions on X.*

Sigmoid function and hyperbolic tangent clearly satisfy the condition of the theorem. However, ReLU is not bounded, so a direct application of this theorem does not generate the universal approximation property for ReLU. The theorem cannot be generalized to any unbounded function, but a discussion in [45] assures that ReLU also has universal approximation property: First, we observe that a linear combination of ReLU satisfies the conditions. For example, a spike function $g(x) = \max(x+1, 0) + \max(-x+1, 0) = ReLU(x+1) + ReLU(-x+1)$ satisfies the conditions. Therefore, the set of functions

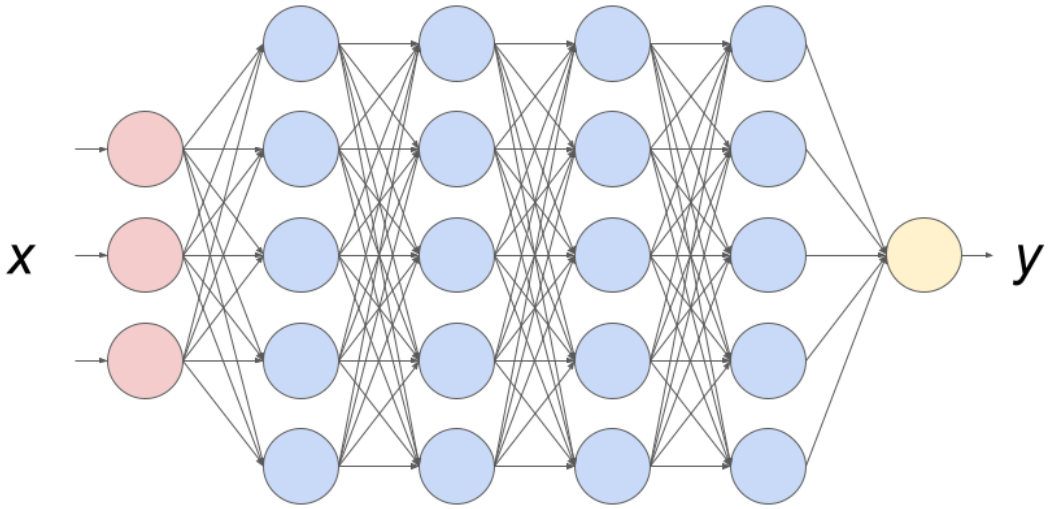


Figure 2.2: Example of layered neural network

that has a representation as a single-layer neural network with activation function g is dense in $C(X)$. Since the set of single-layer neural networks with ReLU activation contains the set of single-layer neural networks with activation function g, the former set is also dense in $C(X)$.

Optimization of Neural Networks

Although a large enough neural network can approximate a given function well, it is not trivial to find the weight, especially for a deep neural network with a lot of parameters. Suppose we are trying to solve a supervised machine learning problem with data $\{(x_i, y_i)\}_{i=1}^n$, where x_i is the input and y_i is the corresponding label. We use a neural network f_θ, where θ represents the parameters of the network, to learn a map between the input x and the corresponding label y. We solve the following optimization problem to achieve it.

$$\min_\theta L(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x_i), y_i)$$

The function L is called a loss function, and l is the sample-wise loss. The choice of L and l depends on the problem. The following list shows the problems and a typical choice of the loss function.

- Regression problem: Mean squared error $l(\hat{y}, y) = |\hat{y} - y|^2$
- Binary classification problem: Binary cross-entropy loss $l(\hat{y}, y) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$
- Multi-class classification: Multi-class cross-entropy loss $l(\hat{y}, y) = -\sum_{k=1}^{K} 1_{y=k} \log \hat{y}_k$, where K is the total number of classes, $1_\bullet$ represents an indicator, and $\hat{y}_k$ is the kth element of y.

The process of finding an optimal weight is called training in machine learning. The training of a neural network is usually done by stochastic gradient descent with the backpropagation algorithm. The gradient descent algorithm updates the parameter by the following rule.

$$\theta \leftarrow \theta - \alpha \nabla_\theta L(\theta).$$

α is called the learning rate or step size, and it is prefixed but can be scheduled. Since computing the exact gradient is computationally expensive with a large number of samples, a stochastic version of the algorithm is used in the training of neural networks. In the mini-batch stochastic gradient descent algorithm, the gradient is estimated by using a few randomly chosen samples. The number of samples used in each step is called the batch size, and it is prefixed. Let B be the batch size and $i_1, \ldots, i_B$ be indices of the random samples. Then, the update rule of the stochastic gradient descent algorithm is

$$\theta \leftarrow \theta - \alpha \nabla_\theta \frac{1}{B} \sum_{j=1}^{B} l(f_\theta(x_{i_j}), y_{i_j}).$$

Namely, the loss function is estimated using randomly chosen samples of size B. The mini-batch stochastic gradient descent algorithm has several advantages over the gradient descent algorithm using all samples like lower requirement of memory and faster convergence. However, this vanilla mini-batch stochastic gradient descent algorithm is rarely used in practice because the loss surface that appears in the deep learning problem has plateaus, saddle points, and other complex structures [31]. The popular choices in practice are

- Stochastic gradient descent algorithm with momentum: It is a common technique to add a momentum term to deal with the bad condition number. The update rule is the following.

$$\begin{aligned} \theta &\leftarrow \theta + m \\ m &\leftarrow \beta m - \alpha g, \end{aligned}$$

  where m is the momentum, α is the learning rate, $\beta \in [0, 1)$ controls the decay of past gradient in momentum, and g is the (estimate of) gradient of the loss function. The momentum term avoids alternating the direction of change too much.

- RMSprop [42]: RMSprop is an adaptive learning rate method. In the gradient descent algorithm, all parameter has the same learning rate α, but an appropriate learning rate is different for different parameters. For example, parameters in a layer close to the input layer usually need smaller change because the change is amplified by the following operations. To account for this, RMSprop uses an estimate of the size of the gradient to approximately normalize the gradient term. The update rule is

$$\begin{aligned} \theta &\leftarrow \theta - \alpha \frac{g}{\sqrt{v}} \\ v &\leftarrow (1-\beta)v + \beta g^2. \end{aligned}$$

  g^2 means taking square element-wise. v is the estimate of the second moment of the gradient, and β balances the current gradient versus the history.

- Adam [51]: Adam improved upon RMSprop by adding an estimate of the first moment of the gradient. Adam updates the parameter by the following rule.

$$\begin{aligned} \theta &\leftarrow \theta - \alpha \frac{\hat{m}}{\sqrt{\hat{v}} + \varepsilon} \\ m &\leftarrow \beta_1 m + (1-\beta_1) g \\ v &\leftarrow \beta_2 v + (1-\beta_2) g^2 \\ \hat{m} &\leftarrow \frac{m}{1-\beta_1^t} \\ \hat{v} &\leftarrow \frac{v}{1-\beta_2^t} \end{aligned}$$

  $\hat{m}$ is an estimate for the first moment and $\hat{v}$ is an estimate for the second moment. ε is a small positive number (for example 10^{-8}) to avoid a division by a very small number, and β_1, β_2 control the amount of update of m, v respectively. Adam is a popular choice for image processing problems.

Now, we look at the computation of the gradient. Let us make an observation first. With $\hat{y} = f_\theta(x)$, the gradient of the loss function with respect to the model parameter is

$$\nabla_\theta l(f_\theta(x), y) = \frac{\partial l}{\partial \hat{y}}(\hat{y}, y) \nabla_\theta f_\theta(x).$$

The second term shows that we need the gradient of f with respect to each parameter in the neural network. The algorithm for computing the gradient on neural networks is called the backpropagation algorithm. As we reviewed, a neural network is a composition of many elementary mathematical transforms. The backpropagation algorithm uses this characteristic of neural networks, and it is essentially an implementation of the chain rule on neural networks. To illustrate how the backpropagation algorithm works, we use a simple

two-layer neural network in Figure 2.3. Given an input x, the neural network compute the output as following.

$$y^{(1)} = g(w_1^T x + b_1),\ y^{(2)} = \begin{bmatrix} g(w_2 y^{(1)} + b_2) \\ g(w_3 y^{(1)} + b_3) \end{bmatrix},\ \hat{y} = y^{(3)} = o(w_4^T y^{(2)} + b_4).$$

Suppose we would like to compute $\dfrac{\partial \hat{y}}{\partial w_1}$. By the chain rule, it is

$$\begin{aligned} \frac{\partial \hat{y}}{\partial w_1} &= \left(\nabla_{y^{(2)}} \hat{y}\right)^T \frac{\partial y^{(2)}}{\partial y^{(1)}} \frac{\partial y^{(1)}}{\partial w_1} \\ &= \frac{\partial \hat{y}}{\partial y_1^{(2)}} \frac{\partial y_1^{(2)}}{\partial y^{(1)}} \frac{\partial y^{(1)}}{\partial w_1} + \frac{\partial \hat{y}}{\partial y_2^{(2)}} \frac{\partial y_2^{(2)}}{\partial y^{(1)}} \frac{\partial y^{(1)}}{\partial w_1}. \end{aligned}$$

This shows that the derivative is a sum of derivatives along all the possible paths from w_1 to $\hat{y}$, and the derivative along a path is a product of the derivatives of each mathematical transform. The backpropagation algorithm on a general neural network works similarly to the example. Basically, it takes derivatives starting from the output node and moves backward until it gets to the parameter of interest. Also, because the algorithm works backward, as opposed to the direction of computing output, it is called backpropagation.

Convolutional Neural Network

A convolutional neural network is a type of neural network specializing in processing data represented as a grid such as image data. A convolutional neural network uses a mathematical operation called convolution to process image data efficiently. To see how convolution works, suppose that the input image, x, has a size of $w_{in} \times h_{in}$, where w is the width and h is the height of the image. A convolution uses a grid of numbers called a filter to extract the image features. Here, we assume the size of the filter, f, is $w_{filter} \times h_{filter}$ ($w_{filter} < w_{in},\ h_{filter} < h_{in}$). Then, the convolution operation $*$ is defined as following. The size of $x * f$ is $(w_{in} - w_{filter} + 1) \times (h_{in} - h_{filter} + 1)$ and the (i, j) element of it is given by

$$(x * f)_{i,j} = \sum_{\alpha} \sum_{\beta} x_{\alpha,\beta} f_{i-\alpha,j-\beta}.$$

Since convolution is commutative, an equivalent definition is given by

$$(x * f)_{i,j} = \sum_{\alpha} \sum_{\beta} x_{i-\alpha,j-\beta} f_{\alpha,\beta}.$$

In most neural network software, a mathematical operation called cross-correlation is used instead of convolution, and also it is called convolution. In this dissertation, we follow this convention. For x and f above, cross-correlation is defined by

$$(x * f)_{i,j} = \sum_{\alpha} \sum_{\beta} x_{i+\alpha,j+\beta} f_{\alpha,\beta}.$$

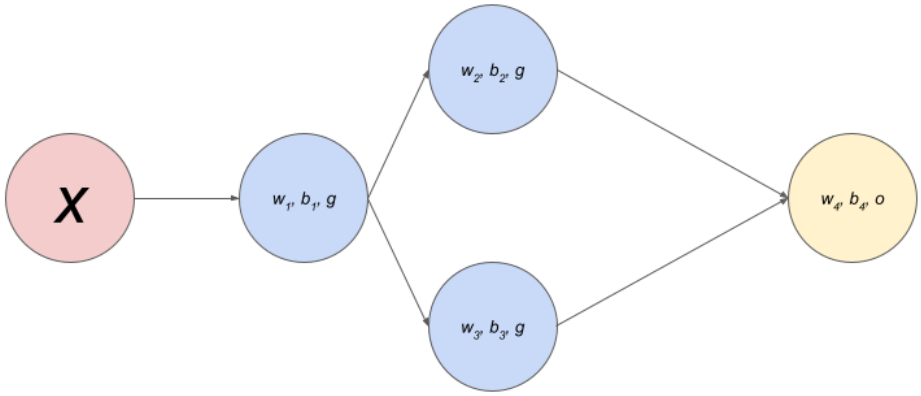


Figure 2.3: Example of backpropagation algorithm

In practice, the input can have multiple channels, the third dimension of the image. For example, a color image has 3 channels for red, green, and blue. The convolution can be easily generalized to this situation. Suppose the input is $x = (x_{i,j,c})$, where the first two indices represent the spatial dimension and the third index represents the channel. The filter f now also has 3 indices as well, $f = (f_{i,j,c})$. Then, the (i, j) element of the convolution is

$$(x * f)_{i,j} = \sum_{\alpha} \sum_{\beta} \sum_{c} x_{i+\alpha,j+\beta,c} f_{\alpha,\beta,c}.$$

A convolutional layer applies this generalized convolution to the inputs. Usually, a number of filters are prepared to capture different features such as lines with different angles. In terms of the size of the output, if a convolutional layer has c_{out} filters, the output has a size of $(w_{in} - w_{filter} + 1) \times (h_{in} - h_{filter} + 1) \times c_{out}$. This means that the spatial dimension is usually reduced. Usually, the number of channels is increased ($c_{in} < c_{out}$). Since the values in the same output channel are generated by using the same filter, those represent some kind of image feature. In other words, a convolutional layer summarizes the local image features and the summary is stored in the channels. After applying several convolutional layers, the output has a small spatial dimension and many channels, and a vector at a specific location has a summary of some global region.

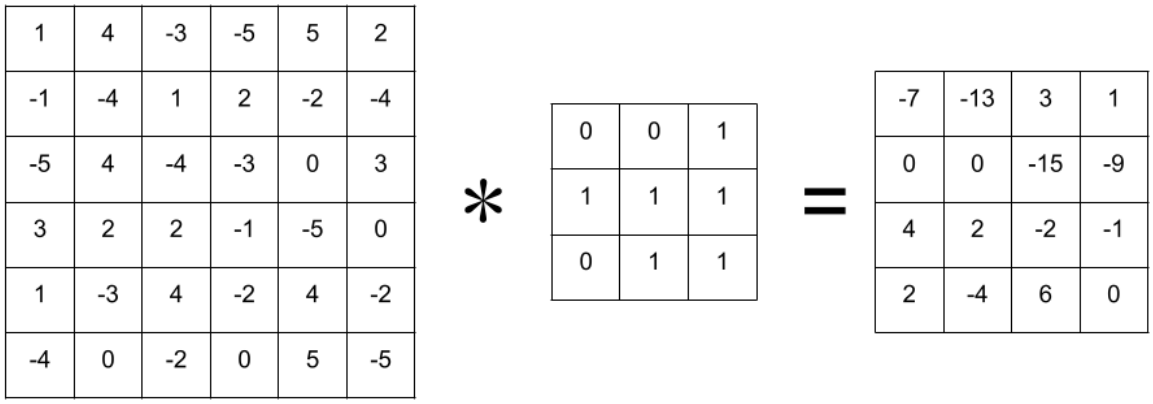


Figure 2.4: Example of cross-correlation

In contrast to the convolutional layer, the standard layer we reviewed is called a fully connected layer. Images can be also processed by the standard neural network we have reviewed if we flatten the input image and regard it as a vector. However, convolutional layer has two key advantages [31].

1. **Shared parameters**. In the standard neural network, a weight is assigned to each pixel. However, this does not make sense too much because image features like blood vessels or disease lesions are usually not tied with specific spatial locations. In the convolutional layer, it has a filter to capture a certain image feature and the filter is applied regardless of the spatial locations. This reduces the number of parameters largely, which is very important in terms of optimization.

2. **Local interaction**. The convolutional operation only considers the local neighbors to compute an output. While, the standard neural network needs all pixel values in the image, which may not make sense in image processing. In image processing, the interaction between the values of two distant pixels is not too important because an image usually consists of local image features.

A parameter called strides is sometimes added to convolution to summarize the image features more aggressively and to lighten the computational cost. The strides parameter has

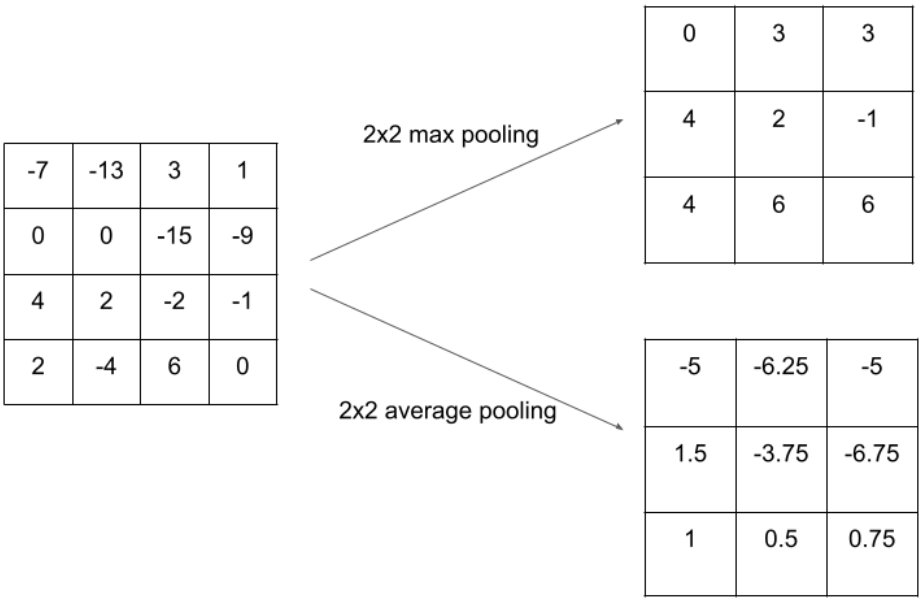


Figure 2.5: Examples of max pooling and average pooling

two dimensions, we write it (s_1, s_2). Roughly, a convolution with strides of (s_1, s_2) applies the filter to the locations of every s_1 pixel in the first dimension and every s_2 pixel in the second dimension. More precisely, the (i, j) element of the output is defined by

$$(x * f)_{i,j} = \sum_{\alpha} \sum_{\beta} \sum_{c} x_{(i-1)\times s_1+\alpha+1,(j-1)\times s_2+\beta+1,c} f_{\alpha,\beta,c}.$$

The size of the output is now $\lfloor (w_{in} - w_{filter} + 1)/s_1 \rfloor \times \lfloor (h_{in} - h_{filter} + 1)/s_2 \rfloor \times c_{out}$.

A pooling layer is usually added after a convolutional layer. The main parameter of the pooling layer is the pooling size. It specifies the size of the neighbor to summarise. A large pooling size means more aggressive summarizing. For each possible location of a grid of the pooling size, it outputs the average or the maximum of the inputs in the grid (Figure 2.5). By adding the pooling layer, the combination of convolution and pooling becomes approximately invariant to a small translation. Pooling can be also combined with strides.

A convolutional neural network (CNN) is a type of neural network specializing in image processing. A CNN consists of two parts. The first part extracts image features and the second part uses the extracted features to make output. The first part is called the feature extractor and the second part is called the decoder (Figure 2.6). However, the architecture of CNN used in real world applications is not just a pile of convolutional layers and fully connected layers. Here, we briefly review the history of the advancement in CNN architectures.

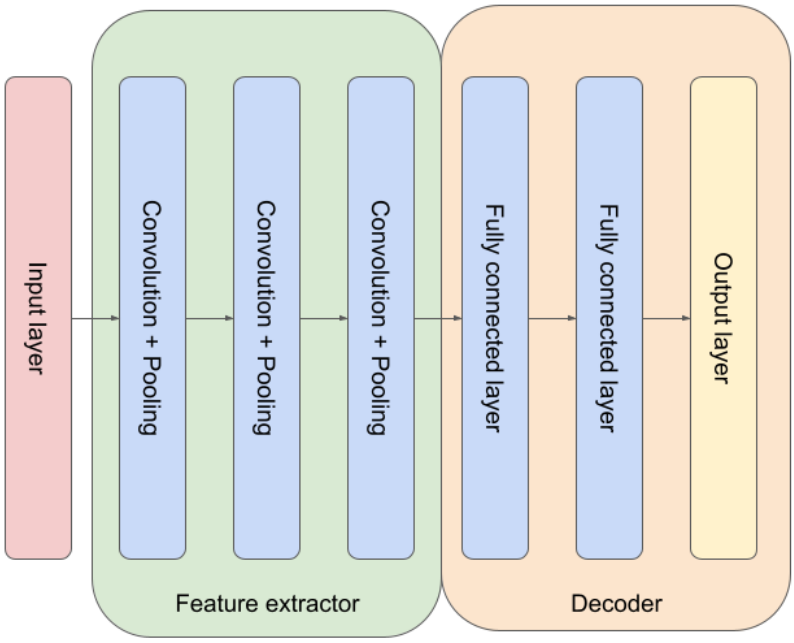


Figure 2.6: Example of convolutional neural network

The advancement of CNN architecture has been led by the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [75] to a great extent. The ImageNet data set consists of more than 1.2 million natural images collected online. Each image is labeled with one of the one thousand classes like koala, sports car, lobster, and so on. This multi-class classification problem has been used as a benchmark of the performance of CNNs. The following three are the architectures that achieved state-of-art performance at the time of their publication, and also the last two are widely used in the application in medical image analysis.

- VGG [82]: The VGG net has significantly more depth compared to the state-of-art architecture before VGG. The VGG architecture uses small convolutional filters (3×3) and has 16-19 weighted layers. VGG won the ILSVRC competition in 2014. The main contribution of VGG is to show the effectiveness of using a deep CNN with small convolutional filters.

- InceptionNet [86]: The basic building block of the inception architecture is called Inception module (Figure 2.7). There are two important features of the Inception module. The first point is that it applied different sizes (1×1, 3×3, 5×5) of convolution at the same time. This helps the neural network to learn features at different scales. The second point is in the 1×1 convolution before 3×3 or 5×5 convolution. The purpose of the 1×1 convolution is to reduce the number of input channels, which

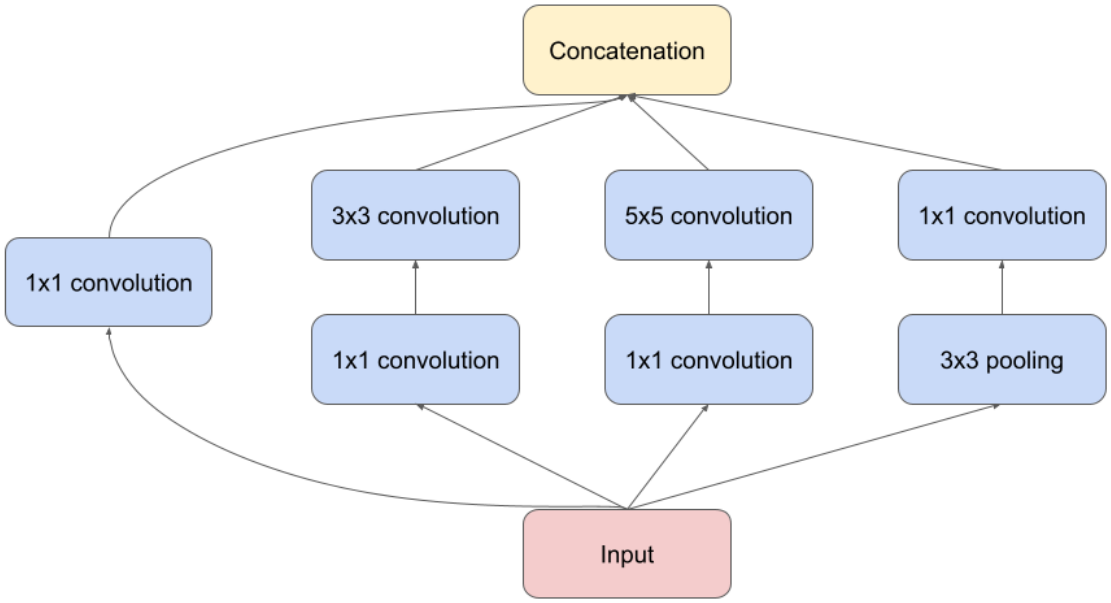


Figure 2.7: Inception module

reduces the number of parameters in the convolutional layer largely. As a result, InceptionNet achieved a better result than the previous state-of-art architecture with 12 times fewer parameters.

- ResNet [37]: Although VGG showed the benefit of deepening architecture, that approach had two problems. The first problem is the vanishing gradient problem, and the second problem is that adding layers does not necessarily improve the performance. The authors of ResNet tackled those problems by adding residual connections. The idea is simple: Each layer learns the residual of the previous layers. Consider a CNN and take a layer in it. Suppose x is the output of the previous layer, or equivalently, the input of the layer. Then, the layer learns $f(x) = h(x) - x$, where h is a desired mapping of the layer. A visual explanation of residual connection is in Figure 2.8. Because the residual connection lets the gradient go through each layer, the gradient vanishing problem is less challenging for ResNet. Also, because each layer learns the residual of the previous layers, adding a layer is theoretically guaranteed to improve the performance given that the neural network is optimized well. As a result, the depth of ResNet architecture was increased to 50 or 101. Later, with more improvement in the architecture, the number of layers was increased to 152 successfully [38].

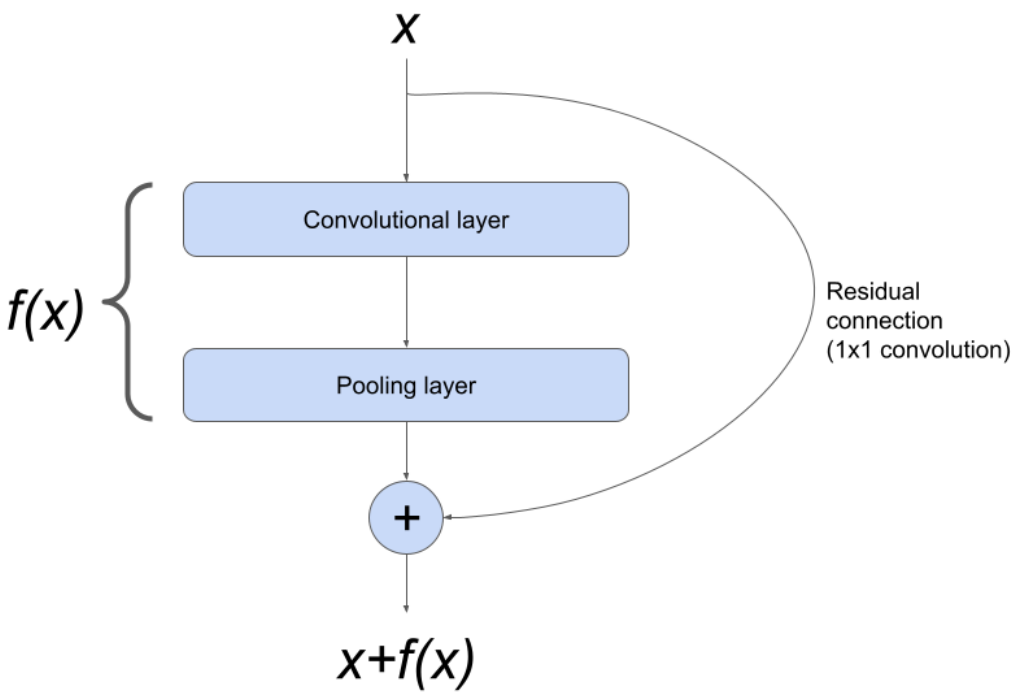


Figure 2.8: Residual connection

2.2 Medical Image Analysis and Deep Learning

Medical imaging is the technique and process of generating an image of the human body or part of it for clinical or medical uses. Medical images show the inside of the human body which is not visible from the outside. Hence, medical images are a vital part of clinical procedures. Indeed, the annual number of medical images taken in the U.S. keeps increasing in recent years [83].

The interpretation of the medical images has been done by human physicians and radiologists. However, it is predicted that there will be a huge shortage of radiologists in near future [1]. Artificial intelligence (AI) based on deep learning for medical image analysis is thought to improve this situation. Furthermore, AI can improve accessibility to the healthcare system. The general advantages of AI over humans are

- Efficiency: In an appropriate setting, a neural network takes only less than a second to make a prediction, whereas a human doctor needs more time to grasp the image features and make a decision. In addition, a deep neural network will never get tired after the repetition of similar works.

- Short Training Time: With a labeled data set, training a neural network takes much less time than training a human expert. Usually, the development of a neural network

takes up to months, while training a human expert takes years. Also, the performance of a trained neural network will not deteriorate as long as the data source stays the same.

- Transferability: Because neural networks are registered on computers, it is easy to copy and send them. Sending a human expert is not that easy as it involves many other factors like personal matters. This could increase the accessibility to healthcare and expert knowledge especially in developing countries or in rural areas.

The eye is a small and sensitive organ, so medical imaging is very important in ophthalmology. In ophthalmology, various types of medical image modalities are used based on the disease. Among those, most of the current applications of deep learning use color fundus photographs or optical coherence tomography. As in the other medical fields, interpreting those images is not an easy task. Among all the subfields in medicine, ophthalmology is a field with one of the highest numbers of deep learning applications. The following overviews the existing application of deep learning in ophthalmology.

- Diabetic retinopathy (DR) is a diabetic complication that affects the retina. DR is one of the leading causes of legal blindness among working age adults [65]. Some of the earliest applications of deep learning in ophthalmology were done in the detection of DR in color fundus photograph [2, 28, 33, 91].

- Retinopathy of prematurity (ROP) is a retinal disease seen in prematurely-born infants and is the most common cause of vision loss in children [40]. The applications of deep learning in ROP have two major directions. A direction focuses on detecting and grading ROP. The other direction focuses on detecting ROP plus disease, which is a key feature to choose the treatment. The works in the first direction are [92, 94], and those in the second direction are [10, 68, 87, 96, 99].

- Age-related macular degeneration (AMD) is a retinal disease affecting macular and affects the central vision in elderly generation [39]. AMD is a leading cause of legal blindness in elderly people. AMD has diverse list of applications: detection of AMD [11, 32, 91], disease progress prediction [97], treatment response prediction for neovascular AMD [48], and lesion segmentation in dry AMD [18].

- Other applications include: detection of glaucoma [55, 91], detection of diabetic macular edema [100, 93], detection of cataract [20], segmentation in optical coherence tomography [53, 64].

Chapter 3

Early Detection of Retinopathy of Prematurity (ROP) in Retinal Fundus Images Via Convolutional Neural Networks

In this chapter, an application of deep learning in detecting retinopathy of prematurity (ROP) is covered.

ROP is an abnormal blood vessel development in the retina of a prematurely-born infant or an infant with low birth weight. ROP is one of the leading causes of infant blindness globally. Early detection of ROP is critical for at-risk infants to receive appropriate treatment and to slow down and avert the progression to vision impairment caused by ROP. Yet there is limited awareness of ROP even among medical professionals, especially in developing countries. Consequently, the data set for ROP is limited if ever available and is in general imbalanced in terms of the ratio between negative images and positive ones.

In this study, we formulate the problem of detecting ROP in retinal fundus images in an optimization framework and apply state-of-art convolutional neural network techniques to solve this problem. In addition, our study shows that as the network gets deeper, more significant features can be extracted for early ROP diagnosis.

Our work, aided by advanced machine learning techniques, achieves for the first time the perfect sensitivity score for early ROP diagnosis, along with comprehensive studies showing significantly improved performance over human diagnosis. Moreover, our algorithm is capable of extracting features of the elevated ridge in the retina, making our prediction results explainable for clinical use.

3.1 Introduction

Retinopathy of prematurity (ROP) is an abnormal blood vessel development in the retina of prematurely-born infants or infants with low birth weight; in all term infants, the vasculature in the retina is fully established, while the development of retinal vasculature in premature infants is not complete, and it possibly progresses to abnormal development [22]. ROP is caused by increased angiogenesis factors as an effect of the decreased amount of oxygen after being discharged from an oxygen chamber [35]. ROP can lead to permanent visual impairment and is one of the leading causes of infant blindness globally. In the U.S., it is estimated that 184,700 preterm infants developed ROP and that 20,000 of them progressed to blind or impaired vision as of 2010 [8]. In developing countries, higher neonatal survival rates have significantly increased the number of premature infants and consequently, the number of ROP for infants [81]. It is estimated that nineteen million children are visually impaired worldwide [8], among which ROP accounts for six to eighteen percent of childhood blindness [29]. Early treatment has confirmed the efficacy of treatment for ROP [21]. Therefore, it is crucial that at-risk infants receive timely retinal examinations for early detection of potential ROP.

Early detection of ROP, however, faces significant challenges. In developing countries, the effective screening system of ROP for preterm infants is not well established due to not full awareness of ROP among pediatricians. In addition, infants' inability of active participation imposes more difficulties in medical diagnosis. To minimize the number of missed diagnoses for ROP in infants, the clinical screening requirement for ROP has an exceptionally high sensitivity level, generally higher than the medical standard of 95%.

3.2 Literature Review

The effectiveness of using deep learning in image analysis in retinal fundus images was first confirmed by [33] for diabetic retinopathy (DR). The imaging modality used in the screening of DR is called color fundus photograph, which is an image of the retina taken by a microscope. The color fundus photograph is also used in the screening of ROP. In the study, the authors developed a convolutional neural network (CNN) using 128,175 images, and the CNN was validated by using 2 data sets. The results showed a high sensitivity score and a high specificity score.

The deep learning-based approach was soon adopted in the area of ROP. The applications of deep learning in ROP have two major directions. The first direction is the detection and grading of ROP. In other words, it is a classification problem in which the output is ROP positive/negative or the stages of ROP. The first work in this direction is by [94], which is based on a training dataset of 742 ROP cases with 5,967 images and 1,484 normal cases with 7,559 images. Their model has achieved impressive results: 96.62% in specificity and 99.32% in sensitivity, and outperforming two ophthalmologists out of three. Recently, [92] has developed convolutional neural networks for ROP detection and grading. Their focus

is to classify the input fundus image into four categories (i.e., normal, mild, semi-urgent, and urgent) based on the requirements of clinical treatment. Their training dataset includes 26,459 images, without explicitly specifying the ratio between normal cases and ROP. Their model has achieved an accuracy of 90.3%, an sensitivity of 77.8%, an specificity of 93.2%, and an F1-score of 76.1% for classifying the ROP cases, with the performance of the system compared to two human experts. The experts resulted in accuracy scores of 0.902 and 0.898, sensitivity scores of 0.748 and 0.659, specificity scores of 0.934 and 0.923, and F1 scores of 0.743 and 0.682.

The second direction is the detection of ROP plus disease. ROP plus disease refers to severe vascular changes and it is often an early sign of severe ROP and potential vision loss [84]. It also plays an important role in the choice of treatment. The first work in this direction is [96]. In that study, the authors developed a CNN which is based on the InceptionNet. The size of the development data set was about 1,500, and the performance was evaluated using 9-fold cross-validation. Also, the learned features were analyzed, and it was shown that the neural network pays attention to vasculature to make an output. Although the performance is not satisfactory compared to the medical standard, this study paved the way for using deep learning in ROP. An extensive investigation of the effectiveness of deep learning in detecting ROP plus disease was done in [10]. This is a large-scale study using 8 study centers across North America. The system consists of two steps, the first step is to segment the blood vessels, and the second step is to detect the plus disease from the segmentation map. The system is evaluated by a 5-fold CV and resulted in the area under receiver operator characteristic curve (AUROC) of 0.98 for detecting the plus disease. This study is further extended in [89] to construct a severity score.

3.3 Methodologies

Data Collection and Processing.

To develop a model for ROP detection, color fundus photographs were collected from infants in the Affiliated Eye Hospital of Nanchang University, which is an AAA (i.e., the highest ranked) hospital in China. All images were de-identified according to patient privacy protection policy, and the ethics review was approved by the ethical committee of Affiliated Eye Hospital of Nanchang University (ID: YLP202103012).

For this work, two data sets were collected. The first de-identified data set, which is called Data_0, consists of random samples of color funds photographs taken at the hospital between 2013 and 2018. A single type of fundus camera, Clarity Retcam3, was used with 130 degree fields of view. All operators are professionally trained. Data_0 includes 2021 negative images and 382 positive images with the resolution of 1600×1200. Images in this data set share a common characteristic: the boundary between the vascular and the non-vascular areas is clear, so is the color difference. As shown in Figure 3.1, there is a clear white dividing line, called the demarcation line, between the vascular and the non-vascular areas of the

peripheral retina. In the early stage of ROP, this demarcation line will get thicker until a ridge occurs. As the ridge gets thicker, a proliferation of abnormal blood vessels will cause the retinal blood vessel to expand, eventually leading to ROP. The appearance of thickened ridges is the main indicator used by ophthalmologists to diagnose ROP. Note that there is no such a thickened ridge in the negative image.

The second de-identified data set, which is called Data_1, consists of 461 negative images and 498 positive images with the resolution of 1600×1200. To collect Data_1, a variety of 130 degree fields cameras were used, including CLARITY Retcam3, SUOER SW-8000, and MEDSO ORTHOCONE RS-B002. This set of data is characterized by the similar appearance of the vascular and non-vascular areas and with similar colors. However, the boundary between the vascular area and the non-vascular area is much clearer than that in Data_0. Figure 3.2 shows a negative image and a positive image from Data_1. In both of the data sets, different images were taken from different patients.

All images were graded by ophthalmologists for the presence of ROP severity and the image quality using an annotation tool. The annotation tool was designed by ophthalmologists and implemented by ourselves. ROP severity was graded as positive or negative. Image quality was assessed by graders, with images of adequate quality considered gradable. The reliability of the grading result was assessed by four ophthalmologists. The final grading results, for which the diagnosis from the hospital agreed with the majority of diagnoses from these ophthalmologists, were used for each color fundus photograph.

Data Processing and Data Rebalancing

The datasets of Data_0 and Data_1 are clearly imbalanced. For instance, negative images in Data_0 dataset are five times more than positive images. This imbalance adds difficulty in the training process: as convolutional neural networks are more exposed to negative images, the training process may be significantly biased towards the negative class. This issue of data imbalance is very common among medical data. There are several data processing approaches to ease the imbalance, including under-sampling [34], re-sampling and fine-tuning [36], oversampling [58], and weight balance and class balance [101].

To mitigate the imbalance problem, we design a hybrid method with a combination of several techniques:

(i) Data enhancement: all images in the data set are first enhanced by brightness adjustment and random flipping. (See Figure 3.3). Afterwards, all images are resized into 300×300.

(ii) Tuning sampling ratio and class weights: we use different class weights in the loss function, to be introduced in the next section. We over-sample the enhanced positive images and re-sample the enhanced negative images so that the numbers of positive images and negative images in the sample batch are kept proportional to the inverse of their class weights. We experiment with different ratios through grid search in the

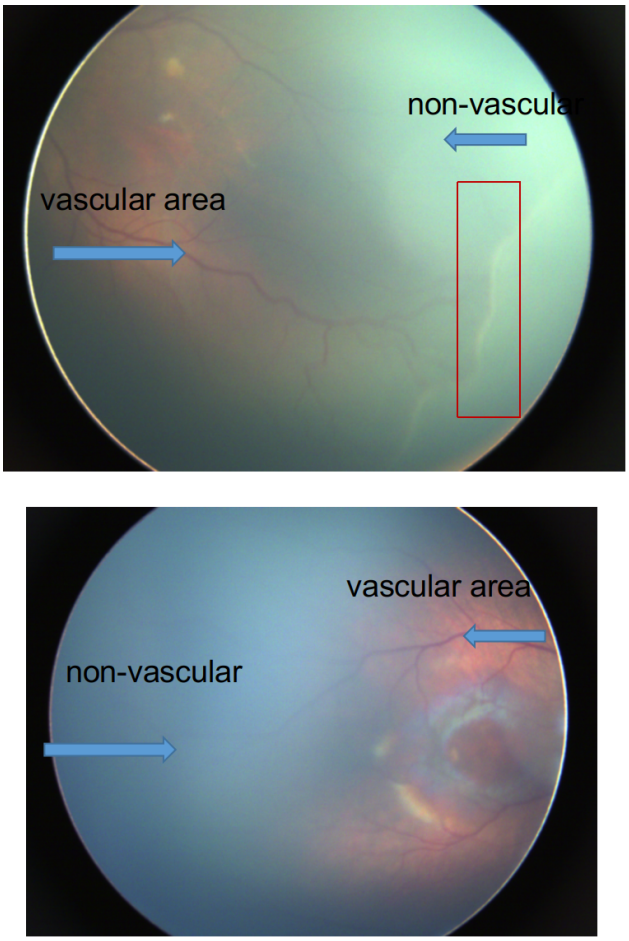


Figure 3.1: Positive image (top) and negative image (bottom) in Data_0; the difference between the vascular area and the non-vascular area is clear; note the apparent thickened ridge (indicated in the red box) in the positive image between the vascular and the non-vascular areas, with no such appearance in the negative image

validation set, and eventually set the ratio of positive and negative images to be $1:2$ in the training process.

Problem formulation

We formulate the problem of detecting ROP as a binary classification problem, where the positive images and the negative images are labeled as 1 and 0, respectively. That is, given a fundus image, instead of labelling the image as either 1 or 0, we assign a score in terms of probability between 0 and 1 to the input image. The higher the score, the higher the

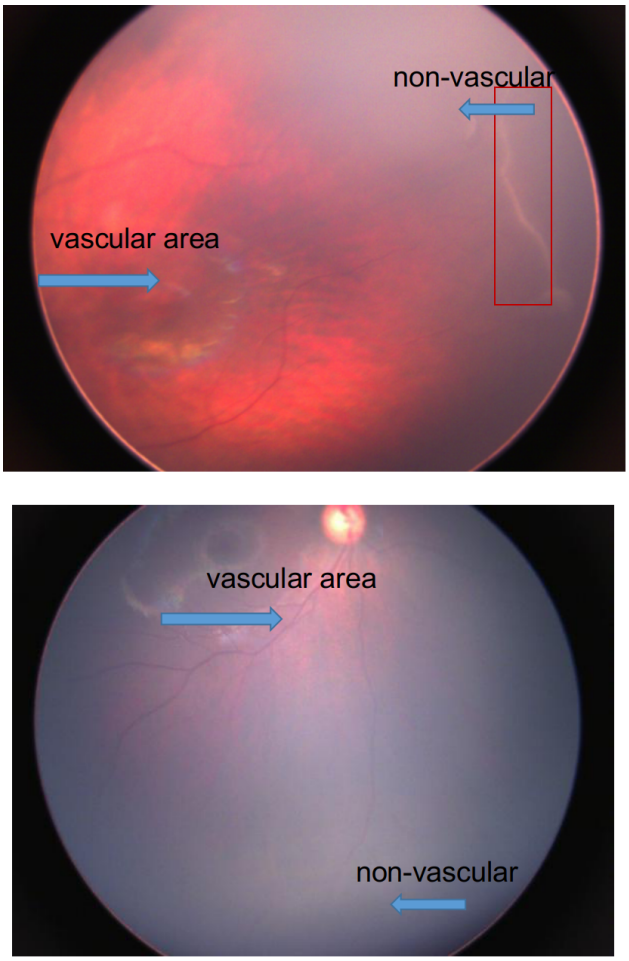


Figure 3.2: Positive image (top) and negative image (bottom) in Data_1; note the thickened white line (indicated in the red box) in the positive image between the vascular and the non-vascular areas

probability that the image has an ROP (i.e., ROP positive). When assigning the label for the input image, if the probability is higher than 0.5, it is then labelled as positive; otherwise, it is negative. This is a natural choice for the neural network which requires the output to be a continuous variable.

Now, suppose the probability is parametrized by θ such that it is denoted as p_θ. This set of parameters could be interpreted as various factors contributing to the probability of having an ROP. Then the training stage is to minimize the cross-entropy loss function over the set of parameters θ. That is, denote the distribution of the pair of the image x and the

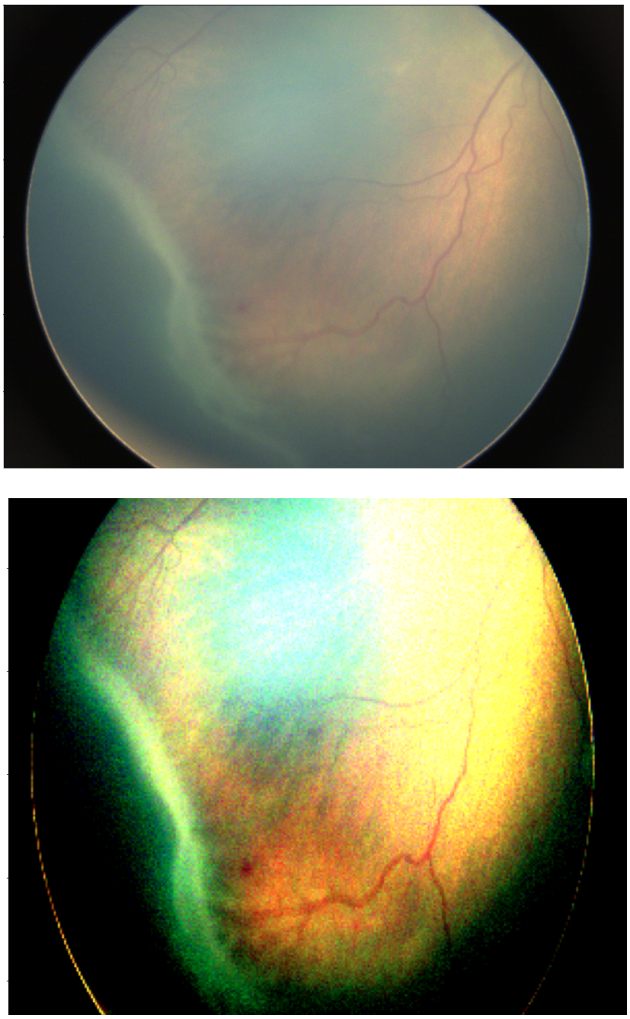


Figure 3.3: The top image is an original color fundus photograph; the bottom image is the same image after preprocessing

0-1 label y by p_{data}, then the training process is to solve the following optimization problem,

$$\min_{\theta} \mathbb{E}_{(x,y)\sim p_{\text{data}}}[-y \log p_\theta(x) - (1-y)\log(1-p_\theta(x))].$$

Given the limited amount of available data and hence the possible issue of overfitting, we add a kernel regularization. In particular, we adopt the L^2 regularization on the weight matrices of the fully connected layers. For each fully connected layer with weight matrix $W = (w_{ij})$, we add the following regularization term to the loss function

$$\lambda \|W\|_{L^2} = \lambda \sum_{i,j} w_{ij}^2,$$

where λ is a hyper parameter to adjust the scale of the regularization. Finally, we adjust the loss function by the $1:2$ class weight from the data processing stage, so that the final optimization problem is to solve the following regularized cross entropy loss function,

$$\min_{\theta} \mathbb{E}_{(x,y)\sim p_{\text{data}}}[-y \log p_\theta(x) - 2(1-y)\log(1-p_\theta(x))] + \lambda \|W\|_{L^2}. \tag{3.1}$$

Model Architectures

Our approach is to use a deep neural network to represent p_θ. In this study, two neural networks are developed, and their performances are compared. The first neural network is a shallow neural network with two convolutional layers called ROPBaseCNN. The architecture detail is shown in 3.1. The first four layers extract image features by applying convolution and pooling, and the following fully connected layers transform the information to output. The fully connected layer is known to be prone to overfitting. To prevent that problem, dropout layers [85] are inserted before fully connected layers in addition to L^2 regularization on the kernel. A dropout layer is a way of regularization and it randomly sets a prefixed portion of inputs zero in the training process. In the inference process, a dropout layer becomes an identity.

The second one is a deep CNN based on ResNet50 [37] called ROPResCNN. ROPResCNN consists of the convolutional part of ResNet50 (i.e. ResNet50 without the output layer), global average pooling [57], and an output layer with Sigmoid activation in this order. A global average pooling layer is a type of pooling layer. It averages all features within a feature map (i.e. a channel), hence the output is a vector. The details of the architecture of ROPResCNN is shown in Table 3.2 and Figure 3.4.

Both models are implemented with Keras (version 2.3.1) [13] on Python (version 3.8) using TensorFlow (version 2.2) [60] as the backend.

Optimization

We used the Adam algorithm [51] to solve our optimization problem (3.1). Adam algorithm is a variant of stochastic gradient descent algorithm, and it combines a momentum method

Type of layer	parameters
Input	shape=(300,300,3)
Convolution	filters=32, kernel size=3×3, strides=(2,2), activation=ReLU
Max pooling	pool size=2×2, strides=(2,2)
Convolution	filters=64, kernel size=3×3, strides=(2,2), activation=ReLU
Max pooling	pool size=2×2, strides=(2,2)
Dropout	dropping probability = 0.25
Flatten	none
Fully connected	neurons=128, activation=ReLU, $\lambda = 0.001$
Dropout	dropping probability = 0.5
Fully connected	neurons=64, activation=ReLU, $\lambda = 0.001$
Output(Dense)	shape=(1), activation=Sigmoid

Table 3.1: Architecture of ROPBaseCNN

Type of layer/block	parameters
Input	shape=(300,300,3)
Residual block	$f = 64$, repeat 3 blocks.
Residual block	$f = 64$, repeat 4 blocks.
Residual block	$f = 128$, repeat 6 blocks.
Residual block	$f = 512$, repeat 3 blocks.
Pooling	global average pooling.
Output(Dense)	shape=(1), activation=Sigmoid

Table 3.2: Architecture of ROPResCNN

and an adaptive learning rate method. It uses the first order and the second order moments to adaptively modify the raw gradient. Empirically, it is known that Adam algorithm is efficient in training CNNs. The parameters of Adam algorithm used here are: learning rate $\alpha = 0.001$, and the exponential decay rate for the first and the second moment $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively.

To train our models more efficiently, we adjust the learning rate with respect to the validation loss. More specifically, the learning rate is reduced by 20% when the validation loss does not improve for 5 epochs.

Training Settings

Data_0 is used for training ROPBaseCNN and split into three sets: the training set has 187 positive images and 990 negative images, the validation set has 80 positive images and 425

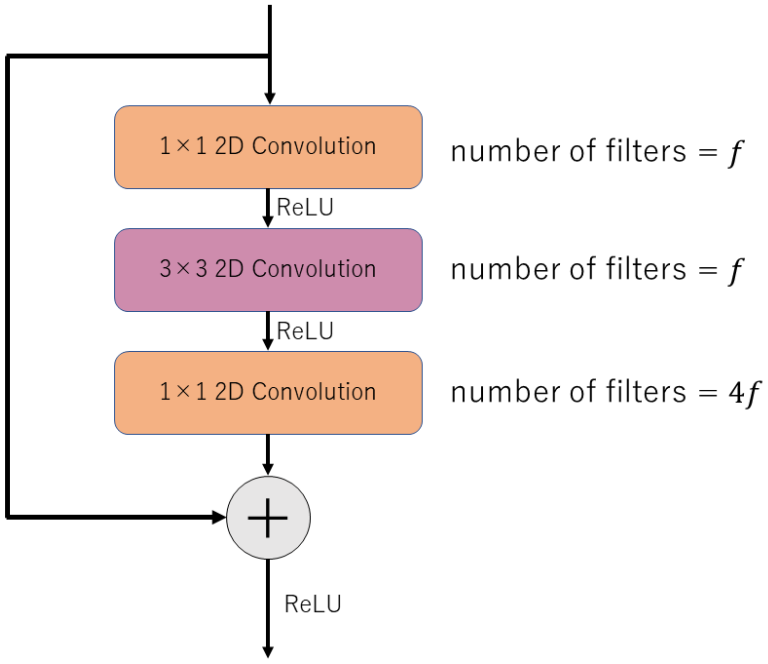


Figure 3.4: Structure of the building block (residual block) of ResNet50.

negative images, and the testing set has 115 positive images and 606 negative images with held-out class labels. A combination of Data_0 and Data_1 is used for ROPResCNN, and is split into three sets: the training set of 431 positive images and 1216 negative images, the validation set of 185 positive images and 521 negative images, and the testing set of 264 positive images and 745 negative images with held-out class labels.

Preferably, a larger image data set is needed to train a deep neural network such as ROPResCNN. Following a standard approach in the field, we used a transfer learning approach from a large image data set. That is to use a pretrained weight on ImageNet data set as the initial point of the optimization.

A single GTX 1080 GPU and 8GB of memory are used for training both neural networks. With appropriate data processing, one GPU turns out to be sufficient to fit the training process with 3000 images per epoch. For ROPBaseCNN, the batch size is 32, and the training is stopped after 25 epochs; for ROPResCNN, the batch size is 64, and the training is stopped after 30 epochs.

model	ROPBaseCNN	ROPBaseCNN	ROPResCNN
Train data	Data_0	Data_0+Data_1	Data_0+Data_1
Test data	Data_0	Data_0+Data_1	Data_0+Data_1
Precision	0.95	0.81	0.96
Sensitivity	0.91	0.79	1.0
Specificity	0.91	0.93	0.96
Accuracy	0.93	0.89	0.98
F1 score	0.93	0.80	0.98

Table 3.3: Experimental results with Data_0 and Data_1

Evaluation Metrics

We adopt the following five standard metrics to evaluate the performance of our models: precision, sensitivity, specificity, accuracy, and the F1 score. Here

$$\text{Precision} = \frac{TP}{TP+FP},\ \text{Sensitivity} = \frac{TP}{TP+FN},\ \text{Specificity} = \frac{TN}{TN+FP},$$
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN},\ \text{F1} = 2\frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}},$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. As it is described in the introduction, sensitivity is an important metric in the detection of ROP. Also, note that the accuracy score alone does not describe the model's performance well because the data set is imbalanced.

In addition, we calculate the following metric called error reduction to quantify the improvement based on the human expert's performance.

$$\text{Error reduction} = \frac{\text{human error} - \text{model error}}{\text{human error}}.$$

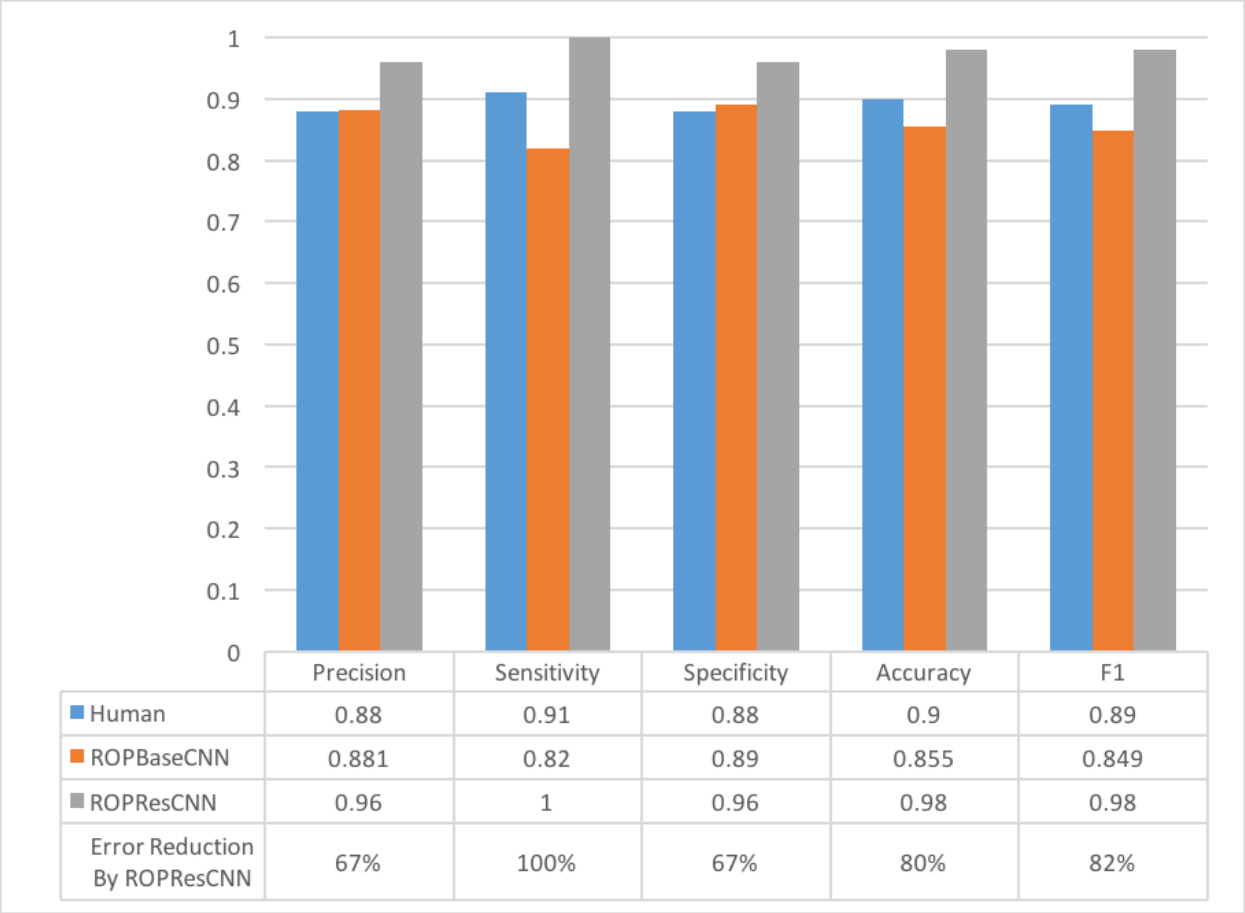
The error reduction is calculated for each of the five metrics.

3.4 Results

The results of ROPBaseCNN-based model with Data_0 are summarized in the first column of Table 3.3, the results of ROPBaseCNN-based model with the combined data sets Data_0 and Data_1 are summarized in the second column of Table 3.3; and the third column shows the results of ROPResCNN-based model with both data sets Data_0 and Data_1.

Next, we take 200 infants' color fundus photographs with confirmed grading results by ophthalmologists. The grading results of the most experienced ophthalmologist are then compared against those generated by our models. Figure 3.5 gives the detailed performance

comparison. We see that ROPBaseCNN-based model manages to achieve comparable performance with the ophthalmologist, especially in terms of precision and specificity.



	Precision	Sensitivity	Specificity	Accuracy	F1
Human	0.88	0.91	0.88	0.9	0.89
ROPBaseCNN	0.881	0.82	0.89	0.855	0.849
ROPResCNN	0.96	1	0.96	0.98	0.98
Error Reduction By ROPResCNN	67%	100%	67%	80%	82%

Figure 3.5: Comparison between the most experienced ophthalmologist and our models

ROPResCNN-based model dominates the ophthalmologist by a wide margin. Its performance on the combined Data_0 and Data_1 dataset shows a perfect score on sensitivity, 96% specificity, 96% precision, and across-the-board improvement of roughly 10% when compared with the ophthalmologist. Most importantly, it cut human errors by over 67% in all categories, and in particular eliminates completely the error in the category of sensitivity, the most critical requirement for the diagnosis of ROP.

Feature Map

A feature map is the output of a layer. We examined the feature maps to see what features are captured by the two models. The feature map from ROPBaseCNN, shown in Figure

3.6, captures an implicit indicator of ROP, the abnormal blood vessel growth. We note that such a disorder from the retinal fundus image has not been used by ophthalmologists as a standard indicator for diagnosis of ROP since it is hard to see by human eyes.

The feature map from ROPResCNN demonstrates that ROPResCNN-based model succeeds in learning and explicitly capturing the well-accepted indicator for the medical diagnosis of ROP: the thickened ridge, see Figure 3.7.

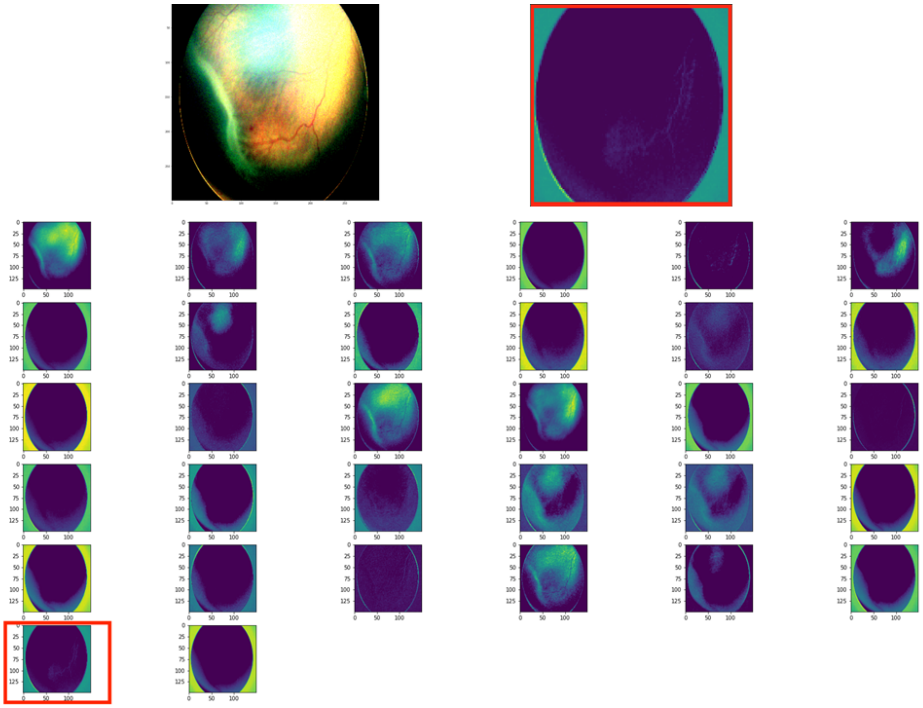


Figure 3.6: Feature maps from ROPBaseCNN; the top left is the preprocessed 300×300 image fed into the ROPBaseCNN; the top right is the extracted feature that shows abnormal blood vessel growth; the bottom is the output from the second layer of ROPBaseCNN

Discussions

There are further questions that are worth future study. One is the stage classification for ROP and another is the localization and the segmentation of the disease feature, both of which are useful for clinical purposes, see for instance [19, 61] and [92]. One may also investigate if the deep network technique in this study may be further developed to extract more significant features for ROP or other related eye diseases such as ROP plus, the latter of which has generated substantial research interests, see for instance [10], [92], [96], and [89]. Finally, we note that it is desired to repeat the experiments to estimate the variance of our method, although it was not feasible due to the limited samples.

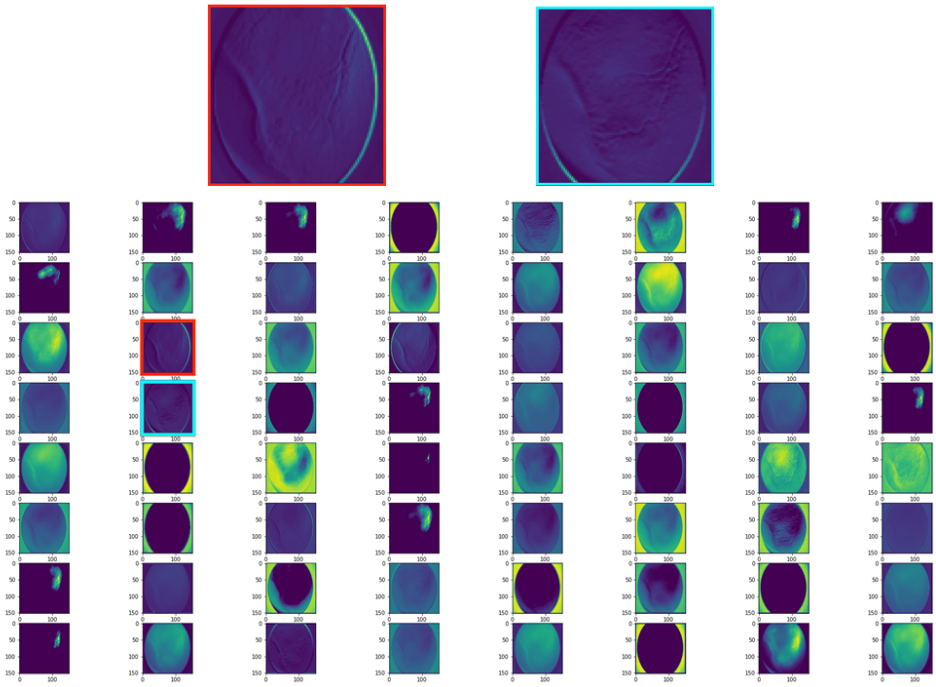


Figure 3.7: Feature maps from ROPResCNN; the top left is the preprocessed 300×300 image fed into the ROPResCNN; the top middle and the top right are the extracted features showing the occurrence of the thickened ridge; the bottom is the output from the fifth layer of ROPResCNN

Conclusion

Two CNNs with different depths are developed to detect ROP in color fundus photographs in this study. Our approach can provide accurate and early detection of ROP with a perfect sensitivity score and excellent scores in specificity and precision. An interesting finding is that the shallow network model learns the vessel feature, while the deep network model manages to learn the thickened ridge feature.

Chapter 4

Transfer Learning for Retinal Vascular Disease Detection

In this chapter, we approach the problem of the detection of ROP from a different perspective.

Retinal vascular diseases affect the well-being of human body and sometimes provide vital signs of otherwise undetected bodily damage. Recently, deep learning techniques have been successfully applied for detection of diabetic retinopathy (DR). The main obstacle of applying deep learning techniques to detect most other retinal vascular diseases is the limited amount of data available.

In this chapter, we propose a transfer learning technique that aims to utilize the feature similarities for detecting retinal vascular diseases. We choose the well-studied DR detection as a source task and identify the early detection of retinopathy of prematurity (ROP) as the target task. Our experimental results demonstrate that our DR-pretrained approach dominates in all metrics the conventional ImageNet-pretrained transfer learning approach, currently adopted in medical image analysis. Moreover, our approach is more robust with respect to the stochasticity in the training process and with respect to reduced training samples.

This study suggests the potential of our proposed transfer learning approach for a broad range of retinal vascular diseases or pathologies, where data is limited.

4.1 Introduction

Problem

The health of eyes is beyond being simply an integral part of the well-being of human body. In particular, retinal vascular disease, referring to a condition that affects the blood vessels of the retina, is well recognized to provide early signals of bodily damage. For example, a retinal vascular disease called hypertensive retinopathy is sometimes the only symptom of a person with a serious cardiovascular condition [3]. This is because the arrangement of blood

vessels at the back of the eye, known as the retinal vasculature, is closely connected to the health of heart [16] and also because the retina is the only part where the vasculature is visible from the outside. Diabetic retinopathy (DR) is another example of retinal vascular disease., with the lesions of fragile retinal blood vessels caused by diabetes complications [7, 79].

Recently, deep learning techniques have been applied to detecting retinal vascular diseases. The most notable success is detection of DR [33]. One of the key reasons behind this success is the vast amount of the data set, in [33], more than 128 thousand images are used to develop a deep convolutional neural network (CNN). Compared to other retinal vascular diseases, DR has a high awareness because it is one of the leading causes of legal blindness for working-age adults. As a result, the number of accumulated retinal images is large, and there are many experienced doctors who can label the images.

In general, developing high-performance deep learning algorithms requires a large number of samples. This is because neural networks have many parameters, even small networks have more than one million parameters and the state-of-art models have more than twenty million parameters. Furthermore, the development of a deep learning algorithm in medical image analysis requires collections of large data sets with tens of thousands of abnormal (positive) cases. As the prevalence of a disease is usually low, this adds a significant challenge.

Unfortunately, data sets for most retinal vascular diseases are limited (often less than several thousand), and generally imbalanced between negative and positive images. This is because labeling medical images is very costly and time-consuming compared to labeling natural images. Labeling natural images are often crowdsourced and ordinary people label the images [78]. However, labeling medical images cannot be crowdsourced because it requires training under experienced clinicians. The unavailability of a reasonable amount of data is one of the main obstacles that prohibit the replication of similar advances for detecting other retinal vascular diseases.

The problem is whether it is possible, and if so, how to build upon the techniques and knowledge of DR detection for other retinal vascular diseases, given their limited amount of data?

Our work

In this work, we propose and apply a transfer learning technique for retinal vascular disease detection. The basic idea of transfer learning is to identify a well-studied source task that shares some similar features with the target task for which there is limited data. Here we choose the well-studied DR detection as a source task, and transfer the learned knowledge to the early detection of retinopathy of prematurity (ROP), as the target task.

The transfer learning approach proposed here is different from the traditional transfer learning approach widely adopted for medical image analysis. The former focuses on feature similarities between the source task and the target task, while the latter uses a large natural and generic image data set such as ImageNet [75] for pretraining, with the belief that transfer learning from a large image data set helps improve the model performance. Clearly, due to the

large difference in image features, effectiveness and robustness of this ImageNet-pretrained transfer learning vary and depend on the size of the pretraining data set and the size of the architecture [66, 62].

To validate our DR-pretraining transfer learning approach, we compare its performance with the ImageNet-pretrained transfer learning approach, against the baseline results from the direct training (training from random initialization) approach. To investigate the robustness of our approach, we conduct a series of experiments with reduced training samples in the target task as well.

Our experimental results show the superior performance of the DR-pretrained approach, not only in all metrics of AUROC, accuracy, precision, and sensitivity but also in robustness. The robustness is with respect to both the stochasticity in the training process and reduction in training samples.

Our studies suggest the effectiveness of our proposed transfer learning approach and its potential for a broad range of retinal vascular diseases or pathologies, where data is limited.

Why Retinopathy of Prematurity?

There are several reasons why ROP is chosen as the target task.

As we reviewed in the previous chapter, ROP has the following features. Firstly, ROP is a common retinal vascular disease. It is an abnormal blood vessel development in the retina of prematurely-born infants or infants with low birth weight [22]. ROP can lead to permanent visual impairment and is one of the leading causes of infant blindness globally. It is estimated that nineteen million children are visually impaired worldwide [8], among which ROP accounts for six to eighteen percent of childhood blindness [29]. Early treatment has confirmed the efficacy of treatment for ROP [21]. Therefore, it is crucial that at-risk infants receive timely retinal examinations for early detection of potential ROP.

Secondly, early detection of ROP is particularly challenging, due to infants' inability of active participation in medical diagnosis. To minimize the number of missed diagnoses for ROP in infants, clinical screening for ROP requires exceptionally high discriminatory power.

In light of these, ROP presents itself as an ideal testbed for the feasibility of the transfer learning technique utilizing feature similarities for detecting retinal vascular diseases. And the success of transfer learning from DR to ROP, especially in comparison with existing approaches, is a barometer for the potential of this transfer learning technique.

Related Work

As mentioned earlier, the standard approach of transfer learning in medical image analysis is to use the ImageNet data set as a pretraining data set. However, the structure of the images is very different in the natural images and medical images. For example, in a natural image, most of the time, the target object located in the center is the most important. On the other hand, in a medical image, the basic structures are shared (e.g. bone structure, the number and the relative location of the organs) among patients and what is important is

Figure 4.1: Examples from ImageNet data set; unlike medical images, the main object is located in the center and images do not have a common structure

the small details. Hence, understanding the effectiveness of transfer learning from a natural image data set to a medical image data set is not straightforward. Given that fact, the following studies investigated the effect of transfer learning from natural image data sets including ImageNet to medical image data sets in various ways. In [66], the effect of transfer learning from the standard ImageNet data set to medical image data sets is investigated in terms of the architecture size, learned features, and feature reusing. The detection of DR and the detection of 5 diseases (Atelectasis, Cardiomagaly, Consolidation, Edema, and Pleural Effusion) in chest X-ray. The DR data set is the one used in [33], and the training set consists of 250,000 images and the test set consists of 70,000 images. The chest X-ray data

set is called Chexpert, and it is collected for a public challenge [47]. The data set consists of 224,000 images. For the architecture, InceptionV3 [86], ResNet50 [37], and a family of small architectures. In terms of the performance, their experiment shows that ImagaNet-pretrained transfer learning offers little benefit to the performance of state-of-art models (ResNet50 and InceptionV3) and almost no improvements to the small architectures both compared to the training from random initialization. Furthermore, a post-analysis showed that the feature reusing is concentrated in the lowest layers, which means only the basic image features are directly transferred. Another benefit confirmed in the experiments is the shortened training time.

Based on [66], [62] conducted similar and extended experiments at a larger scale in terms of the pretraining data set size and the architecture size. Three natural image data sets are used as the source data set. The first data set is the standard ImageNet data set, which again consists of more than 1.2 million images in 1,000 classes. The second source data set is an expanded version of the standard ImageNet data set. It consists of 14 million images in more than 21 thousand classes. The last and the largest source data set is called JFT-300M. The data consists of 300 million images, and each image is labeled with one or more of 18 thousand classes. The source data sets are mammography, chest X-ray, and dermatology. The architectures used in the study are based on ResNet50 and ResNet101, but the hidden dimension is expanded up to 3 times bigger than the base architectures. The number of parameters ranges from 24 million to 380 million. Their results show that using larger architecture and a larger pretraining data set are the keys to benefit by transferring from natural image data sets. As well as the improved in-domain performance, the results show improvements in robust generalizability to domain shift and data efficiency, while not hurting the subgroup fairness and model calibration.

Very limited studies have been done beyond ImageNet-pretrained transfer learning except for the following studies. [14] aggregated the data set from several public challenges with diverse imaging modalities, target organs, and diseases to develop a 3D lung segmentation model. The authors designed a neural network to make a series of pretrained models to extract common 3D features in medical images.

[41] utilized the in-domain transfer learning approach to develop a liver lesion segmentation and classification problem. Their results showed superior performance of in-domain pretraining than ImageNet-pretraining. [5] also showed the improved performance of in-domain transfer learning over pretraining on a natural image data set. The authors applied transfer learning between two data sets of histopathology images, from colon cancer to breast cancer. Another work in this direction is [6], in which transfer learning was performed from a dermatology data set to a diabetic foot ulcer data set to show its higher performance compared to the standard transfer approach of using a natural image data set.

Lately, [4] proposed a transfer learning technique to utilize unlabeled data.

To the best of our knowledge, our work is the first that applies the supervised transfer learning method from one retinal vascular disease to another.

4.2 Methodologies

Transfer Learning

Transfer learning (TL) is a technique in machine learning which aims to transfer the learned knowledge from one domain to another [102]. In transfer learning,

- a domain is a pair of a measurable space and a probability distribution on this space: $\mathcal{D} = (X, P(X))$, where X is called a feature space and $P(X)$ is the distribution of the feature;
- a task in a domain $\mathcal{D}$ is a pair of a measurable space Y and a function from X to Y: $\mathcal{T} = (Y, f)$, where Y is called the label space and f is called a decision function.

The problem that machine learning tries to solve in a domain $\mathcal{D}$ for a task $\mathcal{T}$ is to learn f from the samples drawn from $P(X)$. Transfer learning utilizes the knowledge of a machine learning problem in a source domain $\mathcal{D}_s$ for a task $\mathcal{T}_s$ to improve performance of the learned decision function in a target domain $\mathcal{D}_t$ for a task $\mathcal{T}_t$. The way of transferring the knowledge from the source domain to the target domain depends on the machine learning algorithm.

In transfer learning with neural networks, the knowledge transfer is done by transferring the weight of the models. There are two major ways to achieve the knowledge transfer when a convolutional neural network is used. The first approach is called fine-tuning. Fine-tuning is the most popular approach when transfer learning is applied to deep learning. In this approach, the network is first trained for the source task and then the weight is transferred to the target task. Namely, given the source domain $\mathcal{D}_s = (X_s, P_s(X))$, source task $\mathcal{T}_s = (Y_s, f_s)$, and the network $\hat{f}(\cdot; \theta)$ with the network weight parameter θ, transfer learning training is a bi-level optimization problem. The first step is to solve the following optimization problem

$$\min_{\theta} \mathbb{E}_{x \sim P_s(X)}[loss(f_s(x), \hat{f}(x; \theta))].$$

The training in the source domain is called pretraining. Suppose that an optimal weight θ_s^* is obtained from solving the above optimization problem, then the second step is to solve the following optimization problem

$$\min_{\theta} \mathbb{E}_{x \sim P_t(X)}[loss(f_t(x), \hat{f}(x; \theta))], \ \theta_0 = \theta_s^*,$$

where $\mathcal{D}_t = (X_t, P_t(X))$ is the target domain and $\mathcal{T}_t = (Y_t, f_t)$ is the target task. In other words, the trained network weight in the source task is used as the initial point of the optimization. This approach is very popular in computer vision problems because the image features such as edges or corners are universal across the image domains, and the fine-tuning approach promotes the reuse of these learned features. Note that when the output layer is not compatible with the target domain, the output layer is removed and trained from scratch in the target domain.

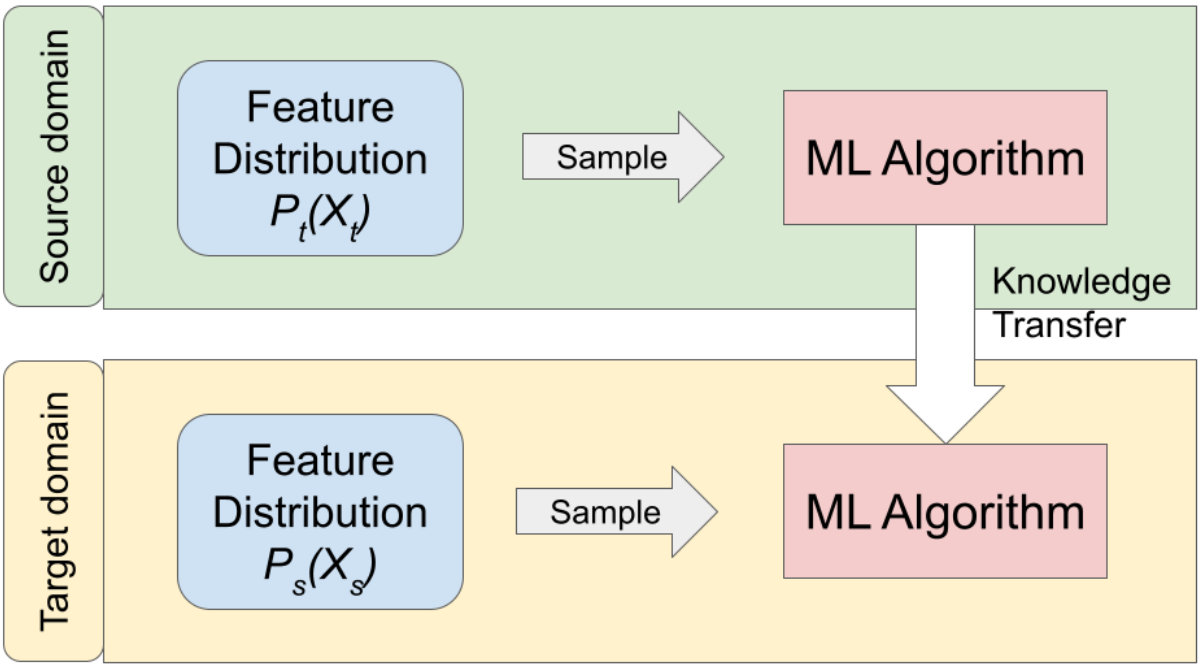


Figure 4.2: Transfer learning; In each domain, samples are drawn from the feature distribution and a machine learning (ML) algorithm learns the relation between the feature and the label; Transfer learning aims to transfer the knowledge learned in the source domain to the target domain

The second way, which is less popular is called freezing. As we have seen in Chapter 1, a convolutional neural network consists of two parts; the image feature extractor consists of convolutional layers and the decoder that interprets the extracted features and makes an output. In the freezing approach, the image feature extractor is frozen, which means the weight will not be trained in the target domain, and the decoder part is retrained or trained from scratch. This approach is used when the high-level image features are common to the source domain and the target domain. In that situation, using the same image feature extractor can be justified and only the decoder part can be tuned to fit into the target domain. The freezing approach is not as effective in medical image analysis, especially when a data set of natural images is used for pretraining.

Target task

The target is to develop a deep neural network that correctly classifies input color fundus photograph as ROP positive or ROP negative (Fig 4.3). In the transfer learning framework,

- The feature space X_t is a space of 3D tensors of a particular size.
- The feature distribution $P(X_t)$ is the distribution of the color fundus photographs taken from infants.
- The label space $Y_t = \{\text{ROP positive, ROP negative}\}$.

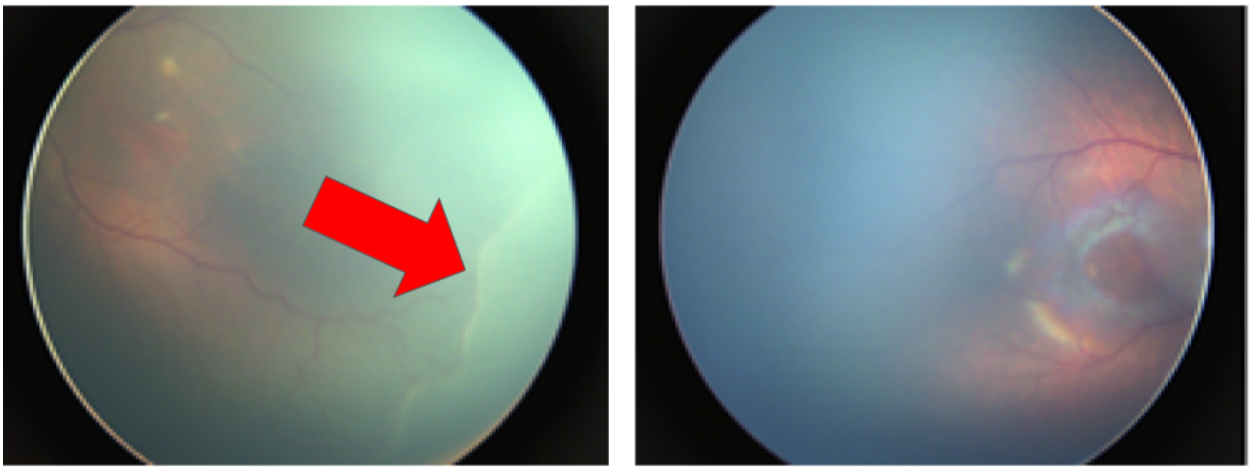


Figure 4.3: ROP positive sample (left) and negative sample (right); the white line pointed by a red arrow in the positive image is a disease feature called thickened ridge

Source tasks

We used two different transfer learning approaches. The first one is the standard approach of using the standard ImageNet data set as a source task. The second one is what we propose: the DR data set. In the following, we refer the first approach and second approach by ImageNet-pretrained approach and DR-pretrained approach, respectively. We introduce the two source tasks in the following.

In the first approach, the source task is to classify an input natural image into one of the 1,00 classes. In the transfer learning approach,

- The feature space X_{s_1} is a space of 3D tensors of a particular size.

- The feature distribution $P(X_{s_1})$ is the distribution of natural images.
- The label space $Y_{s_1} = \{\text{Gold Fish, Sports Car, ...}\}$.

The second source task is the detection of DR in color fundus photograph.

- The feature space X_{s_2} is a space of 3D tensors of a particular size.
- The feature distribution $P(X_{s_2})$ is the distribution of color fundus photographs taken from diabetes patients.
- The label space $Y_{s_2} = \{\text{DR positive, DR negative}\}$.

Here, note that we chose the target task to be a binary classification task although there are 5 stages in DR. This is because the purpose of the pretraining is to learn the basic features, so making the source problem unnecessarily complicated is not desirable.

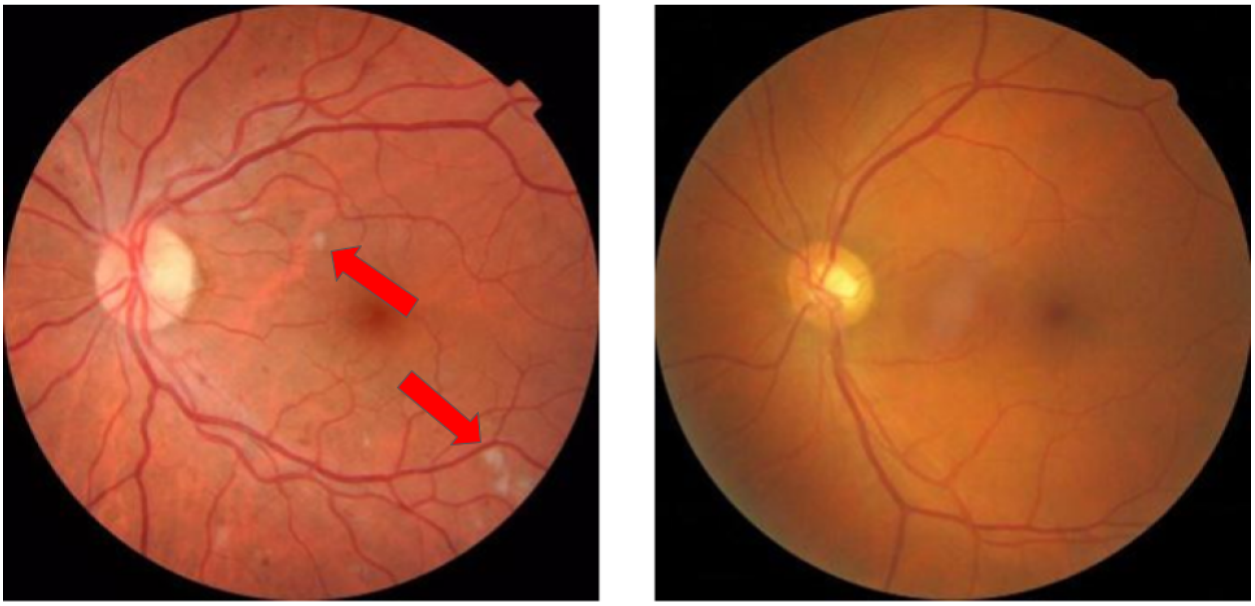


Figure 4.4: DR positive sample (left) and negative sample (right); the pathological region is pointed by red arrows

4.3 Experiment

We investigate the effect of transfer learning by comparing our DR-pretrained approach with the ImageNet-pretrained transfer learning, against the baseline results from the direct training with random initialization. Throughout our experiment, ResNet50 architecture is used.

Data collection

The color fundus photographs taken from infants were collected from the Affiliated Eye Hospital of Nanchang University, which is an AAA (i.e., the highest ranked) hospital in China. All images were de-identified according to patient privacy protection policy, and the ethics review was approved by the ethical committee of Affiliated Eye Hospital of Nanchang University (ID: YLP202103012). The images were graded by 4 experienced ophthalmologists. As a result, the data set consists of 9,727 images (2,310 positive samples, 7,417 negative samples). The data set is randomly split into a training set and a test set by a ratio of 4:1.

The ImageNet data set consists of 1.3 million images collected online [75]. We used the public weight available in Keras [17] because it is how usually it is done and training a large neural network on a large data set is computationally heavy and time-consuming.

The color fundus photographs for DR were collected from the same hospital as the ROP data set, and the images were graded by experienced ophthalmologists whether DR positive or DR negative. The DR data set consists of 36,126 images with 26,548 positive samples and 9,578 negative samples.

Data augmentation

Each image in the ROP data set or the DR data set is applied brightness adjustment and random flipping. Afterwards, each image is resized to 300×300.

Class rebalance

The ROP data set and the DR data set are imbalanced in terms of the ratio between the positive samples and the negative samples. To mitigate the class imbalance issue in the DR and ROP data sets, we use a hybrid class balancing method: First, the class weight in the loss function is set to be $1 : r$ for the negative class to the positive class; Second, when generating a minibatch, images from each class are sampled at the ratio of $r : 1$ so that each class has an equal impact on the training process. The parameter r is treated as one of the hyper parameters to be tuned.

Metrics for performance evaluation

The trained models are evaluated by four metrics: the area under the receiver operator characteristic curve (AUROC), the accuracy, the precision, and the sensitivity. The training of neural networks is stochastic because the mini-batch stochastic gradient algorithm is used. To account for the stochastic nature of the training, each experiment is iterated three times with different random seeds and the metrics are averaged.

Experiment with reduced training samples

Often the time, the sample size in the target domain is in the order of thousand, which is smaller than our ROP data set. To understand the effectiveness and robustness of transfer learning with limited data set, we further train the models with reduced training samples. In this series of experiments, the same pretrained weights are used for each training i.e., pretraining data set is fully utilized, but the training samples in the target task are reduced by factors ranging from 0% to 90% with 10% interval. The test set is kept the same to ensure consistency for comparison.

4.4 Results

The results are shown in Figure 4.5 and tables 4.1, 4.2, and 4.3. We observe three critical advantages of our proposed approach via DR-pretraining over the traditional approach via ImageNet-pretraining.

Improved performance Firstly, DR-pretraining demonstrates superior performance compared with ImageNet-pretraining. In Figure 4.5, the two transfer learning approach clearly dominate the baseline of the direct training approach. Table 4.1 shows their mean percentage improvements from the direct training. DR-pretraining dominates ImageNet-pretraining by all four metrics.

Pretraining	AUROC	Accuracy	Precision	Sensitivity
DR	16.4%	17.9%	53.5%	29.2%
ImageNet	15.7%	16.3%	47.8%	26.9%

Table 4.1: Mean percentage improvement from direct training

Improved robustness to stochasticity in training Secondly, the DR-pretraining is more robust with respect to the stochasticity in the training process. Table 4.2 shows the mean percentage reduction of standard deviation from direct training. DR-pretraining

reduces the standard deviation by at least nearly 50% for all metrics. In contrast, ImageNet-pretraining adds more standard deviation (reduction of -46.6%) in precision and shows almost no improvement (reduction of 2.94%) in accuracy.

Pretraining	AUROC	Accuracy	Precision	Sensitivity
DR	75.4%	64.3%	47.5%	53.7%
ImageNet	57.0%	2.94%	-46.6%	25.1%

Table 4.2: Mean percentage reduction of standard deviation from direct training: the standard deviation calculated over different runs and the mean calculated over different training sizes

Improved robustness to reduced training samples Lastly, DR-pretraining is more robust with respect to the reduction of training sample size. The percentage changes of metrics from 100% training size to 10% training size are shown in Table 4.3.

Pretraining	AUROC	Accuracy	Precision	Sensitivity
DR	3.63%	4.61%	10.9%	8.82%
ImageNet	4.26%	6.82%	15.8%	11.8%

Table 4.3: Percentage changes in all metrics from 100% training size to 10% training size

These observations suggest 1) DR-pretraining dominates the traditional ImageNet-pretraining in all four metrics (AUROC, accuracy, precision, and sensitivity), 2) DR-pretraining is more robust with respect to both the stochasticity in the training process and reduced training samples.

4.5 Conclusion

The deep learning algorithm is data-hungry. To develop a high-quality deep neural network, a large data set is necessary. However, it is difficult to collect a large medical image data set as labeling medical images is very specific, costly, and time-consuming.

As a result, transfer learning from natural image data sets is a popular approach in medical image analysis using deep learning. However, the appearance of natural images and medical images are very different. Hence, this approach is shown to be not too effective or requires hard requirements for the computational environment.

In this study, we propose a transfer learning approach that uses the detection of well-studied retinal vascular disease as a source task to transfer the learned knowledge to the detection of an under-studied retinal vascular disease as a target task. Our experimental

results demonstrate the superior performance of the DR-pretraining approach when compared with the traditional transfer learning and direct training approaches. In addition, our approach showed more robustness to the stochasticity in the training process and the reduction of the training sample size. Our study shows promises of transfer learning techniques utilizing feature similarities for general studies of retinal vascular diseases or other pathologies from different medical fields, where a shortage of data is the main bottleneck for developing efficient deep-learning algorithms for medical image analysis.

Appendix

Experiment with different resolutions

In the main experiments, all the color fundus photographs are resized to 300×300 as explained in the previous section. However, generally, there is a trade-off between the input dimension and the information loss. Using a high-resolution image as the input increases the input dimension and hence more computational burden, but it decreases the information lost in the process of resizing. Usually, the input dimension is decided based on the sample size and memory constraint. In medical image analysis, often the time, the raw medical image is resized to a few hundred by a few hundred with consideration of the small sample size.

Here, to understand the effect of resolution in the transfer learning setting, we trained the models with different resolutions of 200×200, 300×300 (the resolution of the main experiment), and 400×400. The experiment was done using the full ROP training set.

The results of the experiment with different resolutions are shown in Figure 4.6. We see that the curves are concave for the DR-pretraining approach and the ImageNet-pretraining approach, but the curves are convex for the direct training approach. This suggests that there is a trade-off between the input dimension and the model performance for the transfer learning approaches. However, whether this applies to the direct training approach is not clear given the results.

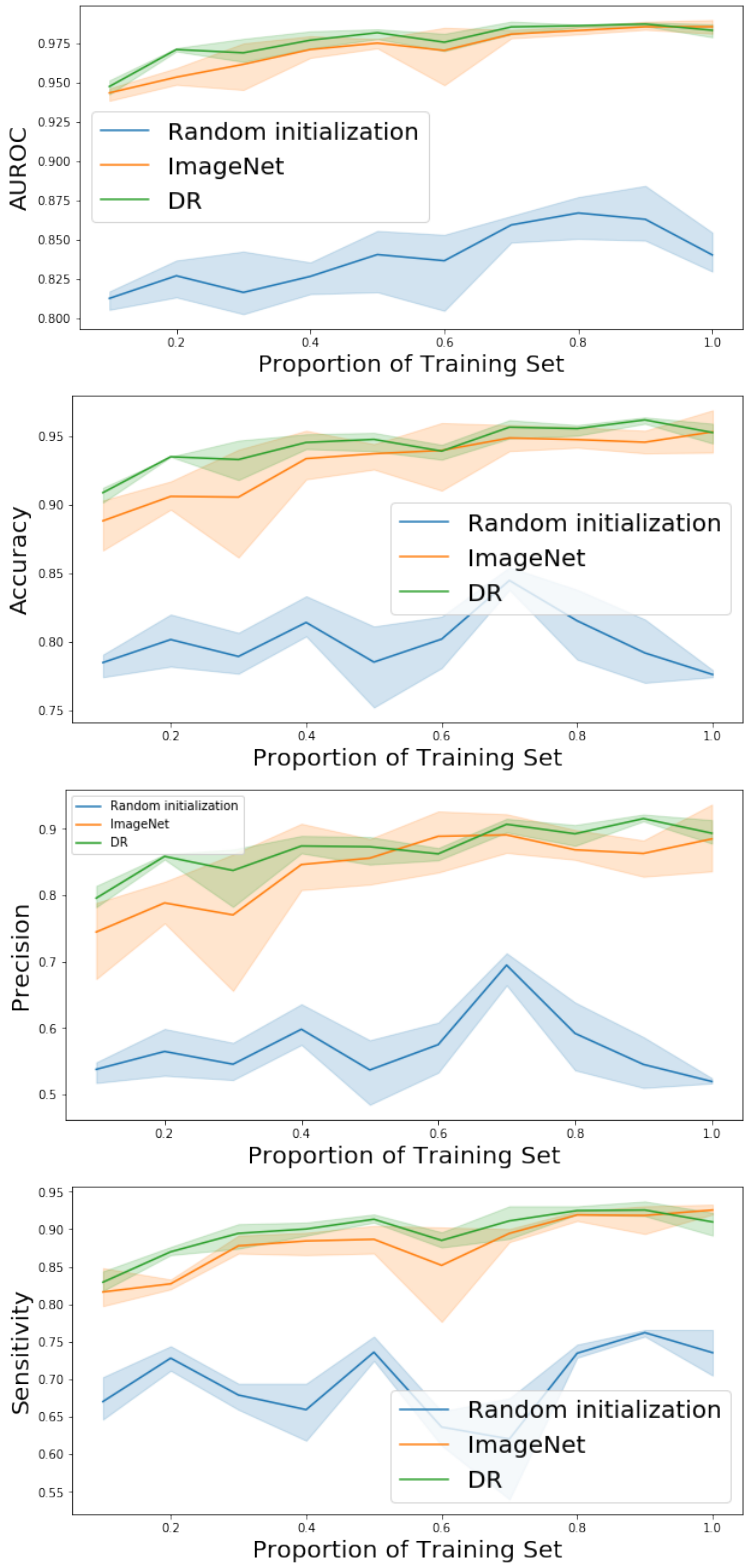


Figure 4.5: Changes in four metrics over training sample reduction, with the dark curves averaged over 3 experiments and the area around the curves showing the minimum and the maximum values in 3 experiments

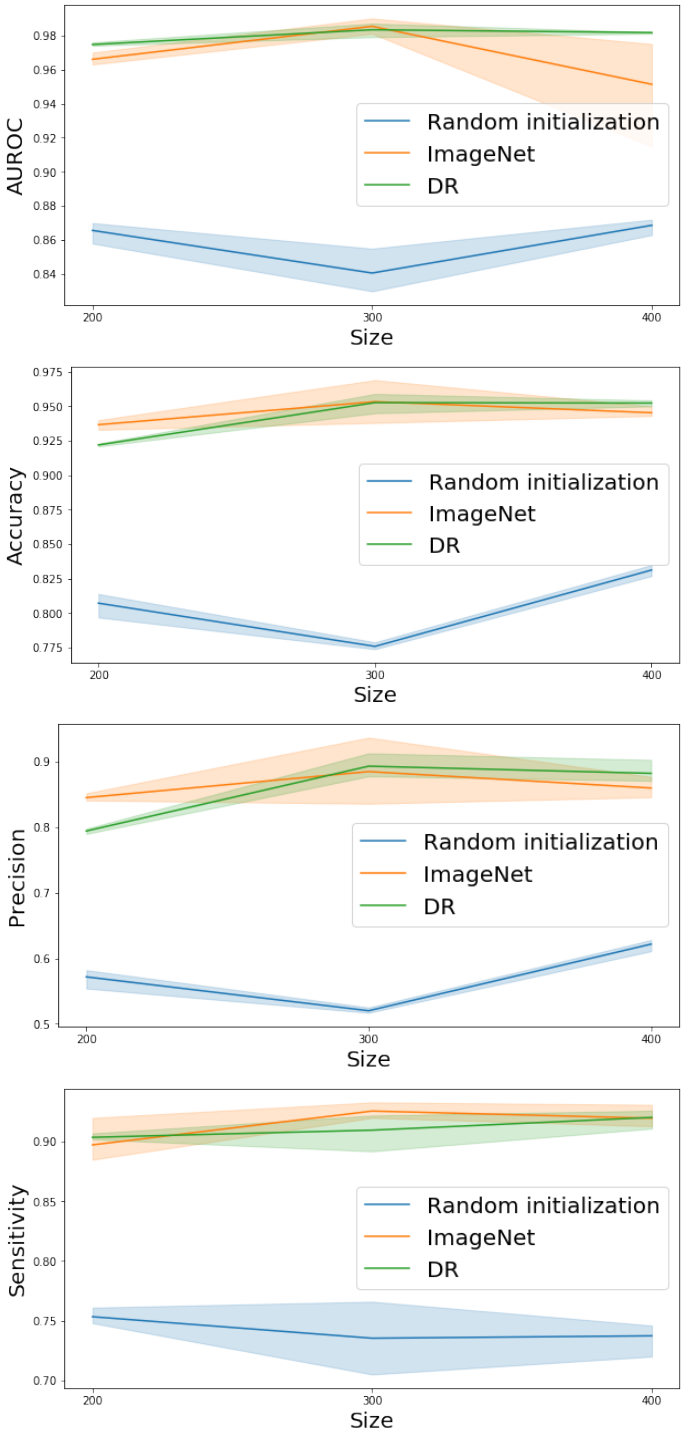


Figure 4.6: Changes in four metrics over different resolution, with the dark curves averaged over 3 experiments and the area around the curves showing the minimum and the maximum values in 3 experiments; the vertical axis is the size of one side of the input image; for example, the size of 300 means the input image size is 300×300

Chapter 5

Predicting Treatment Outcomes in Patients with Neovascular Age-Related Macular Degeneration

In this chapter, a problem of prognostic prediction in neovascular age-related macular degeneration is investigated using machine learning.

5.1 Introduction

Neovascular age-related macular degeneration

Age-related macular degeneration (AMD) is an eye disease that affects the macular. The macular corresponds to the central vision and aging damages the region. The central vision of a patient with AMD gets distorted and blurred when the disease progresses [39]. AMD is a leading cause of legal blindness for people older than 50 years old [24].

Neovascular age-related macular degeneration (nAMD) is an advanced version of AMD. nAMD is also called the 'wet' AMD, and it is characterized by choroidal neovascularization i.e. new blood vessels are developed in the choroid. It is estimated that 90% of severe vision loss due to AMD is because of nAMD.

The vascular endothelial growth factor (VEGF) is known to be a key in the development of nAMD. Therefore, the standard of care involves monthly or bimonthly intraocular injection (i.e. administering into the eyeball) of anti-VEGF agents i.e. inhibitors of VEGF [23]. The anti-VEGF therapy is also known to be effective for diabetic macular edema (DME) [63]. DME is an advanced version of diabetic retinopathy, and edema is created by leaking from fragile new blood vessels, and VEGF involves in the neovascularization here as well [74].

Although the efficacy and safety of anti-VEGF agents for nAMD were confirmed in clinical trials [12, 54], a real-world study suggested that many patients do not achieve the full potential or maintain vision outcomes [44]. This is because of the high treatment burden.

Because most of the patients are elderly, they need help from their families to receive the treatment. As a result, even though the treatment is monthly or bimonthly, many patients skip the injection, and that leads to a lower outcome than expected.

Faricimab is the first bispecific antibody designed for intraocular use that blocks two angiogenesis factors (a factor that involves the development of blood vessels) VEGF and angiopoietin-2 [69]. Faricimab could extend the treatment interval up to 4 months [50]. The efficacy and safety of faricimab for nAMD and diabetic macular edema were tested in phase III clinical trials [70, 71, 72, 73]. In January 2022, faricimab is approved for both diseases by FDA [25].

Personalized medicine

Although faricimab has the potential of extending the treatment interval in general, the treatment response to it is different for a different patient. This is not only specific to faricimab, but the treatment response to a drug generally depends on the patients [30]. The idea of personalized medicine is to use individual patient's profiles to navigate clinical decisions in the disease prevention, disease diagnosis, and treatment of disease, as opposed to the conventional one-dose-for-all approach.

Machine learning and personalized medicine are a very good match. In general, a machine learning model finds out the hidden patterns in the input data to achieve the given task. A successfully trained machine learning model can make sample-level predictions that can help implement the idea of personalized medicine. The effectiveness and the capability of machine learning in personalized medicine are confirmed in early works [56, 88].

Key question

Faricimab could extend the treatment interval. However, whether it is possible to extend the treatment interval, and if it is possible, how long is highly dependent on the patient's profile. The question that we focus on is the following: Can a machine learning model predict treatment outcome with faricimab in nAMD based on the baseline characteristics of patients? This question is very important toward the personalized medicine for nAMD treatment.

Related work

A few studies have been done in the treatment response to the anti-VEGF agents (the standard of care). [80] developed a random forest model to evaluate the potential of machine learning algorithm to predict best-corrected visual acuity (BCVA). In the study, a fully automated segmentation algorithm was used to extract the biomarkers from spectral-domain optical coherence tomography (SD-OCT). In addition to the extracted biomarkers, BCVA is also used as an input feature. The input features from the baseline, month 1, month 2, and month 3 are used to predict the BCVA at month 12. The source of the data is a phase III

clinical trial on ranibizumab (an anti-VEGF agent) for nAMD patients. The results showed an R^2 score of 0.34 when only baseline features are used and R^2 score of 0.70 when the features up to month 3 are used. [9] investigated the problem of predicting low and high anti-VEGF injection requirements using random forest. SD-OCT images and BCVA from the baseline, month 1, and month 2, and demographic profile are used as input to the model. The data source is the pro re nata (PRN), meaning treatment is given as needed, arm in the HARBOR trial, which consists of 317 patients. The patients are divided into a group of low, medium, or high injection requirements. The low injection requirement is defined as less than or equal to 5 injections from month 3 to month 23, and the high injection requirement is defined as more than 16 injections over the same period. The trained model showed AUROC of 0.7 and 0.77 for classifying low injection requirement and high injection requirement, respectively. [67] developed a deep neural network to predict the treatment response to anti-VEGF therapy for diabetic macular edema (DME). In this study, the treatment response is defined by the reduction rate of the total retinal thickness after 3 months of anti-VEGF treatment. The target label is a binary variable of the reduction rate of the total retinal thickness being more than 10% or not. The authors only included SD-OCT as the input for the model. 127 patients are included in the analysis, and the developed model showed an AUROC score of 0.866, a precision score of 85.5%, a sensitivity score of 80.1%, and a specificity score of 85.0%. [49] developed deep neural networks solely on the SD-OCT image to predict BCVA. There are two target variables. The first variable is the concurrent BCVA (i.e. BCVA measured at the same visit as the SD-OCT) and BCVA at month 12. Both regression model and classification model are considered for both of the concurrent BCVA and BCVA at month 12. For the classification problem, the threshold was the Snellen equivalent of $< 20/40$, $< 20/60$, and $\leq 20/200$. The HARBOR trial is used as the data source for the analysis. The results were $R^2 = 0.67$ for concurrent BCVA regression, $R^2 = 0.33$ for month 12 BCVA regression, and AUROC of 0.84 for the best-performing classification model.

To the best of our knowledge, our work is the first study to develop machine learning models to predict treatment response for faricimab in nAMD.

Our work

We developed machine learning models to predict the treatment response to faricimab in nAMD patients using the characteristics from the baseline. The target variable (i.e. the treatment response) is defined from two perspectives. The first target variable is a functional response measured by BCVA. The second variable is an anatomical response measured by the central subfield thickness (CST) reduction from the baseline. For the BCVA variable, a regression problem is considered, and for the CST reduction variable, a binary classification problem is considered.

The data is taken from a phase II clinical trial on faricimab for nAMD patients called AVENUE. 185 patients in AVENUE who are treated with faricimab are included in the study. The input consists of two groups. The first group is symbolic features such as the

baseline BCVA, the baseline CST, demographic features, treatment arm, and so on. The second group is image input. SD-OCT taken at the baseline is included in the input features.

Corresponding to the two groups of features, we used two groups of machine learning models to process each type of input as well. To process the symbolic features, symbolic models such as the linear model and eXtreme Gradient Boosted tree (XGB) are used. To process the image data, deep neural networks are used. Furthermore, to merge the two groups of inputs, we suggested two approaches called model staking and model averaging.

On the holdout test set, the best regression model has an R^2 score of 0.32, and the best classification model has an AUROC score of 0.87. Although merging two groups of input features did not show improvement in the performance, this study highlighted the potential of machine learning to predict the treatment response for faricimab in nAMD patients.

5.2 Methodologies

Data set

AVENUE trial AVENUE (NCT02484690) is a phase II clinical trial for faricimab in nAMD patients [77]. 271 patients enrolled in the study in 58 study sites in the United States. The length of AVENUE is 36 weeks (9 months), and the patients were randomly assigned to one of the five treatment arms including an arm in which the patients are treated fully with the standard of care (ranibizumab) (Figure 5.1). Patients who were treated with faricimab partially or entirely are included in this study. As a result, 185 patients are included in the analysis.

Data split The data set is split into 80% (148 patients) training set and 20% (37 patients) test set. In the split, the target variable is used in stratification.

The training set was split into equal-sized 5 folds, and 5-fold cross-validation was performed. Here again, the target variable is used in stratification when the folds were made. The hyper parameters of the models are tuned using 5-fold cross-validation.

Target variables

In this study, the treatment response is defined with two target variables. The first target variable is the functional response measured by the best-corrected visual acuity (BCVA) letter score at month 9 (i.e. at the end of the clinical trial). In AVENUE, the BCVA is measured using the Early Treatment Diabetic Retinopathy Study (ETDRS) chart. The BCVA letter score is the combination of the total number of letters correctly read at 1 meter or 4 meters. If the number of letters correctly read at 4 meters is more than or equal to 20, the BCVA letter score is that plus 30 letters. Else, it is the total of the number of letters read correctly at 4 meters and that at 1 meter [52].

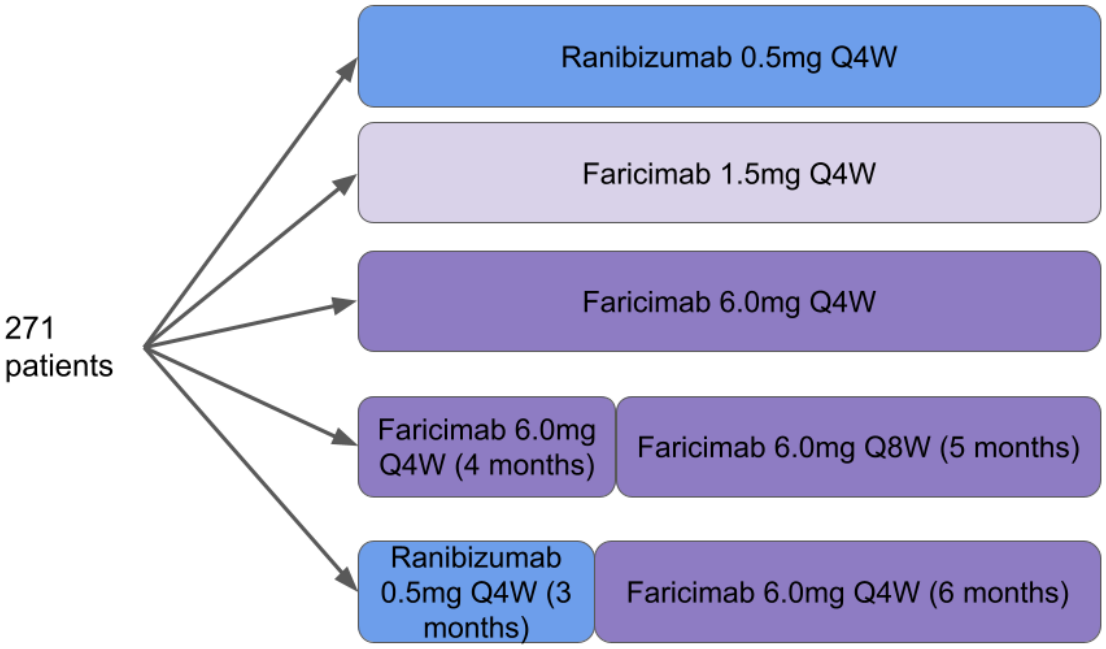


Figure 5.1: Structure of AVENUE trial; QXW means the patients received the treatment every X weeks; Patients in the top treatment arm is exclusively treated with ranibizumab and are excluded from the analysis

The second target variable is the anatomical response measured by the central subfield thickness (CST). The central subfield is a disk with a diameter of 1mm around the central point of the macular. CST is an important anatomical measure to know the status of the disease because the new blood vessels made by nAMD leak and thicken the central subfield. The treatment response can be measured by the reduction rate of CST. In this study, the CST reduction rate variable is transformed into a binary variable with a threshold of 35%. Namely, the variable is 1 if the CST reduction rate is more than 35%, and 0 otherwise. This threshold is chosen to balance the class ratio. In the AVENUE data, the unit for CST is micrometer.

Input features

To predict the treatment response, features at the baseline (i.e. at the beginning of the clinical trial) are included in the model. The inputs can be classified into two groups: symbolic features and image data. A symbolic feature refers to a feature whose number itself has meaning. For example, the height and the weight are symbolic features, but a profile picture is not. This is because the representation of the profile picture on computers

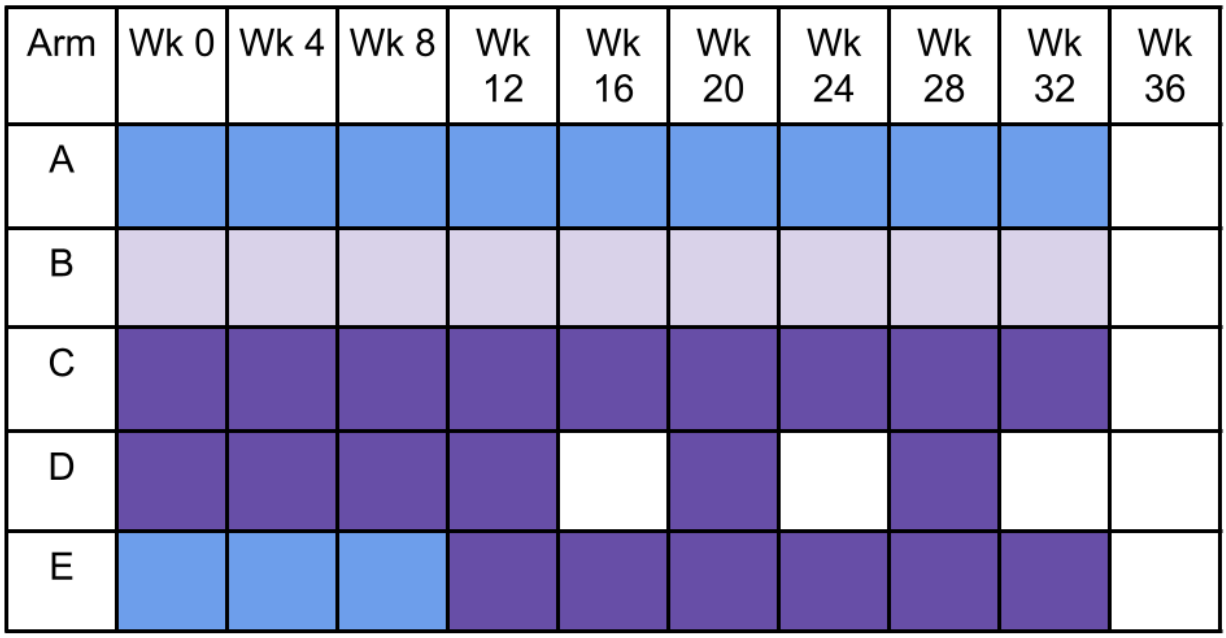


Figure 5.2: Administration schedule of each arm; Empty means sham;Arm A is ranibizumab 0.5 mg Q4W; Arm B is faricimab 1.5 mg Q4W; Arm C is faricimab 6.0 mg Q4W; Arm D is faricimab 6.0 mg Q4W for 4 months and faricimab 6.0 mg Q8W for 5 months; Arm E is ranibizumab 0.5 mg Q4W for 3 months and faricimab 6.0 mg Q4W for 6 months

does not have meaning, and what is meaningful is the structure of the image.

Symbolic features The symbolic features can be further divided into clinical features, demographic features, and the treatment arm. For the clinical features, we included BCVA, CST, and low-luminance deficit (LLD). The LLD is the difference between BCVA and low-luminance visual acuity, which is the number of letters correctly read in a dark room. The LLD is known to have an association with the BCVA gain after the series of treatments [26]. The demographic features include age counted in years and sex. Finally, the treatment arm (faricimab 1.5 mg Q4W, faricimab 6.0 mg Q4W, faricimab 6.0 mg Q4W for 4 months and faricimab 6.0 mg Q8W for 5 months, or ranibizumab 0.5 mg Q4W for 3 months and faricimab 6.0 mg Q4W for 6 months) is one-hot encoded and included in the model to capture the effect of the treatment regimen.

Image data The spectral-domain optical coherent tomography (SD-OCT) taken at the baseline is an image input for the models. OCT is a 3D noninvasive cross-sectional imaging technique in biomedical systems [46]. In SD-OCT, the image is reconstructed by analyzing

the strength and the delay of back-scattered light in spectral domain [98]. In AVENUE, SD-OCT is used to capture the cross-sectional view of the macular. SD-OCT is used in the diagnosis and the monitoring of AMD [59]. The scans were taken using Spectralis (Heidelberg Engineering, Inc. Heidelberg, Germany).

Models

We propose three approaches to model the relationship between the input features and the target variable.

Benchmark models Each benchmark model uses either only the symbolic features or image data. The benchmark models with the symbolic features are the linear model and eXtreme Gradient Boosted trees (XGBoost) [15]. Each model has the following characteristics.

- **linear model**: In a regression problem, a linear model captures a linear relationship between the target variable and the input features.

  Ordinary least square (OLS) method solves the following optimization problem to obtain the coefficients.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} - \beta_0 \right)^2$$

  Here, n is the number of samples, p is the number of input features, and y_i and $x_{i,j}$ are the target variable and the jth input feature of ith observation. Without regularization, the OLS is prone to overfitting. To mitigate that problem, the following regularization techniques are usually used. LASSO [90] uses L^1 norm:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} - \beta_0 \right)^2 + \lambda \sum_{i=j}^{p} |\beta_j|.$$

  With L^1-regularization, the coefficient of unimportant feature is typically set to 0. Ridge regression [43] uses the L^2 norm:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} - \beta_0 \right)^2 + \lambda \sum_{i=j}^{p} |\beta_j|^2.$$

  The Ridge regression shrinks the absolute value of the coefficients. Elastic Net [103] is a mixture of them

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} - \beta_0 \right)^2 + \lambda_1 \sum_{i=j}^{p} |\beta_j| + \lambda_2 \sum_{i=j}^{p} |\beta_j|^2.$$

The Elastic Net inherits the properties of LASSO and Ridge regression.

For the binary classification problem, the linear model captures the linear relationship between the logit and the input features:

$$\hat{y} = \sigma\left(\sum_{j=1}^{p} \beta_j x_{i,j} + \beta_0\right),$$

where $\hat{y}$ is the predicted probability and σ is the Sigmoid function. Also, the loss function is replaced with the binary cross-entropy loss. For example, the Elastic Net for a binary classification problem finds the coefficients by solving

$$\min_{\beta} \sum_{i=1}^{n} -\left(y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right) + \lambda_1 \sum_{i=j}^{p} |\beta_j| + \lambda_2 \sum_{i=j}^{p} |\beta_j|^2.$$

- **XGBoost**: XGBoost is a tree-based model. The core part of XGBoost is the gradient boosting machine [27]. Gradient boosting is an ensemble technique that uses weak models like stumps (single-level decision trees) or small decision trees to make a strong "committee". Gradient boosting method seeks an additive model

$$\hat{y} = \sum_{m=1}^{M} \eta h_m(x),$$

where M is the number of weak models and h_m is the ith weak model. η is called the learning rate. It controls the amount of contribution from each model. In gradient boosting, each weak model is added upon the previous models. The details are the following. Suppose l is the loss function we are trying to minimize over the training samples $\{(x_i, y_i)\}_{i=1}^{n}$. The first weak model is a constant model that solves the following.

$$h_1(x) = \arg\min_{\rho} \frac{1}{n} \sum_{i=1}^{n} l(y_i, \rho).$$

Now, suppose t iterations are done. Let

$$\hat{y}_i^{(t)} = \sum_{m=1}^{t} h_m(x_i).$$

Then, an ideal new weak model satisfies

$$h_{t+1} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)} + h(x_i)\right),$$

where $\mathcal{H}$ is the set of weak models like stumps or decision trees. Since the weak models are usually trees, the use of the gradient descent algorithm is not an option. Thus, the optimization problem above needs to be converted into an optimization problem of suitable form for trees. In gradient boosting used in XGBoost, the first-order and second-order derivatives are used to approximate the objective function locally. Namely,

$$l\left(y_i, \hat{y}_i^{(t)} + h(x_i)\right) \approx l\left(y_i, \hat{y}_i^{(t)}\right) + g_i h(x_i) + \frac{1}{2} H_i h(x_i)^2,$$

where

$$g_i = \frac{\partial}{\partial z} l(y_i, z)\Big|_{z=\hat{y}_i^{(t)}}, \; H_i = \frac{\partial^2}{\partial z^2} l(y_i, z)\Big|_{z=\hat{y}_i^{(t)}}.$$

With this approximation, h_{t+1} is defined by

$$h_{t+1} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(g_i h(x_i) + \frac{1}{2} H_i h(x_i)^2 \right).$$

Or, equivalently, h_{t+1} is a solution to the regression problem with data of $\left\{ \left(x_i, -\frac{g_i}{H_i} \right) \right\}_{i=1}^{n}$. After all, the problem is reduced to a standard regression problem that can be solved using standard decision tree learning algorithms.

XGBoost has additional features on gradient boosting. Examples of the additional features are various ways of regularization, parallel learning, sparsity awareness, and optimization on the engineering side.

The benchmark models only use either one of symbolic features or image data. However, combining those two could increase the predictive power. To merge the two types of inputs, we propose two approaches.

Model stacking The model stacking approach is one of the two approaches to merge the symbolic features and the image data. The model stacking approach has two stages. In the first stage, the deep neural network makes a prediction using the image data. In the second stage, the prediction from the deep neural network is used as one of the input features of the symbolic model in addition to the symbolic features (Figure 5.3). The symbolic model is either the linear model or XGBoost.

Model averaging The model averaging approach is the other approach to merge the symbolic features and the image data. In this approach, the deep neural networks and the symbolic models are trained separately, and the predictions from the deep neural network and the symbolic model are averaged to generate a final prediction (Figure 5.4). The symbolic model is again either the linear model or XGBoost.

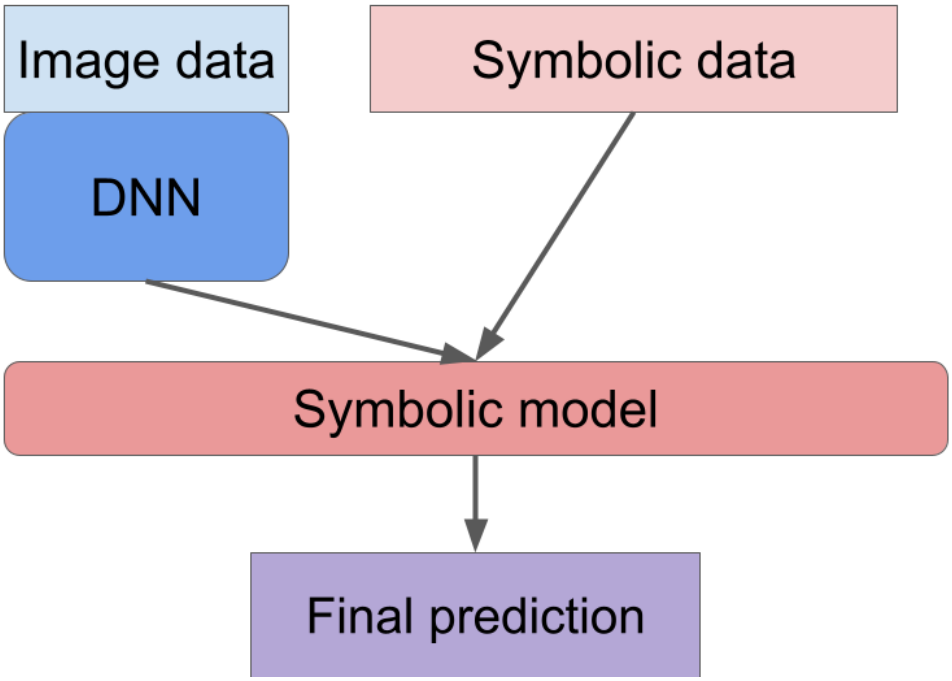


Figure 5.3: Model stacking; DNN, deep neural network

General stacking The model stacking approach and the model averaging approach can be understood in a more general framework. The general stacking approach is an ensemble approach in machine learning [76], and it consists of two stages [95] (Figure 5.5); The first stage is to make predictions using different models; The second stage is to combine the predictions in the first stage using a meta model. The basic idea of stacking is i) to capture different aspects of the relationship between the target variable and the input features by using models with different mechanisms like linear model, tree-based model, and neural network model, and ii) to combine the predictions in the next stage. The general stacking is a popular approach in practice and in data science competitions for its higher performance level than single models.

Our model stacking approach may look different from the general stacking approach since only the image data is considered in the first stage. However, we can interpret our approach in the scope of the general stacking approach in the following way. The first stage can be regarded as consisting of two models. The first model is the projection of image data from the whole input and the deep neural network. The second model is just the projection of the symbolic features from the whole input. Note that we need the projection because each model in the first stage in the general stacking approach takes the whole input. Also, note that a projection can be seen as a machine learning model without any parameters. Then, the symbolic model can be regarded as the meta model in the general stacking approach.

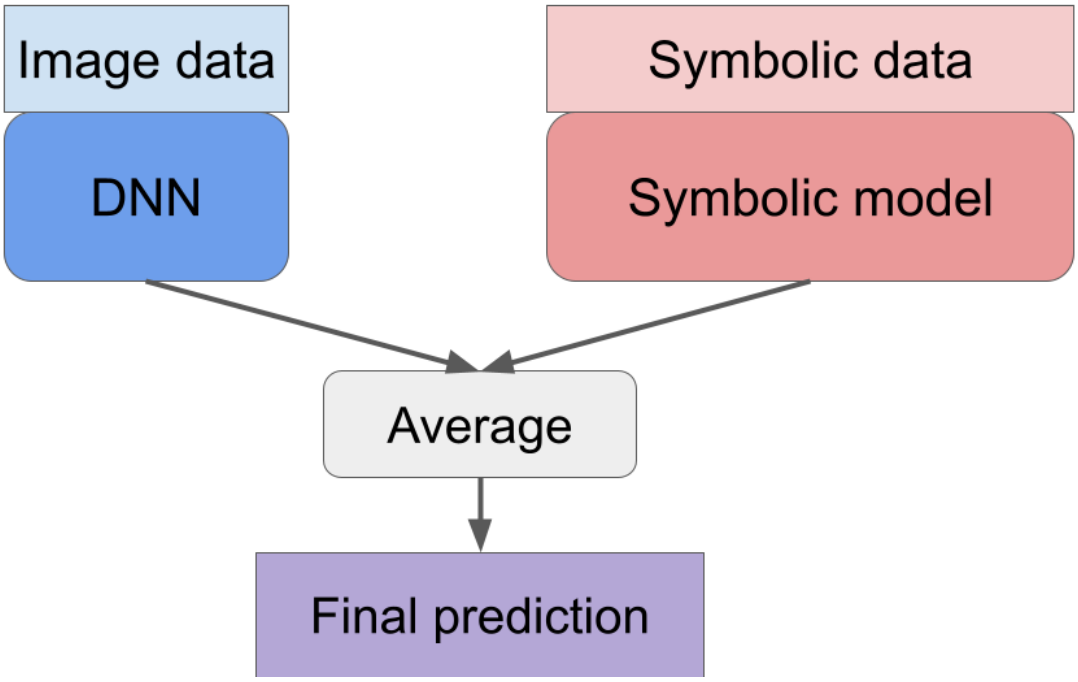


Figure 5.4: Model averaging; DNN, deep neural network

This is also explained in Figure 5.6.

Our model averaging approach can be also seen as an instance of the general stacking approach Figure 5.7. Here, the first stage also consists of two models. The first model is the projection of image data from the whole input data and the deep neural network. The second model is the projection of the symbolic features from the whole input and the symbolic model. The meta model is the simple average. The simple average can be also seen as a machine learning model without any trainable parameters.

Metrics The coefficient of determination (defined below) is used for the BCVA regression problem, and the area under receiver operator characteristic curve (AUROC) is used for the CST classification problem. The definition of the coefficient of determination is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2},$$

where y_i and $\hat{y}_i$ are the true value and the predicted value of ith sample, respectively, and $\bar{y}$ is the mean of $\{y_i\}_{i=1}^{n}$.

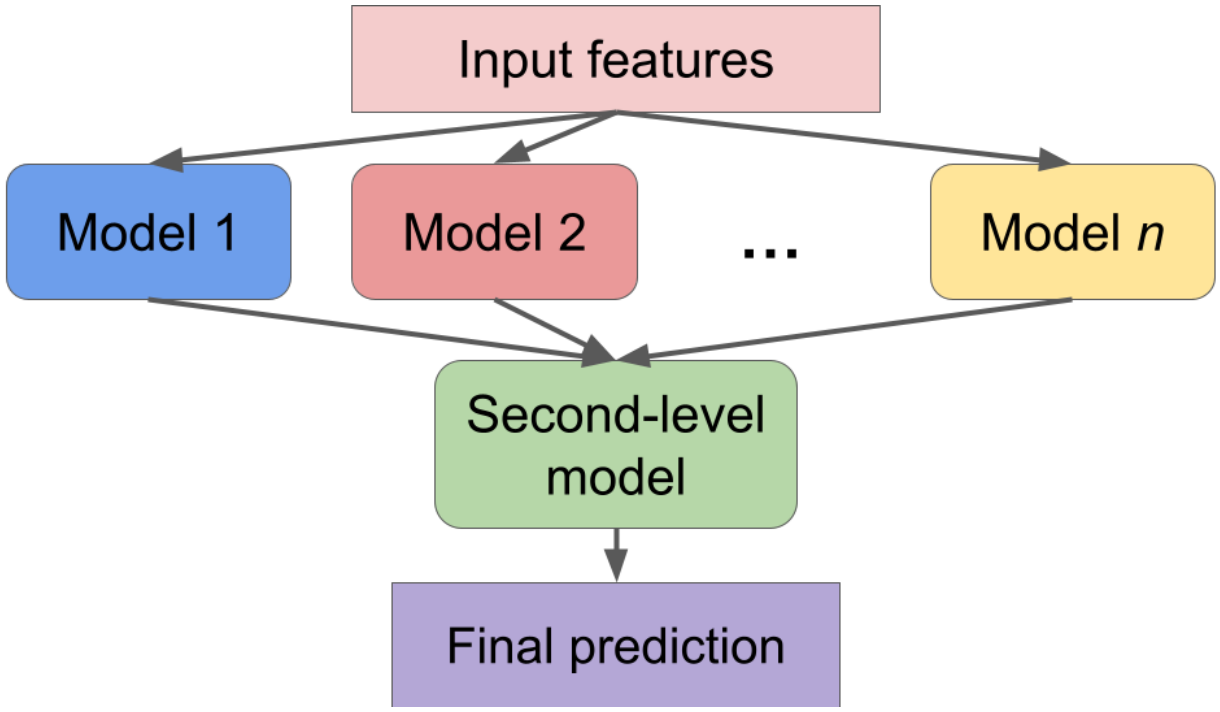


Figure 5.5: General model stacking

5.3 Results and discussion

The results from 5-fold cross validation and test set are shown in Table 5.1 and Table 5.2.

BCVA regression The best performing model on the test set is the linear model with the model stacking approach. It showed the coefficient of determination of 0.32. This is consistent to the results in the previous works [80, 49].

We see no clear improvement after merging the symbolic features and the image data. This would be because of the low performance of the deep neural network; The test result of the deep neural network is a coefficient of determination of 0.079. Since the prediction of the deep neural network is not too helpful, merging it with the symbolic features did not help improve the performance. The low performance of the deep neural network is likely to be due to the small number of samples. Usually, a deep neural network is trained on tens of thousands of samples, but there are only 148 samples for model development in this study. In fact, the big discrepancy between the cross-validation result and the test result of the deep neural network suggests that the model overfitting occurred.

CST classification For the CST classification, the best-performing models are the benchmark linear model and the linear model with model stacking. Both showed an AUROC score

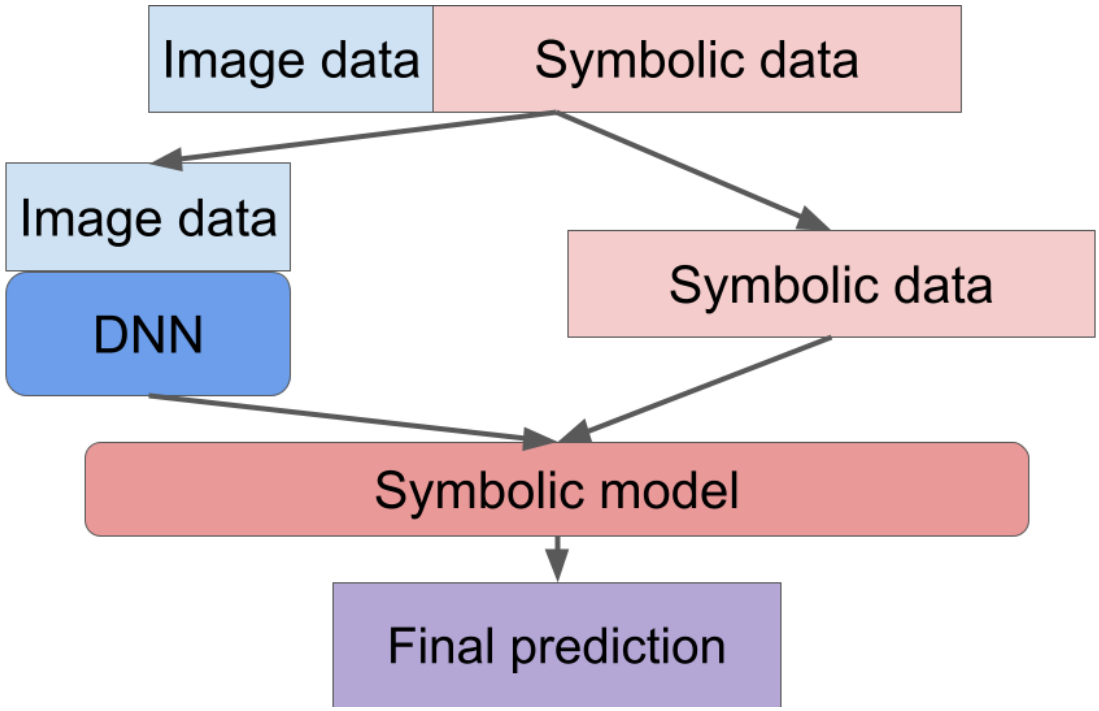


Figure 5.6: Model stacking as a general model stacking; DNN, deep neural network

of 0.87. This is considered as a high performance. Compared to the linear models, the XGBoost model showed a little lower performance.

Here again, there was no clear improvement after merging the symbolic features and the image data. The small number of samples would be surely one of the reasons. However, there is a difference from the BCVA regression results; The benchmark deep neural network showed mild predictive power (AUROC of 0.70). This suggests that the symbolic features and the image data explain the same variance. As a result, the performance did not improve even after merging the two.

5.4 Conclusion

We developed machine learning models to predict the treatment response to faricimab in nAMD patients from the baseline characteristics using data from a phase II clinical trial AVENUE. The treatment response is defined by BCVA and CST. The two types of input are considered; The symbolic features and the image data. We developed symbolic models and deep neural networks. The symbolic models like the linear model and the XGBoost are used to process the symbolic features and the deep neural networks are used to process the image data. In addition, two approaches are suggested and tested to merge the symbolic

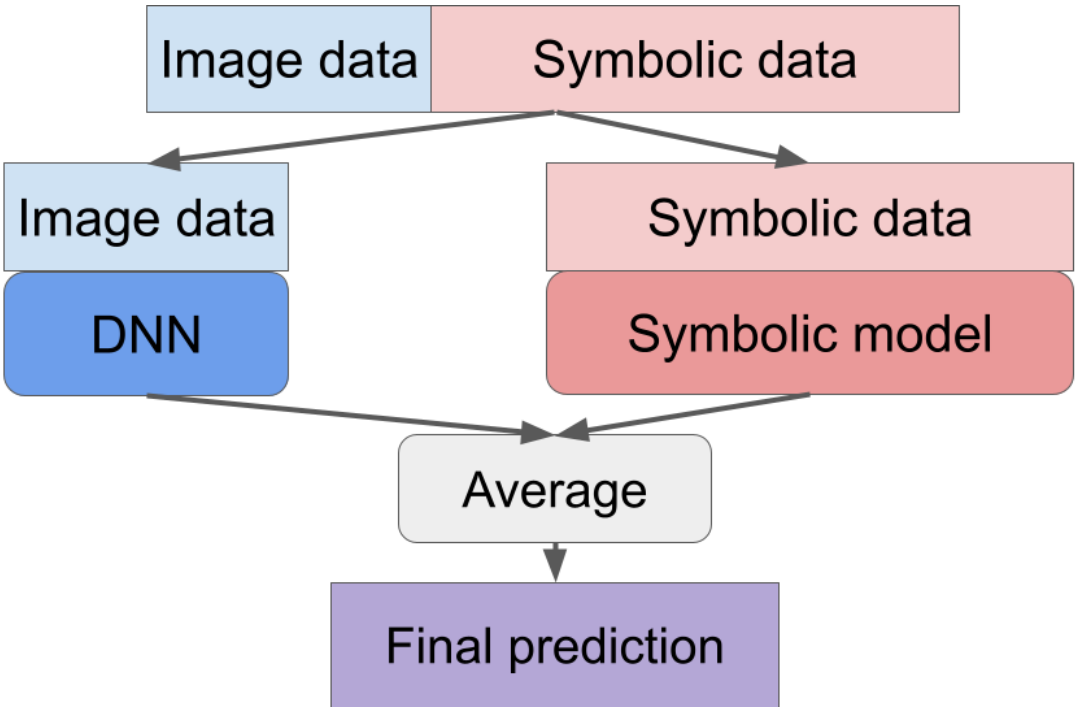


Figure 5.7: Model averaging as a general model stacking; DNN, deep neural network

Model	BCVA regression	CST classification
Benchmark linear model	0.35 (0.16)	0.89 (0.046)
Benchmark XGBoost	0.36 (0.17)	0.90 (0.049)
Benchmark DNN	0.26 (0.078)	0.77 (0.11)
Model stacking linear model	0.42 (0.14)	0.89 (0.046)
Model stacking XGBoost	0.39 (0.14)	0.90 (0.031)
Model averaging linear model	0.38 (0.10)	0.88 (0.070)
Model averaging XGBoost	0.39 (0.11)	0.89 (0.042)

Table 5.1: Mean metrics of 5-fold cross-validation; numbers in parenthesis are standard deviation; DNN, deep neural networks

features and the image data. The model stacking approach has two stages, and the prediction of the deep neural network in the first stage is used as one of the inputs of the symbolic model in the second stage. In the model averaging approach, the deep neural network and the symbolic model are trained separately, and the prediction from each model is simply averaged to make a final prediction.

The result showed mild predictive power for the BCVA prediction problem, and high

Model	BCVA regression	CST classification
Benchmark linear model	0.30	0.87
Benchmark XGBoost	0.29	0.78
Benchmark DNN	0.079	0.70
Model stacking linear model	0.32	0.87
Model stacking XGBoost	0.29	0.76
Model averaging linear model	0.27	0.85
Model averaging XGBoost	0.27	0.80

Table 5.2: Metrics evaluated on the test set; DNN, deep neural networks

predictive power for the CST prediction problem. There was no clear improvement after merging the two types of input with the given sample size. However, this study showed the potential of the suggested approaches to predict the treatment response from the baseline characteristics. To fully explore the predictive capability of the suggested approach, a validation study with more samples, for example, data from a phase III clinical trial or even bigger, is desired.

Bibliography

[1] AAMC. *The Complexities of Physician Supply and Demand: Projections From 2019 to 2034*. June 2021. URL: https://www.aamc.org/media/54681/download?attachment.

[2] Michael David Abràmoff et al. "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning". In: *Investigative Opthalmology & Visual Science* 57.13 (Oct. 2016), p. 5200. DOI: 10.1167/iovs.16-19964. URL: https://doi.org/10.1167/iovs.16-19964.

[3] Saud A. AlAnazi et al. "Effectiveness of in-office blood pressure measurement by eye care practitioners in early detection and management of hypertension". In: *International Journal of Ophthalmology* 8.3 (June 2015), pp. 612–621.

[4] Laith Alzubaidi et al. "Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data". In: *Cancers* 13.7 (Mar. 2021), p. 1590. DOI: 10.3390/cancers13071590. URL: https://doi.org/10.3390/cancers13071590.

[5] Laith Alzubaidi et al. "Optimizing the Performance of Breast Cancer Classification by Employing the Same Domain Transfer Learning from Hybrid Deep Convolutional Neural Network Model". In: *Electronics* 9.3 (2020). ISSN: 2079-9292. DOI: 10.3390/electronics9030445. URL: https://www.mdpi.com/2079-9292/9/3/445.

[6] Laith Alzubaidi et al. "Towards a Better Understanding of Transfer Learning for Medical Imaging: A Case Study". In: *Applied Sciences* 10.13 (June 2020), p. 4523. DOI: 10.3390/app10134523. URL: https://doi.org/10.3390/app10134523.

[7] Diabetes Atlas et al. "International diabetes federation". In: *IDF Diabetes Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation* (2015).

[8] Hannah Blencowe et al. "Preterm-associated visual impairment and estimates of retinopathy of prematurity at regional and global levels for 2010". In: *Pediatric research* 74 Suppl 1 (Dec. 2013), pp. 35–49. DOI: 10.1038/pr.2013.205.

[9] Hrvoje Bogunović et al. "Prediction of Anti-VEGF Treatment Requirements in Neovascular AMD Using a Machine Learning Approach". In: *Investigative Ophthalmology & Visual Science* 58.7 (June 2017), pp. 3240–3248. ISSN: 1552-5783. DOI: 10.1167/iovs.16-21053. eprint: https://arvojournals.org/arvo/content_public/

journal/iovs/936282/i1552-5783-58-7-3240.pdf. URL: https://doi.org/10.1167/iovs.16-21053.

[10] James M. Brown et al. "Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks". In: *JAMA Ophthalmology* 136.7 (July 2018), p. 803. DOI: 10.1001/jamaophthalmol.2018.1934. URL: https://doi.org/10.1001/jamaophthalmol.2018.1934.

[11] Philippe M. Burlina et al. "Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks". In: *JAMA Ophthalmology* 135.11 (Nov. 2017), p. 1170. DOI: 10.1001/jamaophthalmol.2017.3782. URL: https://doi.org/10.1001/jamaophthalmol.2017.3782.

[12] Brandon G. Busbee et al. "Twelve-Month Efficacy and Safety of 0.5 mg or 2.0 mg Ranibizumab in Patients with Subfoveal Neovascular Age-related Macular Degeneration". In: *Ophthalmology* 120.5 (May 2013), pp. 1046–1056. DOI: 10.1016/j.ophtha.2012.10.014. URL: https://doi.org/10.1016/j.ophtha.2012.10.014.

[13] P.W.D. Charles. *Project Title*. https://github.com/charlespwd/project-title. 2013.

[14] Sihong Chen, Kai Ma, and Yefeng Zheng. *Med3D: Transfer Learning for 3D Medical Image Analysis*. 2019. arXiv: 1904.00625 [cs.CV].

[15] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[16] Carol Yim-lui Cheung et al. "Retinal Microvascular Changes and Risk of Stroke". In: *Stroke* 44.9 (Sept. 2013), pp. 2402–2408. DOI: 10.1161/strokeaha.113.001738. URL: https://doi.org/10.1161/strokeaha.113.001738.

[17] François Chollet et al. *Keras*. https://keras.io. 2015.

[18] Yasmine Derradji et al. "Fully-automated atrophy segmentation in dry age-related macular degeneration in optical coherence tomography". In: *Scientific Reports* 11.1 (Nov. 2021). DOI: 10.1038/s41598-021-01227-0. URL: https://doi.org/10.1038/s41598-021-01227-0.

[19] Alexander Ding et al. *Retinopathy of Prematurity Stage Diagnosis Using Object Segmentation and Convolutional Neural Networks*. 2020. eprint: arXiv:2004.01582.

[20] Yanyan Dong et al. "Classification of cataract fundus image based on deep learning". In: *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, Oct. 2017. DOI: 10.1109/ist.2017.8261463. URL: https://doi.org/10.1109/ist.2017.8261463.

[21] Early Treatment For Retinopathy Of Prematurity Cooperative Group. "Revised Indications for the Treatment of Retinopathy of Prematurity: Results of the Early Treatment for Retinopathy of Prematurity Randomized Trial". In: *Archives of Ophthalmology* 121.12 (Dec. 2003), pp. 1684–1694. ISSN: 0003-9950. DOI: 10.1001/archopht.121.12.1684. eprint: https://jamanetwork.com/journals/jamaophthalmology/articlepdf/415949/ecs30202.pdf. URL: https://doi.org/10.1001/archopht.121.12.1684.

[22] Walter M. Fierson. "Screening Examination of Premature Infants for Retinopathy of Prematurity". In: *Pediatrics* 142.6 (2018). ISSN: 0031-4005. DOI: 10.1542/peds.2018-3061. eprint: https://pediatrics.aappublications.org/content/142/6/e20183061.full.pdf. URL: https://pediatrics.aappublications.org/content/142/6/e20183061.

[23] Christina J. Flaxel et al. "Age-Related Macular Degeneration Preferred Practice Pattern®". In: *Ophthalmology* 127.1 (Jan. 2020), P1–P65. DOI: 10.1016/j.ophtha.2019.09.024. URL: https://doi.org/10.1016/j.ophtha.2019.09.024.

[24] Seth R Flaxman et al. "Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis". In: *The Lancet Global Health* 5.12 (Dec. 2017), e1221–e1234. DOI: 10.1016/s2214-109x(17)30393-5. URL: https://doi.org/10.1016/s2214-109x(17)30393-5.

[25] United States Food and Drug Administration. *Highlights of prescribing information, Vabysmo*. Accessed: 02-10-2022.

[26] Ronald E P Frenkel, Howard Shapiro, and Ivaylo Stoilov. "Predicting vision gains with anti-VEGF therapy in neovascular age-related macular degeneration patients by using low-luminance vision". In: *British Journal of Ophthalmology* 100.8 (2016), pp. 1052–1057. ISSN: 0007-1161. DOI: 10.1136/bjophthalmol-2015-307575. eprint: https://bjo.bmj.com/content/100/8/1052.full.pdf. URL: https://bjo.bmj.com/content/100/8/1052.

[27] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[28] Rishab Gargeya and Theodore Leng. "Automated Identification of Diabetic Retinopathy Using Deep Learning". In: *Ophthalmology* 124.7 (July 2017), pp. 962–969. DOI: 10.1016/j.ophtha.2017.02.008. URL: https://doi.org/10.1016/j.ophtha.2017.02.008.

[29] Clare Gilbert et al. "Retinopathy of prematurity in middle-income countries". In: *The Lancet* 350.9070 (1997), pp. 12–14. DOI: 10.1016/s0140-6736(97)01107-0.

[30] Laura H. Goetz and Nicholas J. Schork. "Personalized medicine: motivation, challenges, and progress". In: *Fertility and Sterility* 109.6 (June 2018), pp. 952–963. DOI: 10.1016/j.fertnstert.2018.05.006. URL: https://doi.org/10.1016/j.fertnstert.2018.05.006.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[32] Arun Govindaiah et al. "Deep convolutional neural network based screening and assessment of age-related macular degeneration from fundus images". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, pp. 1525–1528. DOI: 10.1109/ISBI.2018.8363863.

[33] Varun Gulshan et al. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". In: *JAMA* 316.22 (Dec. 2016), pp. 2402–2410. ISSN: 0098-7484. DOI: 10.1001/jama.2016.17216. eprint: https://jamanetwork.com/journals/jama/articlepdf/2588763/joi160132.pdf. URL: https://doi.org/10.1001/jama.2016.17216.

[34] Guo Haixiang et al. "Learning from class-imbalanced data: Review of methods and applications". In: *Expert Systems with Applications* 73 (Dec. 2016). DOI: 10.1016/j.eswa.2016.12.035.

[35] M. Elizabeth Hartnett and Robert H. Lane. "Effects of oxygen on the development and severity of retinopathy of prematurity". In: *Journal of American Association for Pediatric Ophthalmology and Strabismus* 17.3 (June 2013), pp. 229–234. DOI: 10.1016/j.jaapos.2012.12.155. URL: https://doi.org/10.1016/j.jaapos.2012.12.155.

[36] Mohammad Havaei et al. "Brain tumor segmentation with Deep Neural Networks". In: *Medical Image Analysis* 35 (2017), pp. 18–31. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2016.05.004. URL: http://www.sciencedirect.com/science/article/pii/S1361841516300330.

[37] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[38] Kaiming He et al. "Identity Mappings in Deep Residual Networks". In: vol. 9908. Oct. 2016, pp. 630–645. ISBN: 978-3-319-46492-3. DOI: 10.1007/978-3-319-46493-0_38.

[39] National Institute of Health. *Age-Related Maclar Degeneration*. Accessed: 12-28-2021.

[40] National Institute of Health. *Retinopathy of Prematurity*. Accessed: 12-28-2021.

[41] Michal Heker and Hayit Greenspan. "Joint Liver Lesion Segmentation and Classification via Transfer Learning". In: *ArXiv* abs/2004.12352 (2020).

[42] Geoffrey Hinton. *Neural Networks for Machine Learning*. Accessed: 12-24-2021.

[43] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 42.1 (Feb. 2000), pp. 80–86. DOI: 10.1080/00401706.2000.10485983. URL: https://doi.org/10.1080/00401706.2000.10485983.

[44] Frank G Holz et al. "Multi-country real-life experience of anti-vascular endothelial growth factor therapy for wet age-related macular degeneration". In: *British Journal of Ophthalmology* 99.2 (2015), pp. 220–226. ISSN: 0007-1161. DOI: 10.1136/bjophthalmol-2014-305327. eprint: https://bjo.bmj.com/content/99/2/220.full.pdf. URL: https://bjo.bmj.com/content/99/2/220.

[45] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4.2 (1991), pp. 251–257. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(91)90009-T. URL: http://www.sciencedirect.com/science/article/pii/089360809190009T.

[46] David Huang et al. "Optical Coherence Tomography". In: *Science* 254.5035 (1991), pp. 1178–1181. DOI: 10.1126/science.1957169. eprint: https://www.science.org/doi/pdf/10.1126/science.1957169. URL: https://www.science.org/doi/abs/10.1126/science.1957169.

[47] Jeremy Irvin et al. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. 2019. arXiv: 1901.07031 [cs.CV].

[48] Michael G. Kawczynski et al. "Development of Deep Learning Models to Predict Best-Corrected Visual Acuity from Optical Coherence Tomography". In: *Translational Vision Science & Technology* 9.2 (Sept. 2020), pp. 51–51. ISSN: 2164-2591. DOI: 10.1167/tvst.9.2.51. eprint: https://arvojournals.org/arvo/content_public/journal/tvst/938366/i2164-2591-9-2-51_1601619060.40778.pdf. URL: https://doi.org/10.1167/tvst.9.2.51.

[49] Michael G. Kawczynski et al. "Development of Deep Learning Models to Predict Best-Corrected Visual Acuity from Optical Coherence Tomography". In: *Translational Vision Science & Technology* 9.2 (Sept. 2020), pp. 51–51. ISSN: 2164-2591. DOI: 10.1167/tvst.9.2.51. eprint: https://arvojournals.org/arvo/content_public/journal/tvst/938366/i2164-2591-9-2-51_1601619060.40778.pdf. URL: https://doi.org/10.1167/tvst.9.2.51.

[50] Arshad M. Khanani et al. "Efficacy of Every Four Monthly and Quarterly Dosing of Faricimab vs Ranibizumab in Neovascular Age-Related Macular Degeneration". In: *JAMA Ophthalmology* 138.9 (Sept. 2020), p. 964. DOI: 10.1001/jamaophthalmol.2020.2699. URL: https://doi.org/10.1001/jamaophthalmol.2020.2699.

[51] Diederik Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations* (Dec. 2014).

[52] C KNIESTEDT and R STAMPER. "Visual acuity and its measurement". In: *Ophthalmology Clinics of North America* 16.2 (June 2003), pp. 155–170. DOI: 10.1016/s0896-1549(03)00013-0. URL: https://doi.org/10.1016/s0896-1549(03)00013-0.

[53] Jason Kugelman et al. "Automatic choroidal segmentation in OCT images using supervised deep learning methods". In: *Scientific Reports* 9.1 (Sept. 2019). DOI: 10.1038/s41598-019-49816-4. URL: https://doi.org/10.1038/s41598-019-49816-4.

[54] Emily Li et al. "Treatment regimens for administration of anti-vascular endothelial growth factor agents for neovascular age-related macular degeneration". In: *Cochrane Database of Systematic Reviews* (May 2020). DOI: 10.1002/14651858.cd012208.pub2. URL: https://doi.org/10.1002/14651858.cd012208.pub2.

[55] Zhixi Li et al. "Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs". In: *Ophthalmology* 125.8 (Aug. 2018), pp. 1199–1206. DOI: 10.1016/j.ophtha.2018.01.023. URL: https://doi.org/10.1016/j.ophtha.2018.01.023.

[56] Eugene Lin, Chieh-Hsin Lin, and Hsien-Yuan Lane. "Machine learning and deep learning for the pharmacogenomics of antidepressant treatments". en. In: *Clin. Psychopharmacol. Neurosci.* 19.4 (Nov. 2021), pp. 577–588.

[57] Min Lin, Qiang Chen, and Shuicheng Yan. *Network In Network*. 2014. arXiv: 1312.4400 [cs.NE].

[58] Charles X. Ling and Chenghui Li. "Data Mining for Direct Marketing: Problems and Solutions". In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. KDD'98. New York, NY: AAAI Press, 1998, pp. 73–79.

[59] No authors listed. "Optical coherence tomography for age-related macular degeneration and diabetic macular edema: an evidence-based analysis". In: *Ont Health Technol Assess Ser* 9.13 (2009), pp. 1–22.

[60] Martın Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[61] Supriti Mulay et al. "Early detection of retinopathy of prematurity stage using deep learning approach". In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. 2019, pp. 758–764.

[62] Basil Mustafa et al. *Supervised Transfer Learning at Scale for Medical Imaging*. 2021. arXiv: 2101.05913 [cs.CV].

[63] Q. D. Nguyen et al. "Ranibizumab for diabetic macular edema: results from 2 phase III randomized trials: RISE and RIDE". In: *Ophthalmology* 119.4 (Apr. 2012), pp. 789–801.

[64] M. Pekala et al. "Deep learning based retinal OCT segmentation". In: *Computers in Biology and Medicine* 114 (2019), p. 103445. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2019.103445. URL: https://www.sciencedirect.com/science/article/pii/S0010482519303221.

[65] Elena Prokofyeva and Eberhart Zrenner. "Epidemiology of Major Eye Diseases Leading to Blindness in Europe: A Literature Review". In: *Ophthalmic Research* 47.4 (2012), pp. 171–188. DOI: 10.1159/000329603. URL: https://doi.org/10.1159/000329603.

[66] Maithra Raghu et al. "Transfusion: Understanding Transfer Learning for Medical Imaging". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf.

[67] R. Rasti et al. "Deep learning-based single-shot prediction of differential effects of anti-VEGF treatment in patients with diabetic macular edema". In: *Biomed Opt Express* 11.2 (Feb. 2020), pp. 1139–1152.

[68] Travis K Redd et al. "Evaluation of a deep learning image assessment system for detecting severe retinopathy of prematurity". In: *British Journal of Ophthalmology* 103.5 (Nov. 2018), pp. 580–584. DOI: 10.1136/bjophthalmol-2018-313156. URL: https://doi.org/10.1136/bjophthalmol-2018-313156.

[69] Jörg T Regula et al. "Targeting key angiogenic pathways with a bispecific Cross MA b optimized for neovascular eye diseases". In: *EMBO Molecular Medicine* 11.5 (Apr. 2019). DOI: 10.15252/emmm.201910666. URL: https://doi.org/10.15252/emmm.201910666.

[70] Hoffmann-La Roche. *A Study to Evaluate the Efficacy and Safety of Faricimab (RO6867461) in Participants With Diabetic Macular Edema (RHINE)*. Accessed: 01-26-2022.

[71] Hoffmann-La Roche. *A Study to Evaluate the Efficacy and Safety of Faricimab (RO6867461) in Participants With Diabetic Macular Edema (YOSEMITE)*. Accessed: 01-26-2022.

[72] Hoffmann-La Roche. *A Study to Evaluate the Efficacy and Safety of Faricimab in Participants With Neovascular Age-Related Macular Degeneration (LUCERNE)*. Accessed: 01-26-2022.

[73] Hoffmann-La Roche. *A Study to Evaluate the Efficacy and Safety of Faricimab in Participants With Neovascular Age-Related Macular Degeneration (TENAYA)*. Accessed: 01-26-2022.

[74] P. Romero-Aroca et al. "Diabetic Macular Edema Pathophysiology: Vasogenic versus Inflammatory". In: *J Diabetes Res* 2016 (2016), p. 2156273.

[75] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[76] Omer Sagi and Lior Rokach. "Ensemble learning: A survey". In: *WIREs Data Mining and Knowledge Discovery* 8.4 (Feb. 2018). DOI: 10.1002/widm.1249. URL: https://doi.org/10.1002/widm.1249.

[77] Jayashree Sahni et al. "Safety and Efficacy of Different Doses and Regimens of Faricimab vs Ranibizumab in Neovascular Age-Related Macular Degeneration". In: *JAMA Ophthalmology* 138.9 (Sept. 2020), p. 955. DOI: 10.1001/jamaophthalmol.2020.2685. URL: https://doi.org/10.1001/jamaophthalmol.2020.2685.

[78] Supheakmungkol Sarin et al. "Crowdsource by Google: A Platform for Collecting Inclusive and Representative Machine Learning Data". In: Oct. 2019.

[79] Justin Schaneman et al. "The Role of Comprehensive Eye Exams in the Early Detection of Diabetes and Other Chronic Diseases in an Employed Population". In: *Population Health Management* 13.4 (Aug. 2010), pp. 195–199. DOI: 10.1089/pop.2009.0050. URL: https://doi.org/10.1089/pop.2009.0050.

[80] Ursula Schmidt-Erfurth et al. *Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration.* May 2017.

[81] SurajSingh Senjam and Parijat Chandra. "Retinopathy of prematurity: Addressing the emerging burden in developing countries". In: *Journal of Family Medicine and Primary Care* 9.6 (2020), p. 2600. DOI: 10.4103/jfmpc.jfmpc_110_20. URL: https://doi.org/10.4103/jfmpc.jfmpc_110_20.

[82] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv 1409.1556* (Sept. 2014).

[83] Rebecca Smith-Bindman et al. "Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000-2016". In: *JAMA* 322.9 (Sept. 2019), pp. 843–856. DOI: 10.1001/jama.2019.11456. eprint: https://jamanetwork.com/journals/jama/articlepdf/2749213/jama_smithbindman_2019_oi_190085.pdf. URL: https://doi.org/10.1001/jama.2019.11456.

[84] CarlosE Solarte et al. "Plus disease: Why is it important in retinopathy of prematurity?" In: *Middle East African Journal of Ophthalmology* 17.2 (2010), p. 148. DOI: 10.4103/0974-9233.63080. URL: https://doi.org/10.4103/0974-9233.63080.

[85] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[86] Christian Szegedy et al. "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[87] Zachary Tan et al. "Deep Learning Algorithm for Automated Diagnosis of Retinopathy of Prematurity Plus Disease". In: *Translational Vision Science & Technology* 8.6 (Dec. 2019), pp. 23–23. ISSN: 2164-2591. DOI: 10.1167/tvst.8.6.23. eprint: https://arvojournals.org/arvo/content_public/journal/tvst/938258/i2164-2591-8-6-23.pdf. URL: https://doi.org/10.1167/tvst.8.6.23.

[88] Weiyang Tao et al. "Multiomics and Machine Learning Accurately Predict Clinical Response to Adalimumab and Etanercept Therapy in Patients With Rheumatoid Arthritis". In: *Arthritis & Rheumatology* 73.2 (2021), pp. 212–222. DOI: https://doi.org/10.1002/art.41516. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/art.41516. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/art.41516.

[89] Stanford Taylor et al. "Monitoring Disease Progression With a Quantitative Severity Scale for Retinopathy of Prematurity Using Deep Learning". In: *JAMA Ophthalmology* 137.9 (Sept. 2019), p. 1022. DOI: 10.1001/jamaophthalmol.2019.2433. URL: https://doi.org/10.1001/jamaophthalmol.2019.2433.

[90] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[91] Daniel Shu Wei Ting et al. "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes". In: *JAMA* 318.22 (Dec. 2017), p. 2211. DOI: 10.1001/jama.2017.18152. URL: https://doi.org/10.1001/jama.2017.18152.

[92] Yan Tong et al. "Automated identification of retinopathy of prematurity by image-based deep learning". In: *Eye and Vision* 7.1 (Aug. 2020). DOI: 10.1186/s40662-020-00206-2. URL: https://doi.org/10.1186/s40662-020-00206-2.

[93] Avinash V. Varadarajan et al. "Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning". In: *Nature Communications* 11.1 (Jan. 2020). DOI: 10.1038/s41467-019-13922-8. URL: https://doi.org/10.1038/s41467-019-13922-8.

[94] Jianyong Wang et al. "Automated retinopathy of prematurity screening using deep neural networks". In: *EBioMedicine* 35 (Sept. 2018), pp. 361–368. DOI: 10.1016/j.ebiom.2018.08.033. URL: https://doi.org/10.1016/j.ebiom.2018.08.033.

[95] David H. Wolpert. "Stacked generalization". In: *Neural Networks* 5.2 (Jan. 1992), pp. 241–259. DOI: 10.1016/s0893-6080(05)80023-1. URL: https://doi.org/10.1016/s0893-6080(05)80023-1.

[96] Daniel E. Worrall, Clare M. Wilson, and Gabriel J. Brostow. "Automated Retinopathy of Prematurity Case Detection with Convolutional Neural Networks". In: *Deep Learning and Data Labeling for Medical Applications*. Ed. by Gustavo Carneiro et al. Cham: Springer International Publishing, 2016, pp. 68–76. ISBN: 978-3-319-46976-8.

[97] Qi Yan et al. "Deep-learning-based prediction of late age-related macular degeneration progression". In: *Nature Machine Intelligence* 2.2 (Feb. 2020), pp. 141–150. DOI: 10.1038/s42256-020-0154-9. URL: https://doi.org/10.1038/s42256-020-0154-9.

[98] Zahid Yaqoob, Jigang Wu, and Changhuei Yang. "Spectral domain optical coherence tomography: a better OCT imaging strategy". In: *BioTechniques* 39.6S (Dec. 2005), S6–S13. DOI: 10.2144/000112090. URL: https://doi.org/10.2144/000112090.

[99] Veysi M. Yildiz et al. "Plus Disease in Retinopathy of Prematurity: Convolutional Neural Network Performance Using a Combined Neural Network and Feature Extraction Approach". In: *Translational Vision Science & Technology* 9.2 (Feb. 2020), pp. 10–10. ISSN: 2164-2591. DOI: 10.1167/tvst.9.2.10. eprint: https://arvojournals.org/arvo/content_public/journal/tvst/938366/i2164-2591-224-2-1595.pdf. URL: https://doi.org/10.1167/tvst.9.2.10.

[100] Quan Zhang et al. "Identifying Diabetic Macular Edema and Other Retinal Diseases by Optical Coherence Tomography Image and Multiscale Deep Learning". In: *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* Volume 13 (Dec. 2020), pp. 4787–4800. DOI: 10.2147/dmso.s288419. URL: https://doi.org/10.2147/dmso.s288419.

[101] Zhi-Hua Zhou and Xu-Ying Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem". In: *IEEE Transactions on Knowledge and Data Engineering* 18.1 (2006), pp. 63–77.

[102] Fuzhen Zhuang et al. *A Comprehensive Survey on Transfer Learning*. 2020. arXiv: 1911.02685 [cs.LG].

[103] Hui Zou and Trevor Hastie. "Regularization and variable selection via the Elastic Net". In: *Journal of the Royal Statistical Society, Series B* 67 (2005), pp. 301–320.