Few-shot Feature Space Learning for Congenital Retinal Diseases Recognition

by

Siwei Mai

May 18, 2020

A thesis submitted to the
faculty of the Graduate School of
the University at Buffalo, The State University of New York
in partial fulfillment of the requirements for the
degree of

Master of Science

Department of Computer Science and Engineering

I would like to dedicate this thesis to my parents, Xiafen Yao and Chaohui Mai. I want to thank them for being so patient with me and supportive through these years of my studying, and for all the financial support and encouragement throughout these years of study abroad. All great wishes to you.

Acknowledgements

I would like to thank the following people, without whom I would not have been able to complete this research, and without whom I would not have made it through my master's degree!

My supervisor Prof. Mingchen Gao, who gave me the opportunity to commit to this research and whose insight and knowledge into the subject matter steered me through this research.

My dear friend and Ph.D. candidate, Yanjun Zhu, who always gave me a hand when I was in trouble during the research.

My committee member, Prof. Wen Dong, who gave me advices on the data set at the beginning of the experiment and whose passion for research infected me.

Abstract

The goal of this project is to recognize a rare congenital retinal disease, Hereditary Macular Degeneration, based on Optical Coherence Tomography (OCT) images, whose main manifestation is the confusion and adhesion of the layers of the retina. The challenge of using machine learning models to recognize rare diseases comes from the limited number of collected data. To address this problem, we propose to learn a discriminative feature space for the OCT images, on which many classifiers can be applied for various tasks. We formulate this problem as a few-shot learning task as only very limited samples are available. The Siamese training strategy with the triplet loss is employed to maximize the inter-class distance and minimize the intra-class distance. OCT images also have large variations due to different capturing devices and angles. To alleviate such effect, a pipeline of preprocessing is first utilized for image alignment. Tissue images at different angles can be roughly corrected to a horizontal state for better feature representation. There are 58 samples collected in our dataset and half of them are diseased. Extensive experiments on our dataset demonstrate the effectiveness of the proposed approach.

Keywords: Congenital Retinal Diseases Recognition, Feature Learning, Few-shot Learning, Siamese Network

TABLE OF CONTENTS

List of Tables

List of Figures

Introduction

Image classification is a crucial task for many medical applications. Although convolutional neural network, as a state-of-the-art approach, has improved the performance significantly in medical image analysis, limited number of training data is still the obstacle on the way to achieve a satisfying result. This raises a challenge for the case of classifying rare diseases. The goal of this project is to design a computer-aided diagnosis algorithm when only a very limited number of rare disease samples can be collected.

The methods of diagnosing eye diseases like age-related macular degeneration (AMD), diabetic macular edema (DME), etc., through the spectral domain OCT images can be roughly categorized into the traditional machine learning methods and deep learning-based methods.

There are lots of contributions of OCT image analysis based on the traditional machine learning methods like Principal Components Analysis (PCA), Support Vector Machine (SVM), or Random Forest. Liu et al.[1] proposed a series of methods for OCT image preprocessing and applied the multi-scale spatial pyramid to extract dimension-reduced local binary pattern (LBP) histograms as features and classify them by the non-linear SVM in 2011. Anantrasirichai et al.[2] employed the PCA to improve classification performance before the SVM classifier. Srinivasan et al.[3] using the combination of Histogram of Oriented Gradients (HOG) descriptors and SVM to overcome the limitations of Liu et al.[1]. In the pipeline of Sugmk et al.[4], before extraction and classification, the tissues of each layer in the OCT image are detected and segmented in advance to improve the performance of classification. Venhuizen et al.[5] using small descriptive image patches extracted by Interest Point Detection with a Bag-of-Words (BoWs) scheme instead of the original image for feature extraction, which allows for a quick and reliable assessment without the need of accurate layer segmentation algorithms. Sankar et al.[6] proposed a method based

Gaussian Mixture Model (GMM) after PCA dimensionality reduction. Sun et al.[7] obtains global representations of images by using sparse coding and a spatial pyramid. Hussain et al.[8] explored the way by 3D segmented volumes and Random Forest.

Lots of researches also focus on the deep-learning-based methods. Some of them are dedicated to improving existing mature and pre-trained frameworks such as Inception-v3[9–11], VGG16[11–14], PCANet [15], GooLeNet[16,17], ResNet[11,17,18], DenseNet[17] to classify OCT images. Others unify multiple networks together to make classification more robust for diagnosing, for example four-parallel-ResNet system[18] and four-network with two-stage system[19]. Compared with the traditional feature engineering, deep learning has the advantage of learning the hierarchical features.

The goal of this project is to classify a hereditary macular degeneration disease. Because of the rarity of the disease, we have collected 26 diseased OCT images. The limited amount of training data has prevented us from using either a traditional machine learning approach or a deep learning based model.

Consider the aforementioned challenges, we propose a few-shot learning system with Siamese network[20] and triplet loss[21]. To augment the feature learning, we first train a CNN based Siamese network through a much larger auxiliary public OCT dataset. The advantage of triplet loss is to mapping the samples from the same class as close as possible in the feature space, and separates the samples of the different classes as far as possible. The triplet loss function improves the characteristics of the normal OCT samples in the feature space. After completing the Siamese network, we freeze its parameters and conduct the few-shot learning through the classifier designed after embedding CNN, and only let the classifier learn and complete the classification task. Through our research, even if 26 samples are new disease

feature samples, the whole system in Figure 2 can distinguish rare samples in a high-dimensional space.

The Contributions of the thesis is two-fold.

- We propose a few-shot learning classification of Siamese network pipeline.
- The proposed method was validated for a rare retinal disease with OCT images.

Methods

The proposed pipeline consists of three steps. The first preprocessing step is to align images to eliminate variations. This step includes pixel-level adjustments in units or image angle adjustments. The second step is to apply the few-shot feature space learning model on the aligned images. In this model, we propose to use a Siamese network with triplet loss to maximize the margin between images with and without disease and minimize the intra-class distances. Finally, the learned features are used in classifiers to predict whether an image is normal or not.

Image alignment

As shown in Figure 1, the original OCT images are usually captured from diverse angles. Unlike to MRI images that are scanned sequentially along an axis, OCT images are captured along the retina plane without following a particular angle. As the result, OCT images have so many variations. Following a standard series of image manipulations[7], we adjust all pictures to the horizontal position without destroying the original pathological information.

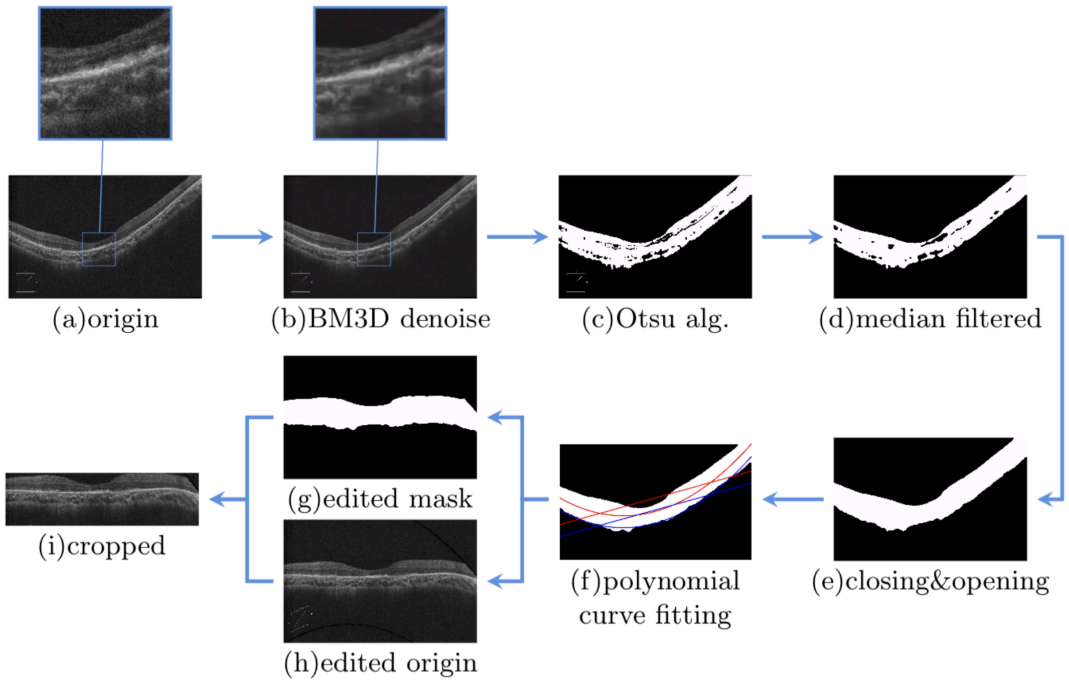


Figure 1 Pipeline for the image alignment

The main idea of image alignment is to generate a mask to attain the region of interest, in this case referring to the retina layered structure. After the mask is obtained, the resulting coefficients of the polynomial fit are used column by column to synchronize the mask and the original image, and finally the pre-processing is done by cropping.

Image pre-processing begins with noise reduction as shown in Figure 1(b) to reduce the irregularly distributed noise. We use Block-matching and 3D filtering (BM3D)[22], the state-of-the-art noise reducing algorithm. The noise reduction pre-processing helps subsequent steps better capture retina structure. After that, the Otsu algorithm in Figure 1(c) is used to allocate the location and morphology of the black background. The Otsu algorithm is independent of the image brightness and contrast. It divides the image pixels into two regions in the background and foreground according to the grayscale values, and such convert the OCT images to binary images. The median filter in Figure 1(d) further reduces the noise area within the tissue.

The morphological operations opening and closing in Figure 1(e) are used to clean noises inside and outside the tissue area so that the binary mask of target tissue is complete and clear. After the contours are obtained, we use polynomial curve fitting to represent the degree of curvature of the tissue area. By the obtained polynomial equation, in particular, the coefficient of the polynomial, we can get the curved state of the retinal tissue. Since a normal retina should be flat or slightly curved downward on an OCT image. A quadratic polynomial is used to fit the upwardly curved abnormal tissue cross section and adjusted to level by column pixels. A primary polynomial was used to fit the normally curved tissue by rotating the angle to horizontal. In the last step, we adjust and crop both the mask and the original image as shown in Figure 1(i).

Few-shot Feature Space Learning

We utilize the Siamese network with a triplet loss function for few-shot feature space learning. The backbone network inside Siamese structure is a convolutional neural network. After training with an auxiliary public dataset, the backbone CNN will be frozen, and two fully connected layers are added to classify rare disease samples.

Siamese network

The main advantage of Siamese network is its strong feature learning capability to be able to extend to classify new categories other than the training labels. Such that, Siamese network is suitable to be applied to some small volume dataset. In our experiments, the Siamese network consists of three ResNets with shared weights. Our goal of Siamese network is to learn a feature y from an input image x. The Siamese network focuses on projecting the feature map into an embedded feature space through minimizing the triplet loss function instead of directly classifying. This procedure determines that the built-in ResNet is to be able learn features of samples from new classes.

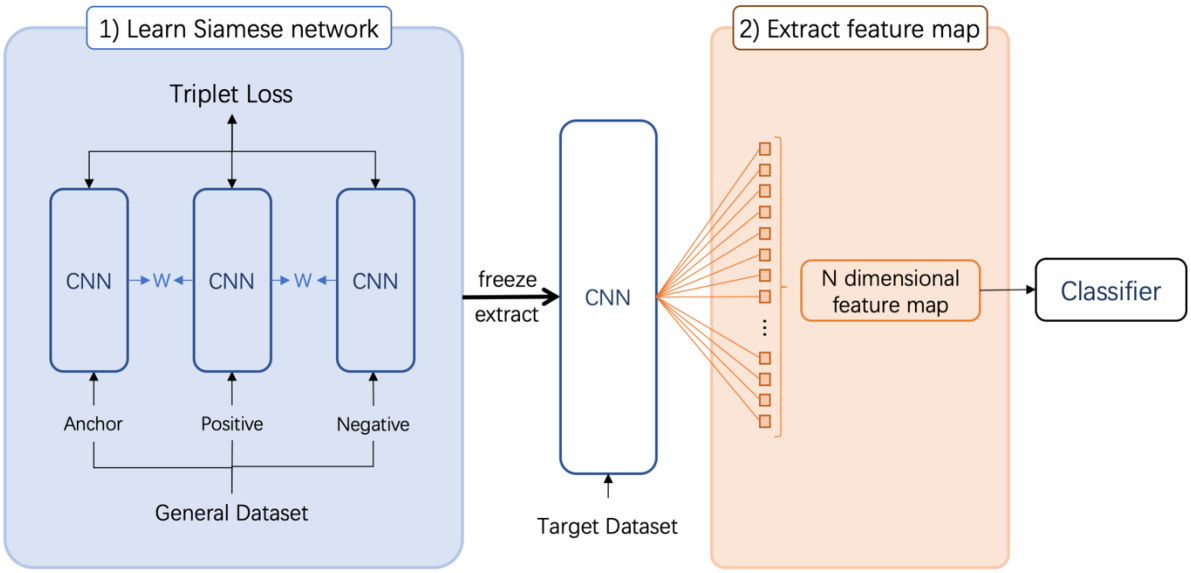


Figure 2 An overview of the proposed method. The Siamese network is trained with the Triplet Loss on an auxiliary dataset. We then use the trained network to extract the feature map on the target dataset for the few-shot classification.

Triplet loss

Triplet loss has the advantage of distinguishing similar inputs by evaluating the three inputs at the same time. Its main purpose is to bring homogeneous samples as close as possible in the embedding space, and to make heterogeneous samples as separated as possible. Compared to the Softmax function usually used for a fixed number of classes, the triplet loss is able to be applied to a variable number of classes.

The input of triplet loss consists of an anchor point, a positive point and a negative point. Let x_a be the anchor sample from a randomly chosen category. Let x_p be the positive sample which means another sample from the same category of x_a and let x_n be the negative sample from the different category besides the category of x_a. The pair of (x_a, x_p, x_n) are randomly chosen from the dataset to ensure the balance of positive and negative samples.

$$\mathcal{L}_{triplet} = max\big(d\big(x_a, x_p\big) - d(x_a, x_n) + margin, 0\big)$$

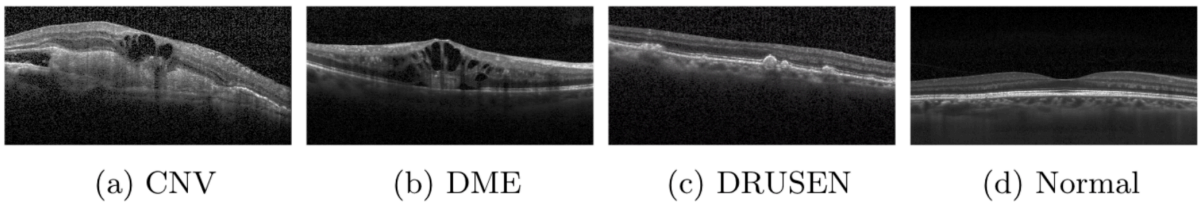
Equation 1 Triplet Loss

The $d(\cdot,\cdot)$ in the Equation 1 denotes the distance between two points in the two or higher dimensional embedding space. Euclidean metric is usually used for the calculation.

Similar to the margin in the Support Vector Machine (SVM), the margin in the Equation 1 plays an important role in ensuring that the points from different categories in the embedding space should be as far away as possible. The larger the value of margin, the greater the distance between different clusters, and the more they are separated.

Experiments

We demonstrate that our feature extraction method can be used to improve the classification performance for scarce medical images. The preprocess pipeline added in the data enhancement process allows the classifier to better cope with different samples obtained by different devices from different angles.



(a) CNV (b) DME (c) DRUSEN (d) Normal

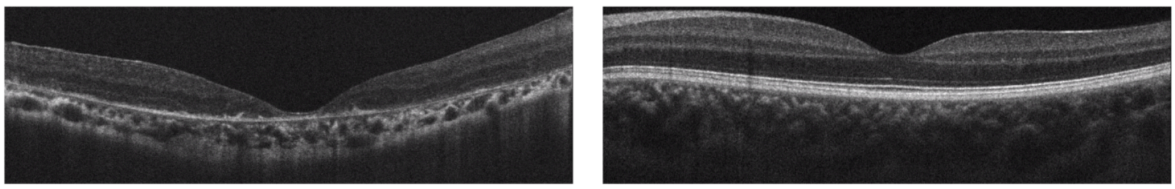
Figure 3 These four figures are the four types of samples in the auxiliary dataset used to train the Siamese network. (a) choroidal neovascularization (CNV) with neovascular membrane (unusually cavernous part) and associated subretinal fluid. (b) Diabetic macular edema (DME) with retinal-thickening-associated intra-retinal fluid (abnormal separation of tissues and fluid content in the layers). (c) Multiple drusen (white irregular bulge) present in early AMD. (d) Normal retina with preserved foveal contour and absence of any retinal fluid/edema.

Auxiliary Dataset

We use a publicly available dataset of OCT images as shown in Figure 3. The data set contains four categories including NORMAL, CNV, DME and DRUSEN. They have a total of 109,309 samples, of which 1,000 are used for testing and the rest are used for training including 51140 for normal and 57169 for abnormal. There are two kinds of sizes in those images, 1536× 496×1 and 1024×496×1. Then during training, they are resized to 224×224×1.

Target Dataset

In the target data set as shown in Figure 4, we have 58 samples, of which 32 are normal and 26 are abnormal. The size of the images is 1180×786×1. Our purpose is to use a combination of a classifier and a Siamese network to classify this rare sample collection.



(a) Abnormal (b) Normal

Figure 4 (a) A abnormal sample with congenital retinal disease. (b) A normal sample. As can be seen from the abnormal figure, the middle of the abnormal picture, the thickness of the retinal layer will be thinner than normal, which means that some tissue is missing.

Implementation detail

In our experiments, we first pre-process the auxiliary and the target data using the image alignment approach described in the previous sections to eliminate their variability.

We use ResNet-50[11] as the backbone network to learn the feature map from the auxiliary dataset. The input data is entered by randomly sampling the sample composition (anchor, positive, negative) from both normal and abnormal categories. Each epoch we use 80,000 triplets

with a training rate of 1e-4, and each of the five epochs decreased 10-fold, for a total of 20 epochs trained. Margin as a means of separating targets in the feature map was uniformly set to 5 in this experiment. The final output of the model is fixed at 128 dimensions, and this part of the training does not use category as the output, but rather by Euclidean distance comparison.

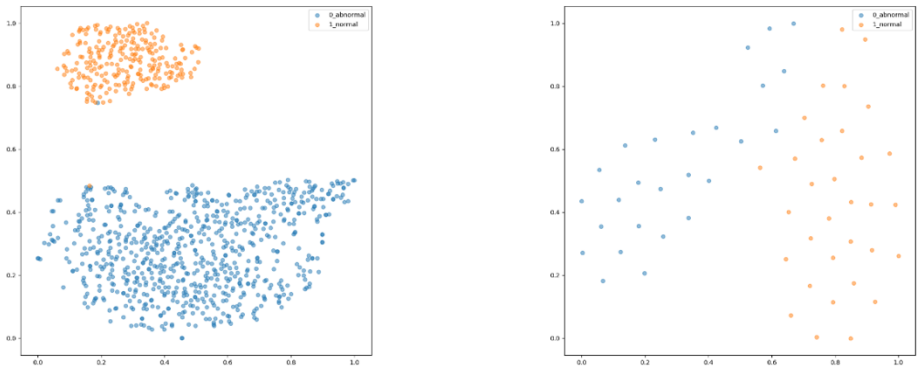
After training the Siamese network, we extract the embedded network and freeze the parameters therein for connecting the classifier. Doing so separates feature learning for large samples from classification for small samples. For the classifier part of the experiment, since the target sample data volume for training was only 44, we use a three-layer fully connected neural network with 64 and 32 and 2 output nodes. Due to the small amount of data, a 5-fold cross-validation method is used for 44 training samples in order to improve classifier performance while preventing fitting.

Extraction of feature map

Since a small amount of data is not enough for the depth neural network to produce a sufficiently valid parameter adjustment, we project samples by category into feature space, which is used to separate the model's identification and classification. Prior to the training of the classifier, the identification section should have the ability to classify the sample points by class.

It can be seen from the Figure 5 that the trained Siamese network is able to distinguish the normal and abnormal samples well in the feature space and the different classes are clustered tightly. By training the auxiliary data, the Siamese network is able to project the data onto the feature map and, specifically, to distinguish between normal and abnormal samples in the feature space, which allows the subsequent classifier to learn the classification directly from the feature space. In the case of scarce data samples, this allows the classifier to work properly compared to feeding the original image directly. Moreover, the reason why the abnormal sample in Figure

5(a) has a larger distribution area than the normal sample is that the abnormal samples were divided into three specific diseases in the original data, for instance, CNV, DME and DRUSEN shown in Figure 3.



(a) Auxiliary Dataset (b) Target Dataset

Figure 5 The 128-dimensional output from Siamese network projected by T-SNE algorithm. Blue dots are abnormal samples, orange dots are normal samples

Classifier Baseline

To experiment with the applicability of the target dataset to existing architectures, we conducted training without pre-training on five existing popular frames after preprocessed images, with 44 samples from the target dataset as training samples and 14 samples as test samples. The results are shown in Table 1.

Table 1 Comparison of various baseline network architectures

Methods	Pre-trained	Accuracy
AlexNet[23]	False	0.5
InceptionV3[24]	False	0.5714
VGG19[25]	False	0.5714
Xception[26]	False	0.5
ResNet-50[27]	False	**0.6428**

As a further implementation of LeNet[28], AlexNet[23] explores for the first time the structure of a convolutional neural network in terms of both width and depth, while practicing the nonlinear activation function of ReLU and Dropout in the paper to reduce overfitting. The VGG[25] network uses smaller convolutional kernels and deepens the network for better results, building on AlexNet. In our experiments, it did show better performance than AlexNet. Compared to AlexNet and VGG, Inception network[24] emerged with multiple branched paths, introduced multiple size convolutional kernel and multiple small size convolutions instead of large size to help reduce the computation of the network, using the inception module to improve training results. Xception[26] adds the concept of depth-wise Separable Convolution on top of Inception to further reduce computation by convolutionally reintegrating each output separately. All of the above network architectures enhance the learning capabilities of the network by increasing the network depth, but this also leads to the degradation problem, which means not only does the test error get higher as the network deepens, but also its training error gets higher. In order to solve this problem, the shortcut connection is implemented, and the input and output of this block are stacked in an element-wise way through shortcut.

The data in the Table 1 does prove that ResNet[27] does have a better performance for OCT images when it absorbs the advantages of Inception and VGG than other network architectures. This led us to use it as the base network in the experiment.

The evolution of architecture

Having identified the adoption of ResNet as our base network, its structure and training methods were explored twice in order to achieve better results. To solve the problem of overfitting due to the model being too large yet the data volume being too small. We decided to use a pre-training

approach, entering a large data set of similar samples to pre-tune the parameters in the model, and then fine-tuning the model to complete the classification task with a small dose of target samples. A three-layer classifier structure is accessed directly after the original structure to form a new network structure.

But ordinary network structures are too huge, and the small amount of data is not enough to adjust the parameters well, and a inappropriate learning rate can mess up the previously pre-trained parameters. So, we decided to introduce the Siamese structure to solve this problem. Use ResNet to project input samples into a sample space of a particular dimension, while having as much distribution of categories as possible i.e. aggregation of samples of the same kind, separation of heterogeneous classes.

Pre-training of 15 epochs was first performed using the general dataset, with a data volume of 83881 for each training and 1000 for the testing, followed by 44 images for few-shot learning. For the Siamese network, the data set is the same, but the Siamese network does not have a classifier for learning feature maps. The classifier is trained with the built-in network parameters frozen and the classifier is only in contact with the target dataset.

Table 2 Comparison of evolutions

Methods	Pre-trained	Accuracy
ResNet	False	0.6428
ResNet+FCNN	True	0.8045±0.0682
ResNet+Siamese+FCNN	True	**0.96±0.04**

From the Table 2, it can be seen that in the feature space projected by the Siamese network, the samples are better identified and classified by the classifier composed of the fully connected network, which serves our purpose.

Variants of classifier

Siamese's embedding network was pre-trained after completing the general data, followed by the classification task. We extracted it from the Siamese structure and made it responsible for mapping, while leaving the classification task to a different classifier independently. Separate feature learning and classifiers both ensure that pre-training is not broken and prevent underfitting for small amounts of data.

In the classification part, two approaches were explored, a fully connected neural network and a combination of principal component analysis (PCA) and a support vector machine (SVM).

Nerual network classifier

Neural network classifiers have inherent nonlinear classification capabilities for high-dimensional data. Combined with the 128-dimensional data, we froze the parameters of the pre-trained built-in network and linked a 3-layer network directly behind it. The neural network classifier consists of a dense layer of length 64, a dense layer of length 32, and an output layer.

SVM classifier

Since the support vector machine is a state-of-the-art method in the field of linear partitioning, we initially applied it directly to the partitioning of 128-dimensional data, but did not achieve satisfactory results after trying Polynomial kernel functions, Gaussian kernel functions, and Radial basis function kernel. So we used principal component analysis (PCA) to implement a downscaling of the feature space data, and SVM to classify the already downscaled feature data.

It can be seen from the table that after the downscaling, the data can be better classified by SVM and the accuracy rate will be greatly improved.

Table 3 Comparison of classifiers

Methods	Pre-trained	Average Accuracy
ResNet+SVM[29]	True	0.5252
ResNet+SVM +PCA[30]	True	0.9440
ResNet+Siamese+FCNN	True	0.9520

You can see from the table that the SVM performs at a faster rate. But neural networks are slightly better at recognition rates and are more flexible, meaning there is more room for adjustment and growth.

Conclusion

We present a few-shot learning based method for classification by using the extracted feature maps, and demonstrated it on a kind of rare congenital retinal diseases images.

We start with a large number of auxiliary images with common retinal diseases to learn the feature maps through a Siamese network. Then input a small amount of target data to let the classifier classify work at the feature map level. It eliminates the problem that too few sample size causes deep neural networks to fail to learn.

This framework proposes an idea of transforming the data space to make it possible to distinguish certain rare diseases. Data preprocessing also plays a critical role that cannot be ignored before training.

References

1. Liu, Y.-Y. *et al.* Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding. *Medical Image Analysis* **15**, 748–759 (2011).
2. Anantrasirichai, N., Achim, A., Morgan, J. E., Erchova, I. & Nicholson, L. SVM-based texture classification in Optical Coherence Tomography. in *2013 IEEE 10th International Symposium on Biomedical Imaging* 1332–1335 (IEEE, 2013). doi:10.1109/ISBI.2013.6556778.
3. Srinivasan, P. P. *et al.* Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed. Opt. Express* **5**, 3568 (2014).
4. Sugmk, J., Kiattisin, S. & Leelasantitham, A. Automated classification between age-related macular degeneration and Diabetic macular edema in OCT image using image segmentation. in *The 7th 2014 Biomedical Engineering International Conference* 1–4 (IEEE, 2014). doi:10.1109/BMEiCON.2014.7017441.
5. Venhuizen, F. G. *et al.* Automated age-related macular degeneration classification in OCT using unsupervised feature learning. in (eds. Hadjiiski, L. M. & Tourassi, G. D.) 94141I (2015). doi:10.1117/12.2081521.
6. Sankar, S. *et al.* Classification of SD-OCT volumes for DME detection: an anomaly detection approach. in (eds. Tourassi, G. D. & Armato, S. G.) 97852O (2016). doi:10.1117/12.2216215.
7. Sun, Y., Li, S. & Sun, Z. Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning. *J. Biomed. Opt* **22**, 016012 (2017).

8. Hussain, M. A. *et al.* Classification of healthy and diseased retina using SD-OCT imaging and Random Forest algorithm. *PLoS ONE* **13**, e0198281 (2018).

9. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402 (2016).

10. Ji, Q., He, W., Huang, J. & Sun, Y. Efficient Deep Learning-Based Automated Pathology Identification in Retinal Optical Coherence Tomography Images. *Algorithms* **11**, 88 (2018).

11. Wang, J. *et al.* Deep learning for quality assessment of retinal OCT images. *Biomed. Opt. Express* **10**, 6057 (2019).

12. Lee, C. S., Baughman, D. M. & Lee, A. Y. Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images. *Ophthalmology Retina* **1**, 322–327 (2017).

13. Awais, M., Muller, H., Tang, T. B. & Meriaudeau, F. Classification of SD-OCT images using a Deep learning approach. in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* 489–492 (IEEE, 2017). doi:10.1109/ICSIPA.2017.8120661.

14. Shih, F. Y. & Patel, H. Deep Learning Classification on Optical Coherence Tomography Retina Images. *Int. J. Patt. Recogn. Artif. Intell.* 2052002 (2019) doi:10.1142/S0218001420520023.

15. Fang, L. & Wang, C. Automatic classification of retinal three-dimensional optical coherence tomography images using principal component analysis network with composite kernels. *J. Biomed. Opt* **22**, 1 (2017).

16. Karri, S. P. K., Chakraborty, D. & Chatterjee, J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. *Biomed. Opt. Express* **8**, 579 (2017).

17. Ji, Q., Huang, J., He, W. & Sun, Y. Optimized Deep Convolutional Neural Networks for Identification of Macular Diseases from Optical Coherence Tomography Images. *Algorithms* **12**, 51 (2019).

18. Lu, W. *et al.* Deep Learning-Based Automated Classification of Multi-Categorical Abnormalities From Optical Coherence Tomography Images. *Trans. Vis. Sci. Tech.* **7**, 41 (2018).

19. Motozawa, N. *et al.* Optical Coherence Tomography-Based Deep-Learning Models for Classifying Normal and Age-Related Macular Degeneration and Exudative and Non-Exudative Age-Related Macular Degeneration Changes. *Ophthalmol Ther* **8**, 527–539 (2019).

20. Koch, G., Zemel, R. & Salakhutdinov, R. Siamese Neural Networks for One-shot Image Recognition. 8.

21. Schroff, F., Kalenichenko, D. & Philbin, J. FaceNet: A unified embedding for face recognition and clustering. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 815–823 (IEEE, 2015). doi:10.1109/CVPR.2015.7298682.

22. Dabov, K., Foi, A., Katkovnik, V. & Egiazarian, K. Image denoising by sparse 3D transform-domain collaborative ltering. **16**, 16.

23. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]* (2015).

25. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2015).

26. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv:1610.02357 [cs]* (2017).

27. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]* (2015).

28. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. in *Proceedings of the IEEE* 2278–2324 (1998).

29. Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273–297 (1995).

30. Jolliffe, I. T. Principal components in regression analysis. in *Principal component analysis* 129–155 (Springer, 1986).