



Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images

An, Guangzhou

Akiba, Masahiro

Omodaka, Kazuko

Nakazawa, Toru

Yokota, Hideo

(Citation)

Scientific Reports, 11(1):4250

(Issue Date)

2021-03-01

(Resource Type)

journal article

(Version)

Version of Record

(Rights)

© The Author(s) 2021. Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) a...

(URL)

<https://hdl.handle.net/20.500.14094/90008100>





OPEN

Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images

Guangzhou An^{1,2,3}, Masahiro Akiba^{1,3}, Kazuko Omodaka⁴, Toru Nakazawa^{3,4} & Hideo Yokota^{2,3}✉

Deep learning is being employed in disease detection and classification based on medical images for clinical decision making. It typically requires large amounts of labelled data; however, the sample size of such medical image datasets is generally small. This study proposes a novel training framework for building deep learning models of disease detection and classification with small datasets. Our approach is based on a hierarchical classification method where the healthy/disease information from the first model is effectively utilized to build subsequent models for classifying the disease into its sub-types via a transfer learning method. To improve accuracy, multiple input datasets were used, and a stacking ensemble method was employed for final classification. To demonstrate the method's performance, a labelled dataset extracted from volumetric ophthalmic optical coherence tomography data for 156 healthy and 798 glaucoma eyes was used, in which glaucoma eyes were further labelled into four sub-types. The average weighted accuracy and Cohen's kappa for three randomized test datasets were 0.839 and 0.809, respectively. Our approach outperformed the flat classification method by 9.7% using smaller training datasets. The results suggest that the framework can perform accurate classification with a small number of medical images.

Artificial intelligence (AI) has been applied in medical image classification via deep learning algorithms trained on massive amounts of supervised data^{1,2}. Gulshan et al.³ used deep learning to create an algorithm for the automated detection of two ocular diseases in retinal fundus photographs based on a dataset of 128,175 retinal images. Rajpurkar et al.⁴ succeeded in classifying 14 different diseases using deep learning based on 112,000 chest X-rays. In addition to applications in disease detection, AI has also been implemented to assist decision making related to treatment. For instance, Esteva et al.⁵ used 130,000 dermatological photographs to build a deep learning model that outperformed human dermatologists in deciding treatment plans for patients with two types of skins cancer. Typically, the more the input data provided, the better a deep learning-based model performs⁶.

However, the supervised data available in the medical field is limited, since high level knowledge is required to prepare such data. Thus, there is a need for techniques for building deep learning models with small amounts of supervised data. Many deep learning techniques are being developed to overcome the obstacle of insufficient supervised data in the medical field. One approach is to synthetically increase the number of available samples for training deep learning models through data augmentation based on the geometrical transformation of images, or to mimic the distributions from which the images are sampled^{1,7}. Another approach is the 'not-so-supervised' learning case, which includes semi-supervised, multi-instance, and transfer learning; among these, transfer learning has recently become the most popular⁸. Transfer learning, inspired by human thought processes, is a method in which model knowledge is effectively transferred across partially related or unrelated tasks to solve a new task with minimal retraining⁹. A study by Kermany, et al.¹⁰ demonstrated the competitive performance of deep learning models built with transfer learning in classifying normal eyes and eyes with three macular diseases using 4000 optical coherence tomography images.

A recent report stated that the doctor's diagnostic decision is hierarchically structured, where a fairly broad initial list of potential diagnoses is then into fewer potential options¹¹. In AI field, there is a similar efficient concept called the hierarchical classification method, in which hierarchically structured local classification models

¹R&D Division, Topcon Corporation, Tokyo, Japan. ²Graduate School of System Informatics, Kobe University, Kobe, Japan. ³Image Processing Research Team, RIKEN Center for Advanced Photonics, RIKEN, Wako, Japan. ⁴Graduate School of Medicine, Tohoku University, Sendai, Japan. ✉email: hyokota@riken.jp

are built according to a predefined hierarchy¹². This method achieved good classification performance not only in fine-grained natural image classification¹³, but also in medical image classification^{14,15}. Although it is effective, there have been few studies to discuss the effectiveness of hierarchical classification to construct deep learning models with small amounts of data.

AI has been developed for supporting the diagnosis of various ophthalmologic diseases. Glaucoma is a leading cause of blindness worldwide, and glaucoma-related blindness is irreversible if not detected early and treated appropriately¹⁶. Glaucoma is regarded as a multifactorial disease, and some ophthalmologists suggest that treatment ought to be categorized by its cause¹⁷. A published guideline that defines four types of retinal optic discs based on their morphology¹⁸ was shown to be a useful addition to the determination of a proper treatment plan in glaucoma management^{19–21}. This classification is difficult for doctors as it is based on subjective assessments of medical images²². Optical coherence tomography (OCT) is a popular imaging modality used in glaucoma diagnosis as it can capture cross-sectional images to form volumetric data of the retina. There have been several studies on diagnosing glaucoma using OCT data with AI^{23–26}; most of these have aimed at automatic detection for classifying normal and glaucoma cases, whereas few have targeted glaucoma classification, which is difficult but crucial for proper treatment of glaucoma patients.

The purpose of this study was to propose a novel training framework for building deep learning models of disease detection and classification by using small datasets. Our approach is based on a hierarchical classification method, where the healthy/disease information from first model is effectively utilized to build subsequent models for classifying the disease into its sub-types using a transfer learning method. The effectiveness of the proposed training framework using a small OCT dataset is demonstrated to classify the glaucoma sub-types.

Methods

Dataset. In this study, images of 156 normal and 798 glaucomatous eyes were obtained from the Tohoku University Hospital database. This study adhered to the tenets of the Declaration of Helsinki, and the protocols were approved by the Clinical Research Ethics Committee of the Tohoku University Graduate School of Medicine (study 2014-1-805). Participants provided their written informed consent to participate in this study. The ethics committees approved this consent procedure, and all methods were carried out in accordance with relevant guidelines and regulations. These images were reviewed and labelled by two glaucoma specialists. The glaucomatous eyes were further classified into four sub-categories based on the optic disc morphology according to the definitions used in Nicoletta classification as focally ischemic (FI), myopic glaucomatous (MY), generalized enlargement (GE), and senile sclerotic (SS) discs. Images with discordant classifications between the two glaucoma specialists were excluded. A total of 156 normal, 118 FI, 266 GE, 307 MY, and 107 SS eye images were used. The demographic data of the final subjects are presented in Supplementary Table S1.

These eye images were captured using swept source OCT (DRI OCT Triton), and a three-dimensional (3D) scan of the disc region (6.0 mm × 6.0 mm) was performed horizontally through 128 B-scans with 512 axial depth scans (A-scans) each. Four types of images, which the doctors used to analyze the 3D data for glaucoma diagnosis, were extracted from the 3D disc scan OCT data and used in our machine learning system: (1) projection images (image 'a' in Fig. 1); (2) en face images (image 'b' in Fig. 1); (3) horizontal B-scan OCT images crossing the disc center (image 'c' in Fig. 1; disc H B-scan); (4) vertical B-scan OCT images crossing the disc center (image 'd' in Fig. 1; disc V B-scan). Extraction and pre-processing details for these four types of images are provided in the subsection titled 'Image Extraction and Pre-processing' in the Supplementary Information.

Proposed approach. First, we analyzed doctors' diagnostic processes of disease detection and classification as follows (shown in the upper part of Fig. 1). (1) Doctors use limited information (e.g., a single type of medical image) to detect diseases. The determination of a treatment plan is difficult, as it requires disease classification based on complex symptoms, and it must be performed after disease detection (excluding normal images). (2) Doctors reuse the knowledge of classifying normal and diseased cases to classify diseased cases into subcategories. (3) For accurate diagnosis, doctors use multiple information sources (e.g., multiple types of medical images) to make optimal diagnostic decisions. Then, according to the analyzed characteristics of doctors' diagnostic processes mentioned above, a novel two-step framework to build deep learning models for disease detection and classification was proposed to achieve high accuracy. In the first step, the model for classifying healthy and disease (disease detection) was built. In the second step, the model for disease classification was built by reusing parameters of the model for disease detection. This two-step framework was further extended to handle multiple input images by combining the machine learning models trained separately on single input images. In detail, we implement a training framework composing of three different methods including hierarchical classification, hierarchy transfer learning, and a stacking ensemble method. The first two components were used to build deep learning classification models with a single type of input image, and the last was used to combine these models. We applied hierarchical classification to separate the training steps of the model classifying normal versus disease cases and the model for disease classification. Hierarchy transfer learning from the model for classifying normal versus disease cases was used to build disease classification models based on the knowledge obtained from classifying normal and disease cases after the step of disease detection. For disease detection, we used transfer learning from a pretrained model on a large visual database (ImageNet) with more than 14 million natural images for visual object recognition. A metamodel was then used to combine the results of the model classifying normal versus disease cases and the model for disease classification using a single type of input images. Finally, a stacking ensemble method was used to combine the separate deep learning models to obtain the overall result (Fig. 1).

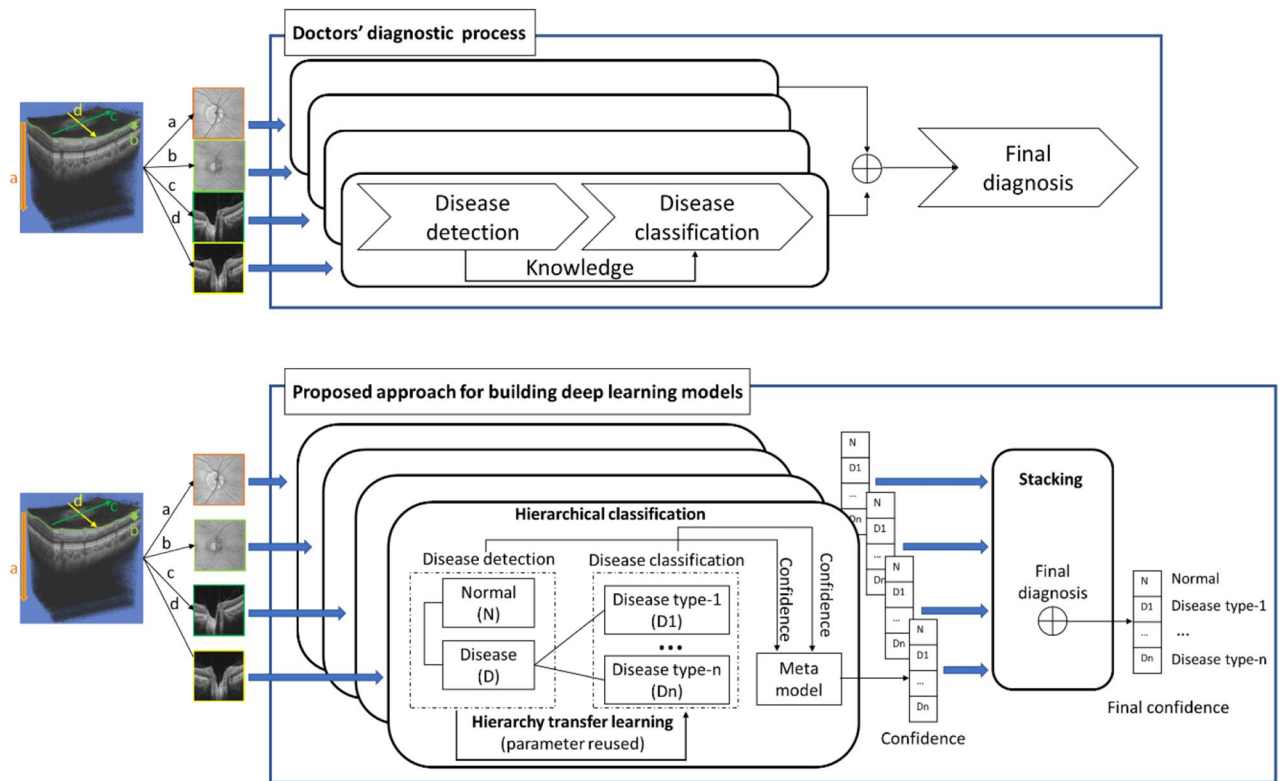


Figure 1. Overview of our main framework for building deep learning models. The top figure illustrates doctors' diagnostic process. The bottom figure illustrates the building of deep learning models based on the analyzed characteristics of the doctors' diagnostic process using three techniques: hierarchical classification, hierarchy transfer learning, and a stacking ensemble method.

Experiments and training. Two experiments were designed to verify our methods. One experiment compared the performance of different training approaches, whereas the other evaluated the applicability of the proposed approaches to small training data.

The first experiment was performed as described below by using the entire set of training data to compare classification performances among models trained using three proposed approaches (Fig. 2). We separately trained deep learning models for each type of extracted image from the 3D OCT data. A convolutional neural network (CNN), which can automatically create efficient image features for the classification, was used as the classifier²⁷. Flat classification models were used to classify eyes as normal, FI, GE, MY, or SS directly with transfer learning from a deep learning model (a CNN) pretrained on the ImageNet dataset (ImageNet-pretrained CNN model) to create Model 1 (Fig. 2a, $n = 4$). We built a hierarchical classification model (Fig. 2b) applying transfer learning from the ImageNet-pretrained CNN model separately with a low-level model (Model 2 in Fig. 2b) for classifying normal versus glaucoma and a high-level model (Model 3 in Fig. 2b) for glaucoma classification ($n = 4$). Furthermore, the normal confidence from Model 2 and the confidence of FI, GE, MY, and SS from Model 3 were concatenated into a confidence vector length of 5. Then, a metamodel of the linear support vector machine (SVM) was trained using the confidence vector data with the supervised labels to combine the models in a cascaded manner²⁸. As described in the section titled 'Proposed Approach', a hierarchical classification model applies transfer learning from the ImageNet-pretrained CNN model to create a low-level model for classifying normal versus glaucoma cases; then, it is hierarchically transferred to build a glaucoma classification model from the low-level model. Finally, a metamodel of the linear SVM was used to calculate the overall result (Fig. 2c).

A stacking ensemble method was used to combine the separately trained single-input image models. The confidence vector calculated by models trained with different single-input images was extracted and concatenated to train the superior metamodel via a linear SVM to combine the single-input models. For comparison with the stacking method, a multiple-input CNN with a direct four-image input was selected to handle multiple images and trained with the three proposed approaches (Fig. 2). A metamodel was also used to combine the results of different-level models to calculate the final classification result.

The second experiment was designed to evaluate the deep learning models' applicability to small amounts of training data built using the proposed approach by using partial training data. Partial training datasets were created using a stratified random sampling strategy using different percentages of the entire training data (25.0%, 37.5%, 50.0%, 62.5%, 75.0%, 87.5%, and 100.0%). With the different training datasets, deep learning models were built with the proposed approaches (Fig. 2) using one type of input image (projection image) and a combination model via a stacking ensemble method for different single-input models.

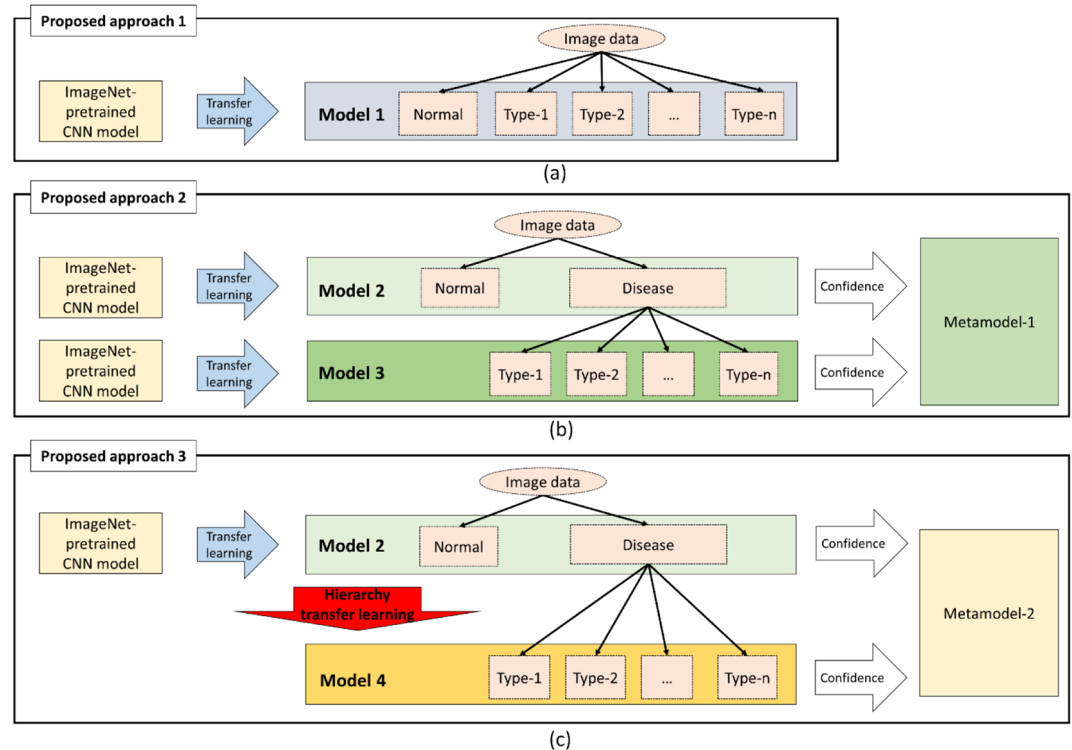


Figure 2. Proposed approaches for building deep learning models. Consider a classification problem (class number = $n + 1$) of normal and disease cases with n subtypes. **(a)** Proposed approach 1: flat classification is used to directly classify normal eyes and those with four subtypes of disease with transfer learning from an ImageNet-pretrained CNN model to create Model 1. **(b)** Proposed approach 2: hierarchical classification is used to create a low-level model (Model 2) for classifying normal versus disease cases and a high-level model (Model 3) for classifying subtypes of disease. Both models apply transfer learning from the ImageNet-pretrained CNN model. The confidence in a 'normal' result from Model 2 and the confidence in disease subtypes from Model 3 are concatenated to calculate the overall result from training Metamodel 1. **(c)** Proposed approach 3: hierarchical classification using hierarchy transfer learning between different-level models is used in the hierarchical classification model is. In contrast with proposed approach 2, the high-level model (Model 4) for classifying disease subtypes, transfer learning is from the low-level model (Model 2) instead of from the ImageNet-pretrained CNN model. The normal confidence from Model 2 and the disease subtype confidence from Model 4 are concatenated to train Metamodel 2 to calculate the overall result.

In this study, we adopted the VGG-16 CNN architecture, which is widely used to solve image classification tasks²⁹, for all the deep learning models and customized it as shown in the Model Architectures and Training Details in the Supplementary Information. In all the model training phases, data augmentation techniques were used to improve the models' classification performance given the limited training data, as presented in the section titled 'Model Architectures and Training Details' in the Supplementary Information.

Evaluation of classification models. The entire dataset was shuffled to create three different training (80%) and test (20%) datasets with a stratified sampling strategy. The training dataset was used to build the classification models, and the test dataset was used to evaluate the models. The average of the classification indexes for the three test datasets was used to evaluate the proposed approaches. Our dataset was not balanced in class distribution; thus, Cohen's kappa and weighted accuracy were used to evaluate the deep learning models. Cohen's kappa coefficient is a statistic that measures the inter-rater reliability of qualitative items by considering the possibility of agreements occurring by chance³⁰. The agreement thresholds used in this study were based on the following guideline: 0.61–0.80 corresponds to good agreement, and 0.81–1 corresponds to almost perfect agreement³⁰. The weighted accuracy is computed by taking the average, over all the classes, of the fraction with the number of correctly predicted instances in that class, divided by the total number of instances in that class³¹. We used a paired t-test P -value (two-tailed) of 0.05 as the significance level to determine whether the mean difference between two sets of Cohen's kappa was zero.

Results

Performance comparison of different training approaches. The Cohen's kappa for the deep learning models of single-input images built with the flat classification strategy and transfer learning from the ImageNet-pretrained model (proposed approach-1 in Fig. 2a) were 0.567 for projection, 0.600 for en face, 0.550 for disc H

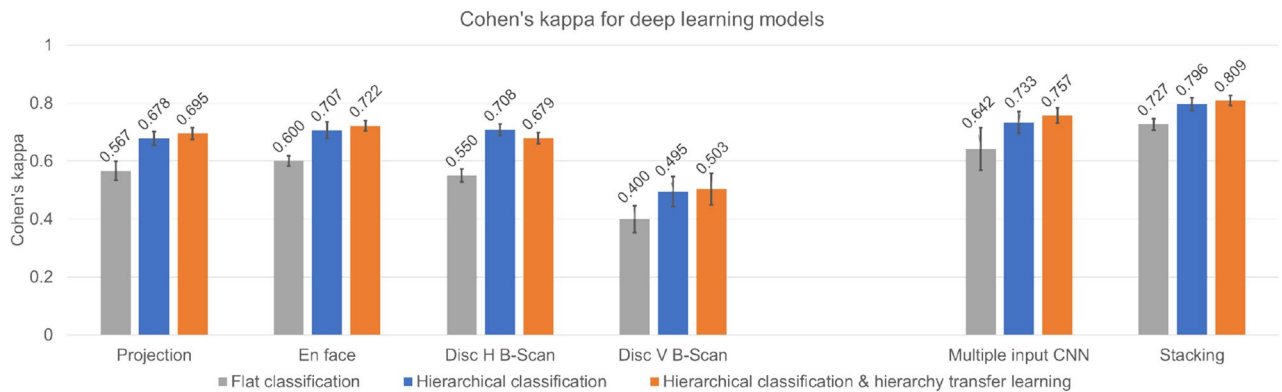


Figure 3. Classification performances of deep learning models built with the proposed approaches. The figure shows Cohen's kappa with standard deviation error bars for the deep learning models for projection, en face, disc H B-scan, and disc V B-scan images and combination models using multiple-input CNN and stacking built with different proposed strategies.

B-scan, and 0.400 for disc V B-scan. Only the deep learning model using en face images achieved good performance with Cohen's kappa over 0.6 (Cohen's kappa: 0.600; weighted accuracy: 66.3%). The Cohen's kappa for the deep learning models built using hierarchical classification without hierarchy transfer learning (proposed approach-2 in Fig. 2b) were 0.678 for projection, 0.707 for en face, 0.708 for disc H B-scan, and 0.495 for disc V B-scan. The deep learning models achieved good performance with Cohen's kappa over 0.6 when using projection (Cohen's kappa: 0.678; weighted accuracy: 75.5%), en face (Cohen's kappa: 0.707; weighted accuracy: 74.6%), and disc H B-scan (Cohen's kappa: 0.708; weighted accuracy: 77.5%) images. The deep learning models using hierarchical classification models provided significantly improved classification performance (P-values of the paired t-test: 0.027, 0.045, and 0.016, respectively) compared to models using the flat classification strategy for the same input images. The Cohen's kappa for the deep learning models built with hierarchical classification and hierarchy transfer learning (proposed approach-3 in Fig. 2c) were 0.695 for projection, 0.722 for en face, 0.679 for disc H B-scan, and 0.503 for disc V B-scan. The deep learning models with the projection, en face, disc H B-scan as the input achieved good performance (projection (Cohen's kappa: 0.695; weighted accuracy: 76.4%), en face (Cohen's kappa: 0.722; weighted accuracy: 77.3%), and disc H B-scan (Cohen's kappa: 0.679; weighted accuracy: 74.4%)). The deep learning models using hierarchical classification models provided significantly improved classification performance (P-values of the paired t-test: 0.014, 0.037, and 0.021, respectively) compared to models using the flat classification strategy for the same input images. There was no considerable difference between hierarchical classification with or without hierarchy transfer learning for all types of input images (left part of Fig. 3).

Our stacked method of four models that were built with transfer learning and pretrained on ImageNet provided satisfactory performance, with a Cohen's kappa of 0.727 and weighted accuracy of 71.8%. The multiple-input CNN models achieved a Cohen's kappa of 0.642 and weighted accuracy of 69.2% with four input images (right part of Fig. 3). The standard deviation of Cohen's kappa for stacking was smaller than that of the multiple-input CNN trained with the same proposed approaches. The multiple-input CNN was not significantly higher than the single-input CNNs of projection images, en face images, and disc H B-scans. However, classification performance significantly improved with the stacking method for each single-input model (P-values of the paired t-test: 0.014, 0.022, 0.006, and 0.012 for projection, en face, disc H B-scan, and disc V B-scan images, respectively).

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

Applicability evaluation of the proposed approaches in small training data. The performance change (as measured using Cohen's kappa) provided by the deep learning classification models using projection input images was compared. All the models trained with flat classification (FC; Fig. 2a), hierarchical classification without hierarchy transfer learning (HC; Fig. 2b), and hierarchical classification with hierarchy transfer learning (HC & HTL; Fig. 2c) using a larger training dataset achieved an increase in Cohen's kappa. Only HC and HC & HTL achieved satisfactory performance when using the entire training dataset. There was no significant difference between HC and FC in the case of a small training dataset. The HC & HTL strategy achieved a good Cohen's kappa when using 62.5% of the entire training dataset (Fig. 4a).

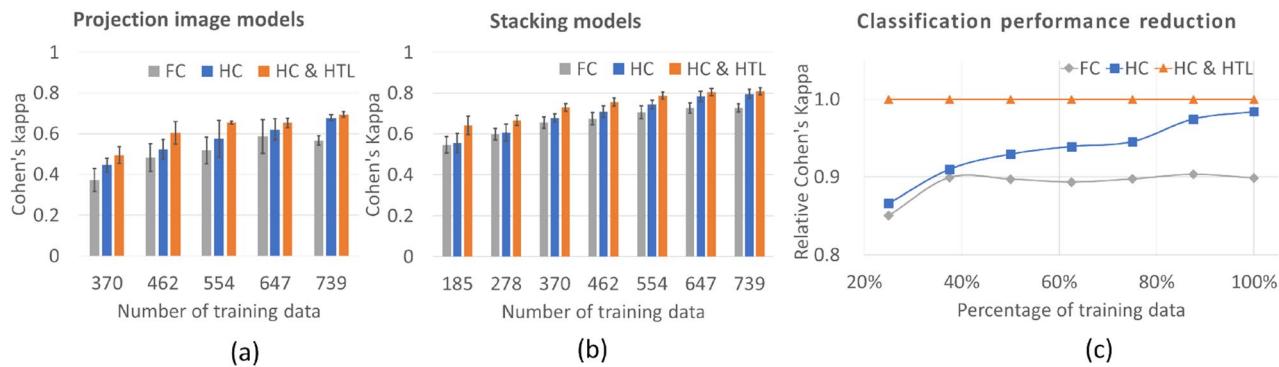


Figure 4. Performance change with different training dataset sizes. **(a)** Cohen’s kappa with standard deviation error bars for deep learning models trained with different training methods based on single-input (projection) images. **(b)** Cohen’s kappa with standard deviation error bars for deep learning models trained with different training methods based on all input images. **(c)** Calculated classification performance reduction for models using different training datasets and other training methods with stacking, with respect to Cohen’s kappa for the CNN model built with the proposed method of HC & HTL with stacking.

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

Table 1. Performance change of stacking models with different training dataset sizes.

The performance changes due to the stacked deep learning classification models using all the types of input images were compared. All the stacked models trained with FC, HC, and HC & HTL achieved a higher Cohen’s kappa (Table 1) compared to the single-input (projection) model using the same number of training datasets. With sufficient training data, all the models achieved convergence classification performance. With the same number of training datasets, the classification performance of HC & HTL was better than that of FC (Table 1, Fig. 4b; all *P*-values of paired *t*-tests < 0.05). The performance of the models built with the HC strategy was better than that of the models built with the FC strategy for large training datasets (more than 75.0% of the training dataset). However, there was no significant difference between HC and FC for a small training dataset (less than 75.0% of training dataset). The HC & HTL strategy achieved a fairly good Cohen’s kappa of 0.642 when using only 25.0% of the entire training dataset (Table 1, Fig. 4b). The performance reduction (relative Cohen’s kappa) for each model using each training dataset was calculated with reference to the Cohen’s kappa for the CNN model built with the HC & HTL strategy (Fig. 4c). The performance reduction of flat classification was greater than $(1.000 - 0.903) \times 100.0\% = 9.7\%$ compared to the use of the proposed method with the same training dataset (Table 1, Fig. 4c). The performance reduction of hierarchical classification was smaller than flat classification for each size of the training dataset, while the rate of performance reduction for hierarchical classification was larger than that with flat classification (Table 1, Fig. 4c).

Discussion

In this study, we built high-accuracy deep learning models for disease detection and classification. Experiments showed that the deep learning models trained with transfer learning from an ImageNet-pretrained CNN model with flat classification and data augmentation performed effectively in disease detection and classification on one type of input images (Fig. 3). The approach proposed by Kermany et al.¹⁰ is similar to proposed approach-1 (Fig. 2) as we adopted data augmentation for model training, which is useful in small datasets. We proposed a two-step framework to build deep learning models for disease detection and classification based on the characteristics of the doctors’ diagnostic processes analyzed; the first step builds the model for disease detection, and the second step builds the model for disease classification by reusing parameters of the model for disease detection. Hierarchical classification was applied first for classifying diseased cases into subcategories after excluding the normal cases. The deep learning models built with hierarchical classification performed well and produced substantially improved results compared to flat classification for three types of input images (Fig. 3). Similar to our proposed approach-2 (Fig. 2), a previous study used hierarchy knowledge in disease

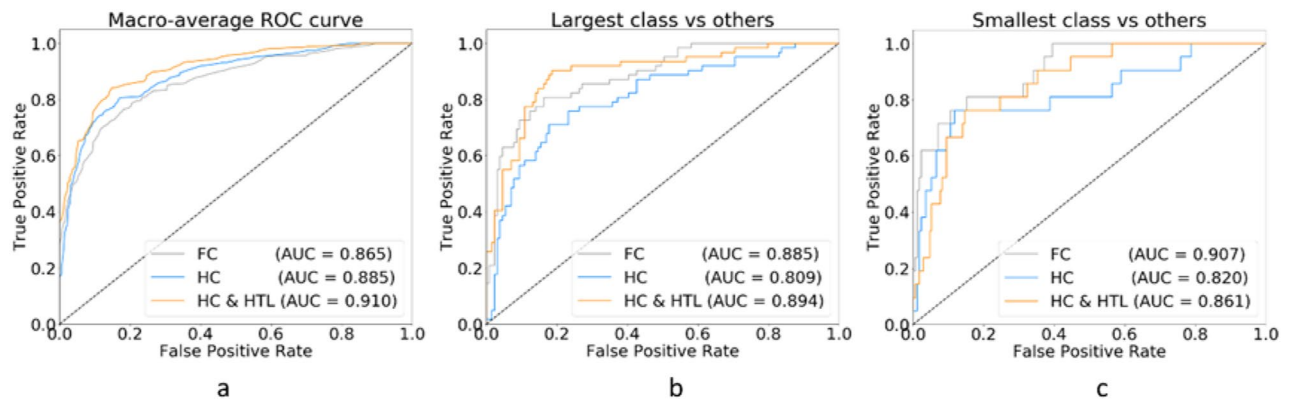


Figure 5. ROC curves for models built using a small training dataset. ROC curves for the stacked hierarchical classification model using hierarchy transfer learning: (a) macro-average ROC curve of all five classes, (b) ROC curves for the classification of the largest data class (MY) versus others, and (c) ROC curves for the classification of the smallest data class (SS) versus others.

classification and transfer learning from a pretrained model with a large natural image dataset and achieved a high accuracy³². We evaluated the applicability of this approach in relatively small datasets and found that it achieved higher performance than flat classification. As the variants for normal and diseased cases are different, we believe that hierarchically building models for disease detection and classification models can contribute to performance improvement. Moreover, we applied similar architectures for low-level and high-level models in the hierarchical classification model to further apply hierarchy transfer learning so that the parameters of models for classifying normal and diseased cases based on the analysis of the doctors' diagnostic processes can be reused for classifying diseases into subcategories, which utilizes the knowledge of classifying normal and diseased cases. The combination of two methods (proposed approach-3) produced higher classification performance using smaller datasets than flat and hierarchical classification. While the knowledge of classifying normal and diseased cases is important to consider when classifying diseases into subcategories, it is believed that the hierarchy transfer learning method that attempts to reuse the parameters contributes to performance improvement when using relatively small training datasets.

We proposed two methods for handling multiple-input images based on doctors' requirement of multiple images to make accurate diagnosis. As described in glaucoma guidelines, multiple sources of information, such as the patient's medical history, visual acuity, assessment of the nerve fiber layer, and visual fields, are required to accurately assess early-stage glaucoma³³. We extracted multiple useful images from one set of volumetric data to develop our deep learning models. The classification performance improved and was good for the multiple-input CNN and stacking methods. In comparison with the multiple-input CNN method, our stacking method considerably boosted the performance of four single-input CNNs with a small standard deviation. Doctors' processes of making a diagnosis after asynchronously interpreting multiple images was imitated with the stacking method.

Our proposed approach of hierarchical classification and hierarchy transfer learning (approach-3; HC & HTL) showed good compatibility with the stacking method and achieved higher classification performance than other proposed approaches. The classification performance was good even for small training datasets. The stacked deep learning models trained with three approaches (Fig. 2) achieved good classification performance using a small dataset (37.5% of the training dataset). The area under the receiver operating characteristic (ROC) curve (AUC) for each class, with the confidence calculated by three different proposed methods, was used to confirm the classification performance without a fixed threshold to predict. We found that our proposed approach of combining all the techniques achieved the best macro-average AUC (Fig. 5a), followed by approach-1 (FC) and approach-2 (HC), in recognizing normal, FI, and GE cases. For the largest class (MY) and smallest (SS) glaucoma-type classes in HC (Fig. 5b,c), the problem of imbalance affected performance, even when weighted loss was applied during training. The HC & HTL method provided the best classification performance of recognizing the largest class among the proposed approaches (Fig. 5b). In future studies, methods for overcoming the imbalance problem of the smallest class should be investigated.

A deep learning model for glaucoma detection and classification was built using the proposed framework that is implemented by hierarchical classification, hierarchy transfer learning, and the stacking ensemble method using different images, achieving a high Cohen's kappa of 0.809. Based on the previous guideline of Cohen's kappa³⁰, this value implies a high performance of the decision support system for glaucoma clinical care. For comparison, we also conducted another diagnosis test. We randomly sampled images of 50 cases from the dataset used in this study, which were then classified into normal and glaucoma sub-categories by three medical ophthalmology interns. The average Cohen's kappa of these classification results with supervised data was 0.408, lower than that of our deep learning model. The model can be used in diagnosis support of glaucoma by providing the confidence level for normal cases and each disc type, which would help doctors to select proper treatment plans according to the optic disc shape^{20,34}.

Our model has a high potential for the automatic diagnosis of glaucoma. First, as more data are collected in future, the performance of deep learning models would improve. Next, many new deep learning techniques have been reported recently, such as deep learning architectures, which can boost the performance of deep

learning models. Hopefully, deep learning models built with these new techniques and our approaches would achieve even better performance for glaucoma detection and classification. Finally, the images for which grading is difficult might be used in a semi-supervised classification manner to improve the performance of our deep learning models.

This study is novel in that it proposes a deep learning training framework using hierarchical classification and hierarchy transfer learning capable of building models that perform accurate disease detection and classification using a small number of images. The high potential of deploying deep learning models with a small number of datasets was demonstrated for practical diagnosis based on the characteristics of the clinical diagnostic processes.

Conclusions

In this study, we proposed a novel two-step framework using hierarchy transfer learning to build deep learning models for disease detection and classification based on an analysis of the characteristics of doctors' diagnostic processes. This training framework enabled the built deep learning models to achieve high accuracies and improved classification performances in comparison with models based on other training approaches. Moreover, the two-step framework was extended to handle multiple input images by combining models created with each type of input image with a stacking method to further improve the accuracy of doctors' diagnoses. We showed that high accuracy and robustness can be achieved with a smaller number of training data using the proposed training framework. The effectiveness of the proposed approach and its applicability to diagnostic classification in medical images were demonstrated using a retinal image dataset relevant in glaucoma diagnosis. These results indicate the high potential of deploying deep learning models with a limited number of medical images for practical diagnoses through the analysis of the clinical diagnostic process.

Data availability

The images and diagnosis data used to support the findings of this study have not been made available because they are real clinical data from Tohoku University Hospital, and the patients' rights to privacy should be protected, as it is possible to identify people from these data, but are available from the corresponding author on reasonable request.

Received: 11 May 2020; Accepted: 11 January 2021

Published online: 01 March 2021

References

- Ker, J., Wang, L., Rao, J. & Lim, T. Deep learning applications in medical image analysis. *IEEE Access* **6**, 9375–9389. <https://doi.org/10.1109/ACCESS.2017.2788044> (2018).
- Altaf, F., Islam, S., Akhtar, N. & Janjua, N. K. Going deep in medical image analysis: Concepts, methods, challenges and future directions. <https://arxiv.org/abs/1902.05655> (2019).
- Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410. <https://doi.org/10.1001/jama.2016.17216> (2016).
- Rajpurkar, P., *et al.* CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. <https://arxiv.org/abs/1711.05225> (2017).
- Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118. <https://doi.org/10.1038/nature21056> (2017).
- Bach, S. H., He, B., Ratner, A. & Re, C. Learning the structure of generative models without labeled data. *ICML* 273–282 (2017).
- Goodfellow, I., *et al.* Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 2672–2680 (2014).
- Cheplygina, V., de Bruijne, M. & Pluim, J. P. W. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **54**, 280–296. <https://doi.org/10.1016/j.media.2019.03.009> (2019).
- Torrey, L. & Shavlik, J. *Transfer Learning*. (IGI Global, 2009).
- Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**(5), 1122–1131. <https://doi.org/10.1016/j.cell.2018.02.010> (2018).
- Committee on Diagnostic Error in Health Care; Board on Health Care Services; Institute of Medicine; The National Academies of Sciences, Engineering, and Medicine. *Improving Diagnosis in Health Care*. (Balogh E.P., Miller B.T. & Ball J.R. Eds.) <https://www.ncbi.nlm.nih.gov/books/NBK338593/> (2015).
- Silla, C. N. & Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* **22**(1–2), 31–72. <https://doi.org/10.1007/s10618-010-0175-9> (2010).
- Zhao, J., Peng, Y. & He, X. Attribute hierarchy based multi-task learning for fine-grained image classification. *Neurocomputing* **395**, 150–159. <https://doi.org/10.1016/j.neucom.2018.02.109> (2020).
- Wang, J. & Yang, Y. A hierarchical learning approach for detection of clustered microcalcifications in mammograms. in *Proceedings of 2019 IEEE International Conference on Image Processing (ICIP)* 804–808. <https://doi.org/10.1109/ICIP.2019.8802984> (2019).
- Wang, J., Bai, Y. & Xia, B. Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning. *IEEE J. Biomed. Health Inform.* **24**(12), 3397–3407. <https://doi.org/10.1109/JBHI.2020.301254715> (2020).
- Quigley, H. A. & Broman, A. T. The number of people with glaucoma worldwide in 2010 and 2020. *Br. J. Ophthalmol.* **90**(3), 262–267. <https://doi.org/10.1136/bjo.2005.081224> (2006).
- Nakazawa, T. Ocular blood flow and influencing factors for glaucoma. *Asia Pac. J. Ophthalmol. (Phila)* **5**, 38–44. <https://doi.org/10.1097/APO.0000000000000183> (2016).
- Nicolela, M. T. & Drance, S. M. Various glaucomatous optic nerve appearances: clinical correlations. *Ophthalmology* **103**, 640–649 (1996).
- Omodaka, K. *et al.* Characteristic correlations of the structure-function relationship in different glaucomatous disc types. *Jpn. J. Ophthalmol.* **59**, 223–229 (2015).
- Nakazawa, T. *et al.* Progression of visual field defects in eyes with different optic disc appearances in patients with normal tension glaucoma. *J. Glaucoma* **21**, 426–430. <https://doi.org/10.1097/IJG.0b013e3182182897> (2012).
- Shin, H. Y., Park, H. Y., Jung, Y., Choi, J. A. & Park, C. K. Glaucoma diagnostic accuracy of optical coherence tomography parameters in early glaucoma with different types of optic disc damage. *Ophthalmology* **121**(10), 1990–1997. <https://doi.org/10.1016/j.optha.2014.04.030> (2014).

22. Omodaka, K. *et al.* Classification of optic disc shape in glaucoma using machine learning based on quantified ocular parameters. *PLoS ONE* **12**(12), e0190012. <https://doi.org/10.1371/journal.pone.0190012> (2017).
23. Asaoka, R. *et al.* Validating the usefulness of the 'Random Forests' classifier to diagnose early glaucoma with optical coherence tomography. *Am. J. Ophthalmol.* **174**, 95–103. <https://doi.org/10.1016/j.ajo.2016.11.001> (2017).
24. Maetschke, S., *et al.* A feature agnostic approach for glaucoma detection in OCT volumes. <https://arxiv.org/abs/1807.04855> (2018).
25. An, G., *et al.* Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. *J. Healthc. Eng.* 4061313, <https://doi.org/10.1155/2019/4061313> (2019).
26. Maetschke, S. *et al.* A feature agnostic approach for glaucoma detection in OCT volumes. *PLoS ONE* **14**(7), e0219126. <https://doi.org/10.1371/journal.pone.0219126> (2019).
27. Becker, S. & LeCun, Y. Improving the convergence of back-propagation learning with second-order methods. in *Proceedings of the 1988 Connectionist Models Summer School* 29–37 (1989).
28. Nakano, F. K., Mastelini, S. M., Barbon, S. & Cerri, R. Stacking methods for hierarchical classification. *Int. Jt. Conf. Neural Netw. (IJCNN)* 539–2546, <https://doi.org/10.1109/ICMLA.2017.0-145> (2017).
29. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. <http://arxiv.org/abs/1409.1556> (2014).
30. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174. <https://doi.org/10.2307/2529310> (1977).
31. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. in *Proceedings of the 20th International Conference on Pattern Recognition*. 3121–3124, <https://doi.org/10.1109/ICPR.2010.764> (2010).
32. Zhang, R., *et al.* Aggressive posterior retinopathy of prematurity automated diagnosis via a deep convolutional network. in *Ophthalmic Medical Image Analysis* (H. Fu, M. K. Garvin, T. MacGillivray, Y. Xu, & Y. Zheng eds.)—6th International Workshop, OMIA 2019, Held in Conjunction with MICCAI 2019, 165–172 (2019).
33. American Optometric Association Original Consensus. Care of the patient with open angle glaucoma. <https://www.aoa.org/documents/optometrists/CPG-9.pdf> (2011).
34. Nakazawa, T. *et al.* Different types of optic disc shape in patients with advanced open-angle glaucoma. *Jpn. J. Ophthalmol.* **54**(4), 291–295 (2010).

Acknowledgements

This study was a joint research project of Topcon Corporation, RIKEN, and Tohoku University. We would like to thank Editage (<https://www.editage.com>) for English language editing.

Author contributions

G.A. performed the experiments and wrote the manuscript. M.A., O.K., and T.N. provided comments and feedback on the paper and the results. H.Y. supervised the project. All authors reviewed the manuscript.

Competing interests

G.A. and M.A. are employees of Topcon Corporation. The other authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-83503-7>.

Correspondence and requests for materials should be addressed to H.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021