ORIGINAL RESEARCH
published: 07 June 2021
doi: 10.3389/fmed.2021.654696

A Novel Hierarchical Deep Learning Framework for Diagnosing Multiple Visual Impairment Diseases in the Clinical Environment

Jiaxu Hong[1,2,3,4*†], *Xiaoqing Liu*[5*†], *Youwen Guo*[6‡], *Hao Gu*[2], *Lei Gu*[7,8], *Jianjiang Xu*[1], *Yi Lu*[1], *Xinghuai Sun*[1], *Zhengqiang Ye*[1], *Jian Liu*[2], *Brock A. Peters*[9] *and Jason Chen*[9‡]

[1] Department of Ophthalmology and Visual Science, Eye, and Ear, Nose, and Thorat Hospital, Shanghai Medical, College Fudan University, Shanghai, China, [2] Department of Ophthalmology, Affiliated Hospital of Guizhou Medical University, Guiyang, China, [3] Key Laboratory of Myopia, Ministry of Health (Fudan University), Shanghai, China, [4] Shanghai Engineering Research Center of Synthetic Immunology, Fudan University, Shanghai, China, [5] AI Laboratory, Deepwise Healthcare, Beijing, China, [6] Wuhan Servicebio Technology, Wuhan, China, [7] Epigenetics Laboratory, Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany, [8] Cardiopulmonary Institute (CPI), Bad Nauheim, Germany, [9] Complete Genomics Inc., San Jose, CA, United States

OPEN ACCESS

Edited by:
Liang Zhang,
Xidian University, China

Reviewed by:
Madhura Ingalhalikar,
Symbiosis International University, India
Guangming Zhu,
Xidian University, China

***Correspondence:**
Xiaoqing Liu
xiaoqing.liu@ieee.org
Jiaxu Hong
jiaxu_hong@163.com

† These authors have contributed equally to this work and share first authorship

‡ORCID:
Youwen Guo
orcid.org/0000-0001-9932-1047
Jason Chen
orcid.org/0000-0002-7627-3019

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *17 January 2021*
Accepted: *07 May 2021*
Published: *07 June 2021*

Citation:
Hong J, Liu X, Guo Y, Gu H, Gu L, Xu J, Lu Y, Sun X, Ye Z, Liu J, Peters BA and Chen J (2021) A Novel Hierarchical Deep Learning Framework for Diagnosing Multiple Visual Impairment Diseases in the Clinical Environment. Front. Med. 8:654696. doi: 10.3389/fmed.2021.654696

Early detection and treatment of visual impairment diseases are critical and integral to combating avoidable blindness. To enable this, artificial intelligence–based disease identification approaches are vital for visual impairment diseases, especially for people living in areas with a few ophthalmologists. In this study, we demonstrated the identification of a large variety of visual impairment diseases using a coarse-to-fine approach. We designed a hierarchical deep learning network, which is composed of a family of multi-task & multi-label learning classifiers representing different levels of eye diseases derived from a predefined hierarchical eye disease taxonomy. A multi-level disease–guided loss function was proposed to learn the fine-grained variability of eye disease features. The proposed framework was trained for both ocular surface and retinal images, independently. The training dataset comprised 7,100 clinical images from 1,600 patients with 100 diseases. To show the feasibility of the proposed framework, we demonstrated eye disease identification on the first two levels of the eye disease taxonomy, namely 7 ocular diseases with 4 ocular surface diseases and 3 retinal fundus diseases in level 1 and 17 subclasses with 9 ocular surface diseases and 8 retinal fundus diseases in level 2. The proposed framework is flexible and extensible, which can be inherently trained on more levels with sufficient training data for each subtype diseases (e.g., the 17 classes of level 2 include 100 subtype diseases defined as level 3 diseases). The performance of the proposed framework was evaluated against 40 board-certified ophthalmologists on clinical cases with various visual impairment diseases and showed that the proposed framework had high sensitivity and specificity with the area under the receiver operating characteristic curve ranging from 0.743 to 0.989 in identifying all identified major causes of blindness. Further assessment of 4,670 cases in a tertiary eye center also demonstrated that the proposed framework achieved a high identification accuracy rate for different visual impairment diseases compared with that of human graders in a clinical setting. The proposed hierarchical deep learning framework would improve clinical practice in ophthalmology and broaden the scope of service available, especially for people living in areas with a few ophthalmologists.

Keywords: artificial intelligence, hierarchical deep learning framework, visual impairment disease, coarse-to-fine, multi-task multi-label

INTRODUCTION

Eye diseases leading to visual impairment are a significant source of social burden. It is estimated that, as of 2017, 1 billion people were living with vision impairment worldwide, including those with moderate or severe distance vision impairment or blindness caused by unaddressed refractive error (123.7 million), cataract (65.2 million), glaucoma (6.9 million), corneal opacities (4.2 million), diabetic retinopathy (3.0 million), and trachoma (2.0 million), as well as near vision impairment caused by unaddressed presbyopia (826.0 million) (1). In China, the most frequent cause of visual impairment is cataract, which is followed by corneal disease and glaucoma (2, 3). In contrast, age-related macular degeneration and diabetic retinopathy are more prevalent in the United States (4). Early detection and treatment of visual impairment diseases are critical and integral to combating this avoidable blindness worldwide.

A slit-lamp investigation of the ocular surface and retina using manual interpretation is a widely accepted screening tool to detect visual impairment diseases. However, this is highly dependent on the ophthalmologist's clinical experience, which is time-consuming and may have an interobserver variation on the same patient. Automated identification of various visual impairment diseases via slit-lamp photography has benefits such as increased efficiency, reproducibility, and access to eye care. To enable this, artificial intelligence (AI)-based approaches for the identification of visual impairment diseases are greatly needed, especially for people living in areas with a limited number of ophthalmologists.

Recent advances in AI, particularly convolutional neural networks (CNN)-based deep learning algorithms, have made it possible to learn the most predictive disease features directly from medical images given a large dataset of labeled examples (5, 6). Esteva et al. (7) proposed a dermatologist-level classification of skin cancer by fine-tuning a pretrained Inception-v3 network (8). Menegola et al. (9) also conducted experiments comparing training from scratch with fine-tuning of pretrained networks on skin lesion images. Their study showed that fine-tuning of pretrained networks worked better than training from scratch. Setio et al. (10) applied a multi-view CNN to classify points of interest in chest computed tomography as nodules or non-nodules. Similarly, Nie et al. (11) used a three-dimensional CNN on magnetic resonance images to assess the survival of patients suffering from brain tumors.

Because of the fine-grained variability in the appearance of eye lesions, most of the existing eye disease identification methods focused on a single disease type (such as retinopathy and macular diseases) via retinal fundus or optical coherence tomography (OCT) images. Gulshan et al. (12) demonstrated the detection of diabetic retinopathy by fine-tuning a pretrained Inception-v3 network on retinal fundus images. Similarly, Gargeya and Leng (13) performed automated identification of diabetic retinopathy using a ResNet-based architecture. Li et al. (14) adopted an Inception-v3 network to detect glaucomatous optic neuropathy using color fundus images, whereas Burlina et al. (15) applied both a pretrained model and a newly trained from a scratch model for automated grading of age-related macular degeneration from color fundus images. Schlegl et al. (16) and Treder et al. (17) proposed automated detection of macular diseases using OCT images. Long et al. (18) developed a technique for the diagnosis of congenital cataracts. However, their method was focused on images covering the pupil area only; therefore, their algorithm could not detect diseases affecting the peripheral cornea and limbus. To date, there have been few studies diagnosing ocular surface diseases or identifying various disease types simultaneously. Ting et al. (19) proposed a deep learning system for diabetic retinopathy and related eye diseases using retinal images. Fauw et al. (20) proposed an Ensemble-based deep learning framework that could make referral suggestions on retinal diseases by analyzing OCT images. Li et al. (21) presented a workflow for the segmentation of anatomical structures and annotation of pathological features in slit-lamp images, which improved the performance of a deep learning algorithm for diagnosing ophthalmic disorders. As most of these algorithms have been derived from datasets of one or a few ocular diseases, they struggle to detect visual impairment diseases accurately in large-scale, heterogeneous datasets.

To maximize the clinical utility of AI, we developed a hierarchical deep learning framework, which enables early screening and differentiation of a large variety of visual impairment diseases simultaneously in a coarse-to-fine manner. Here, a hierarchical architecture means that multiple classification layers are arranged in a hierarchical way for different levels. To test the feasibility of the proposed framework, we identified eye diseases on two different levels of the eye disease taxonomy. Thereby, in our case, the proposed framework would first perform disease classification for a lower level (i.e., level 1) and then perform a higher-level disease classification (i.e., level 2). Also, algorithm performance was tested against 40 ophthalmologists in a clinic-based dataset. Finally, we performed an observational diagnostic assessment comparison of visual impairment disease screening between the algorithm and the ophthalmologists in a tertiary eye center.

MATERIALS AND METHODS

Datasets

Our dataset came from two major eye centers in China: (i) the Eye and ENT Hospital of Fudan University, Shanghai, and (ii)

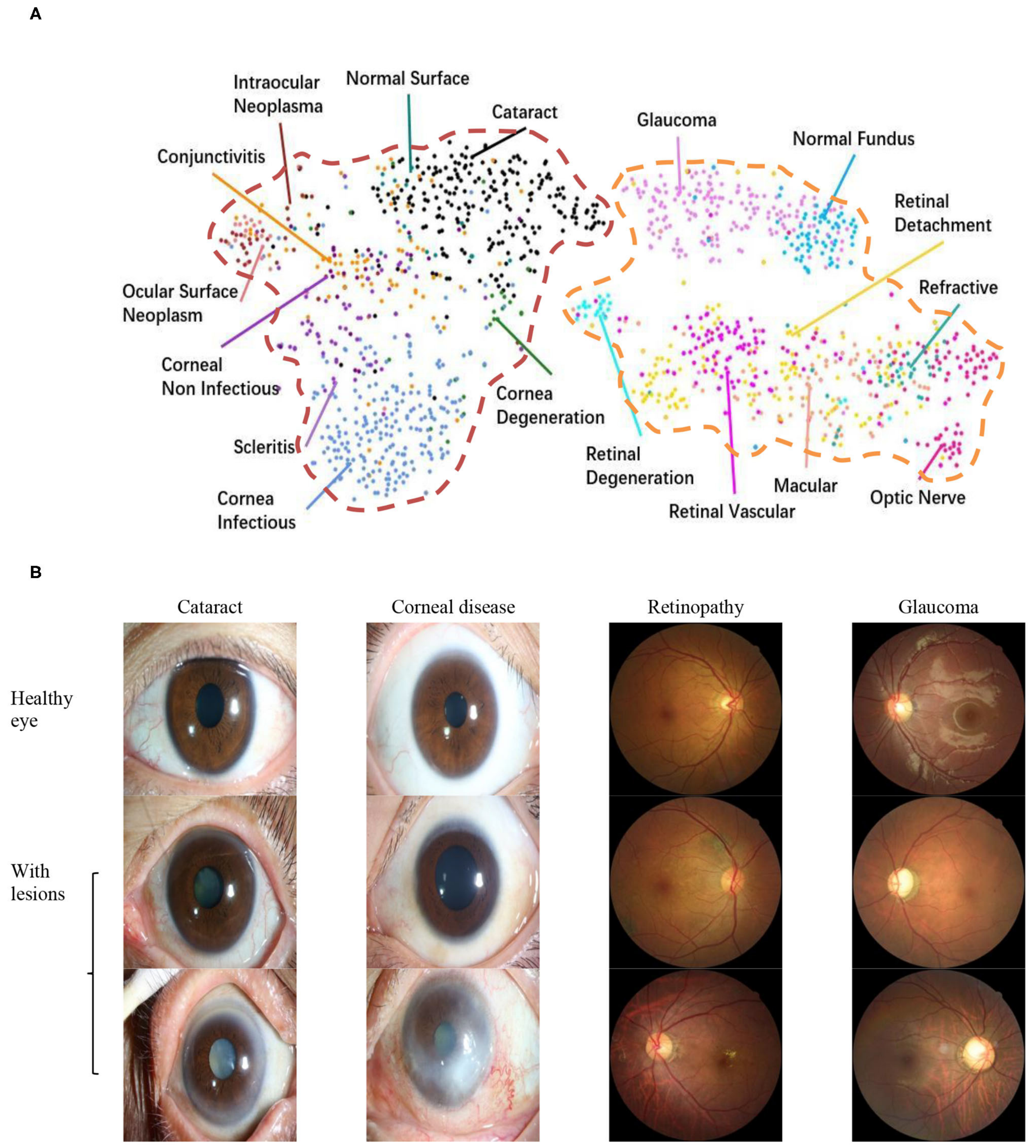


FIGURE 1 | Dataset. **(A)** t-Distributed stochastic neighbor embedding visualization of the collected dataset consisting of 17 major ocular disease classes (100 subtypes), leading to visual impairment, clustered according to deep features generated from the last layer of trained networks. Colored point clouds represent images with different visual impairment diseases. This visualization represents the ability of our method to objectively separate normal patients from early cases of visual impairment diseases for referral. **(B)** Example ocular surface and retinal images for the eye with some common diseases or healthy eye. In this study, the first two levels of the taxonomy consisting of 17 major ocular disease classes (100 subtypes) were used in performance evaluation.

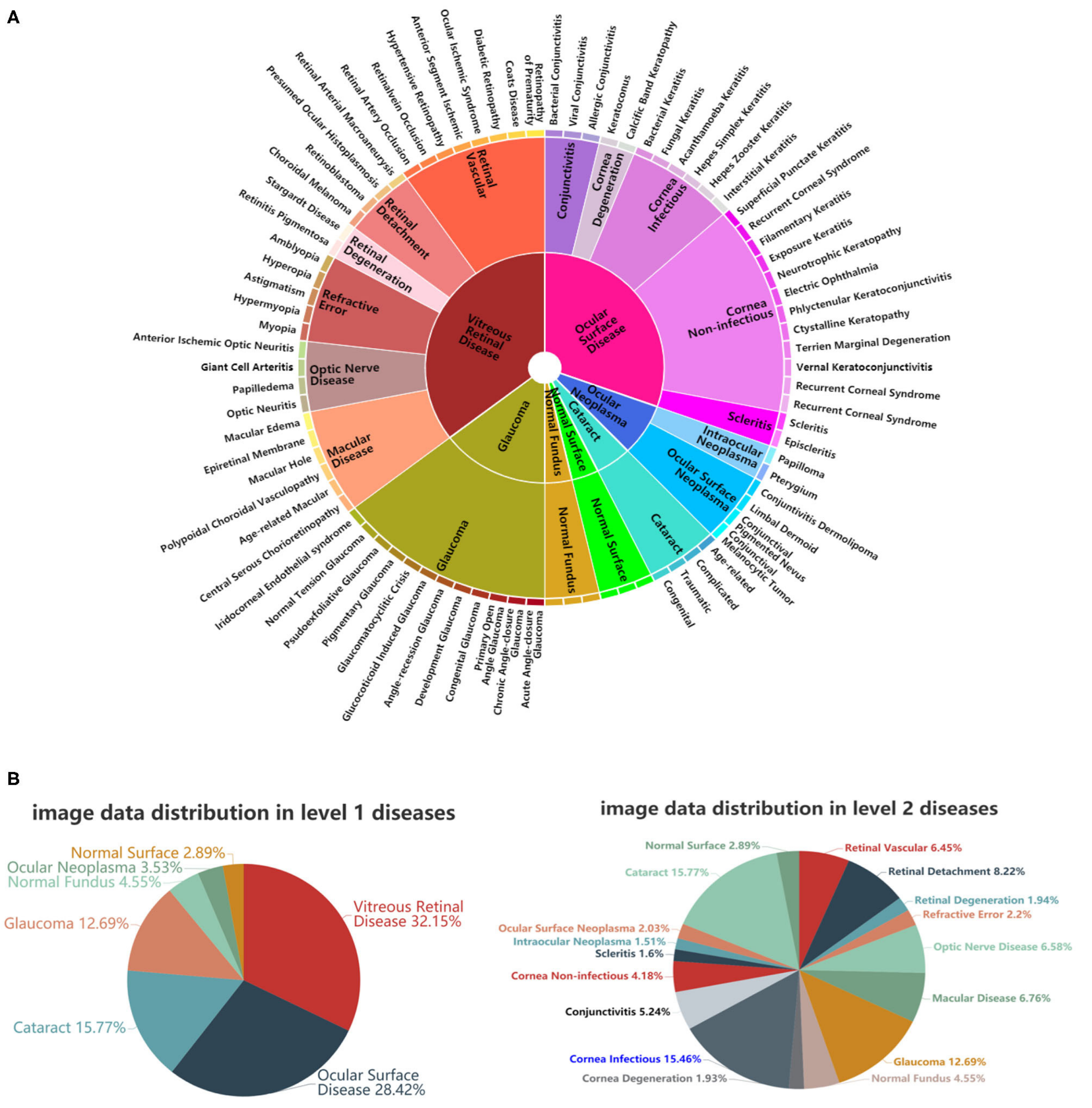


FIGURE 2 | A schematic illustration of the predefined eye disease taxonomy and example test set images. **(A)** Pie-structured eye disease taxonomy. **(B)** Data distribution for the first two levels of diseases.

the Affiliated Hospital of Guizhou Medical University, Guizhou. We used the IM 900 or 600 digital slit-lamp photography system (Haag-Streit, Switzerland) and CR-2 digital non-mydriatic retinal cameras (Canon, Japan). All images were annotated by senior ophthalmologists, where 50% of the proportion included retinal photographs and no images with the dilated pupil were included. Our objective was to provide a fast and cost-effective tool for screening patients with visual impairments. A suspected participant would be referred to a doctor for further assessment, including the dilated examination.

Retrospective Dataset

Thirty-two ophthalmologists were invited to grade the images of the retrospective database. During the training process of ophthalmologists, a dataset of 100 images (including 25 corneal disease cases, 25 cataract cases, 25 glaucoma cases, and 25 retinal

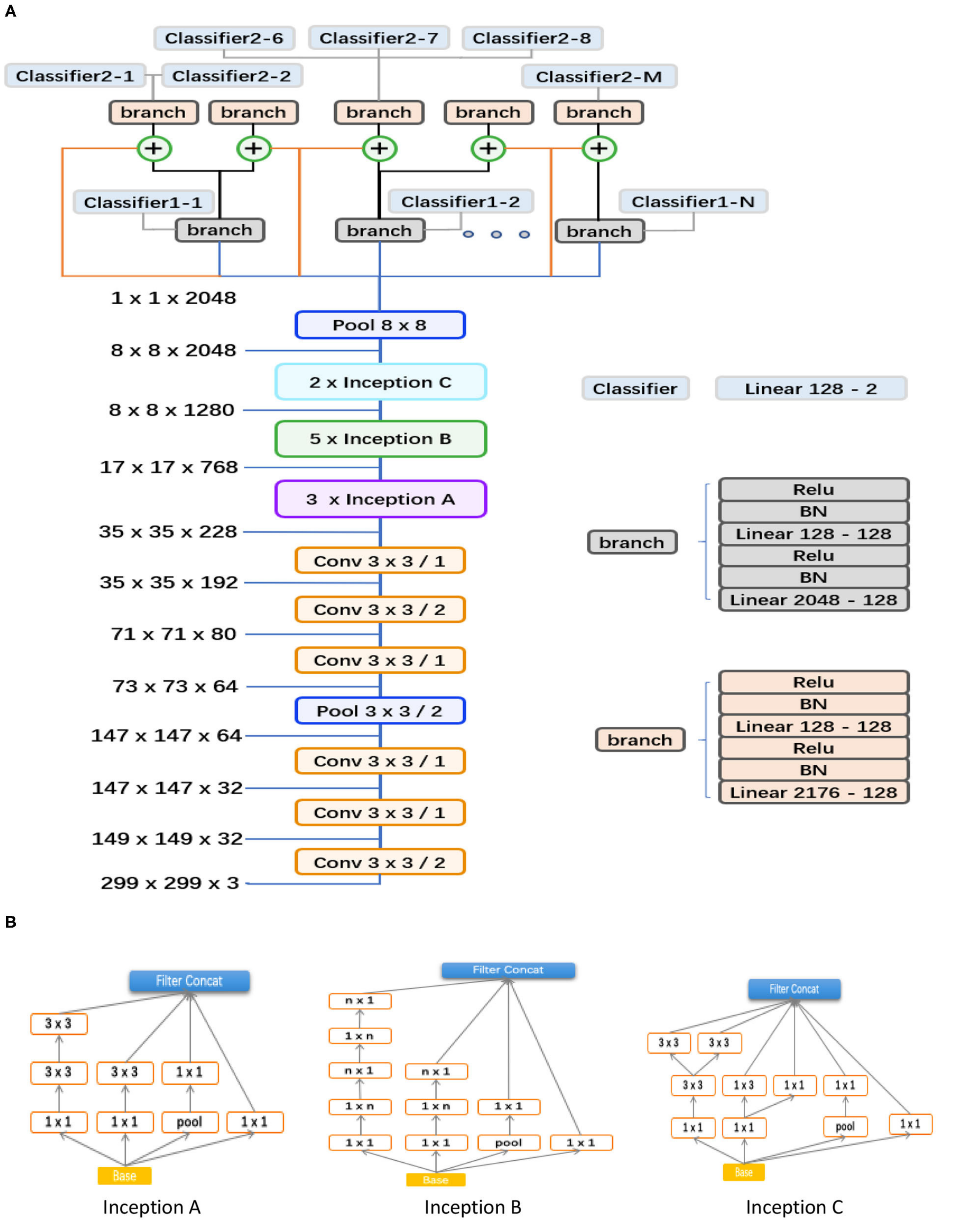


FIGURE 3 | Abstraction of the proposed hierarchical deep learning framework. **(A)** The proposed network architecture based on the feature network of Inception v3 (Conv 3 × 3/2 indicates that a 3 × 3 convolution kernel was used and stride = 2). The corresponding sizes of the input and output for each module are also shown. In *(Continued)*

FIGURE 3 | our framework, a family of multi-task & multi-label classification layers were used hierarchically to represent various levels of eye diseases. The individual multi-task classifier layer is defined on the basis of a predefined eye disease taxonomy. Here, the data flow in blue indicates that the backbone is directly connected to the branch of level 1; the orange means that the backbone is directly connected to the branch of level 2; the flow in black means connecting from the branch of level 1 to the branch of level 2; and the $\bigoplus$ is a feature concatenation operation, where features from the black and orange are superimposed; finally, this 8*8 pooling layer is a global average pooling, which turns the 8*8 feature map into a 1*1 feature map. **(B)** Different spatial factorized Inception modules are presented here. Inception A contains the factorization of the original 5 × 5 convolutions, factorizes general n × n convolutions (*n* = 5 in our study), and has expanded the filter bank outputs.

disease cases) was used for the test. The participants' results were compared with those of two senior corneal specialists (H.G. and J.H.). The participants would not complete the training until they achieved a κ-value of 0.75 or more. A κ-value of 0 indicates that observed agreement is the same as that expected by chance; 1 indicates perfect agreement; 0.75 or more indicates substantial agreement and/or almost perfect agreement. As a result, 20 ophthalmologists were qualified as graders to classify images. Each photograph was reviewed with the same standard and annotated via face-to-face communication between two ophthalmologists. As all 7,100 images from 1,600 patients collected already had original diagnoses recorded in medical charts, graders were asked to review, validate, and classify the images.

Prospective Dataset

A total of 4,670 outpatients agreed to receive the test and got their ocular surface slit-lamp photographs taken before their physician visits. Informed consent was obtained from all the participants. A software practitioner participating in this study fed these images as input to the trained deep learning software model. The algorithm generates a probability/confidence score over the classification nodes in a sequential manner, i.e., level by level. If the probability/confidence score of any disease subtype was greater than a predefined threshold, the disease subtype was diagnosed as positive. To quantitatively compare the sensitivity and specificity of our algorithm to that of the other 40 ophthalmologists on the diagnostic task of these cases, receiver operating characteristic (ROC) curves were plotted where each ophthalmologist was asked about the diagnosis on the basis of the images. Thirteen additional cases were also independently collected from clinics for our direct performance test sets.

To explore the visual characteristics of different clinical classes, we examined the internal image features learned by the proposed framework using t-distributed stochastic neighbor embedding (22). As demonstrated in **Figure 1A**, each point represents an eye image projected from the n-dimensional output of the last hidden layer of Inception-v3 backbone into two dimensions. We see clusters of points of the same clinical classes. This visualization represents the ability of our method to objectively separate normal patients from early cases of visual impairment diseases for a referral. **Figure 1B** shows a few examples of images that demonstrate the visual features using which the proposed hierarchical deep learning framework can identify and make a diagnosis.

Taxonomy

Inspired by Esteva et al. (7), who defined skin diseases in a tree structure, we adopted a similar approach to define our domain taxonomy structure for eye diseases, taking advantage of fine-grained information embedded within the images. Our taxonomy represented 100 individual diseases hierarchically arranged in a Pie structure. It was derived based on the collected retrospective database with 7,100 images from 1,600 patients by ophthalmologists using a bottom-up procedure: Individual diseases—initialized were defined as leaf nodes, and then were merged on the basis of clinical and visual similarity until the entire structure was connected.

As shown in **Figure 2A**, the taxonomy is useful in generating hierarchical training classes that are both well-suited for machine learning classifiers and medically relevant. In this study, the first two levels of the taxonomy were used in performance validation. **Figure 2B** illustrates the corresponding data distributions. It is worth mentioning that due to insufficient numbers of images for each of the level 3 diseases, we did not perform the level 3 classification. However, the extension to more levels can be implemented via our flexible and extensive framework with sufficient training data.

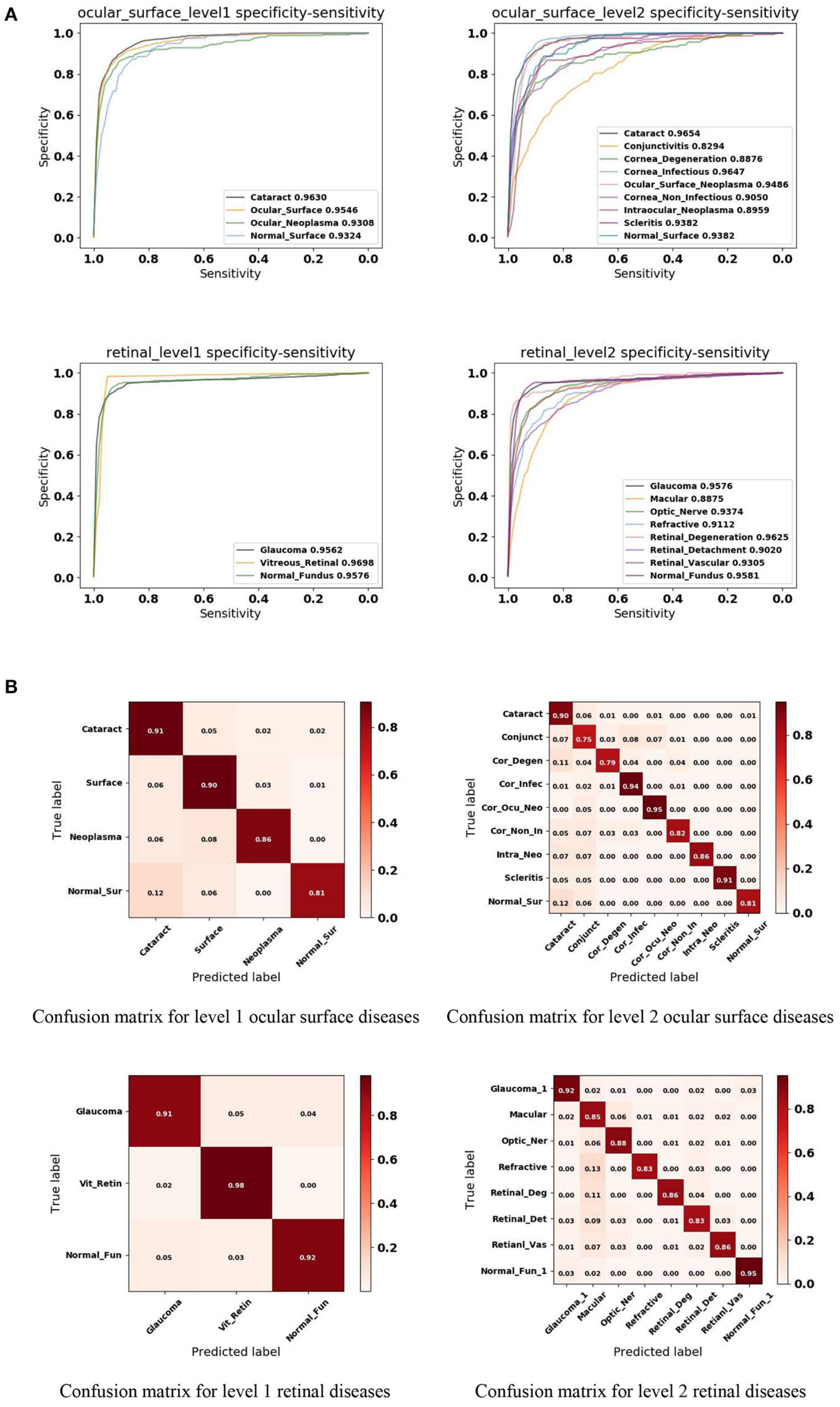
Proposed Hierarchical Deep Learning Framework

As shown in **Figure 3**, the proposed hierarchical deep learning framework is composed of a family of multi-task & multi-label learning classifiers representing different levels of eye disease classification derived from the hierarchical eye disease taxonomy. Here, we used an Inception-v3 CNN as the backbone of the proposed framework, and the final classification layer of the Inception-v3 network was replaced with our novel hierarchical multi-task & multi-label classification layers. Each task branch consists of several stacked fully connected units, hierarchically representing various levels of eye disease classification. As a result, the classification results of lower levels of classifiers can be used as priors for higher levels of classifiers, thereby improving the final classification performance.

We trained the model by minimizing our novel multi-level eye disease–guided loss function consisting of multiple levels of losses. The objective function for two levels can be represented as follows:

$$Loss_T = \alpha^* loss_{l1} + (1 - \alpha)^* loss_{l2} \tag{1}$$

where the term $Loss_T$ is the total loss of the final model, and $loss_{l1}$ and $loss_{l2}$ represent the corresponding losses for levels 1 and 2 of eye disease identification, respectively. α is a weight parameter that is used to control the balance between the two losses. For the two levels, $\alpha \in (0, 0.5)$, setting more weight for the higher level because the ultimate goal was to classify



Confusion matrix for level 1 ocular surface diseases

Confusion matrix for level 2 ocular surface diseases

Confusion matrix for level 1 retinal diseases

Confusion matrix for level 2 retinal diseases

FIGURE 4 | Performance of the proposed hierarchical deep learning framework. **(A)** The mean receiver operating characteristic (ROC) curve for various eye diseases of the first two levels of the eye disease taxonomy. AUC is the area under the ROC curve. **(B)** Confusion matrices for the first two levels of the eye disease taxonomy. Conjunct, Conjunctivitis; Cor_Degen, Corneal_Degeneration; Cor_Infec, Corneal_Infectious; Ocu_Cor_Neo, Ocular_Corneal_Neoplasma; Cor_Non_In, Corneal_Non_Infectious; Intra_Neo, Intraocular_Neoplasma; Normal_Sur, Normal_Surface; Optic_Ner, Optic_Ner; Retinal_Deg, Retinal_Degeneration; Retinal_Det, Retinal_Detachment; Retinal_Vas, Retinal_Vascular; Normal_Fun, Normal_Fundus.

higher levels of diseases. Through experiments, we found that $\alpha = 0.3$ performed well (i.e., the loss weight ratio 3:7 between level 1 and 2 classifiers). In this study, we used the sigmoid function for each class instead of the commonly used SoftMax function, for multiple diseases may simultaneously exist. Because of the unbalanced property of data, we applied the focal loss (23) for the loss function of each level, which reduced the impact of data imbalance and made the training focus on hard negatives as well. The focal loss function can be represented as follows:

$$FL\left(p_t\right) = -\left(1 - p_t\right)^{\gamma}\ log\left(p_t\right) \tag{2}$$

where

$$p_t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise \end{cases} \ (3) \tag{3}$$

$\left(1 - p_t\right)^{\gamma}$ is a modulating factor of the cross-entropy loss, with a tunable focusing parameter $\gamma \geq 0$, $p \in [0, 1]$. During the training process, various data augmentation methods (including horizontal and vertical flipping, color jitter, rotation, etc.) were also applied to all classes independently on-the-fly. It is worth mentioning that the online data augmentation was aimed at increasing the diversity of data for generalization rather than balancing and/or increasing the amount of training data.

Instead of training from scratch, we applied a fine-tuning strategy on a pretrained model using a multi-step retraining strategy. In this study, all images were resized to the size of 299 × 299 since that is the default input size for the Inception-v3 model. We used the Inception-v3 model pretrained on the ImageNet dataset (24) as the initial model and fine-tuned all layers with our dataset. First, the multi-task branches were trained by freezing the backbone's weights for 5 epochs. The Adam optimizer and a learning rate of 0.0001 and epsilon of 0.1 were used. Then, we performed a multi-step retraining strategy. In this strategy, we gradually unfroze the layer weights in steps, with the first few layers being unfrozen last. The learning rates were progressively reduced from 0.0001 to 0.000001, whereas other parameters were kept unchanged. Every step lasted 20 epochs. We used Facebook's PyTorch deep learning framework (25) to train, validate, and test the algorithm networks.

RESULTS

Performance Evaluation

Algorithm performance was measured by the area under the ROC curve (AUC) and the accuracy rate. The accuracy rate calculated the percentage of correctly predicted individuals among the whole test set, whereas the ROC curve was generated by plotting the curve of sensitivity against specificity, which can be defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative rates, respectively. TP and TN represent correctly predicted positives and negatives with respect to the ground truth labels. FP and FN represent incorrectly predicted positives and negatives with respect to the ground truth labels.

In this study, we applied a 5-fold cross-validation strategy to evaluate the effectiveness of the proposed framework. This strategy randomly divides the entire dataset into five subsets, each containing around 20% of the data. Model training and validation were performed five times. **Figure 4A** shows that our framework achieved high sensitivity, specificity, and AUC for most of the identified diseases. **Figure 4B** illustrates the corresponding confusion matrices for disease classification. As shown in level 1 confusion matrices, the CNN model performed extremely well on all three retinal fundus diseases, with an accuracy of 0.91 for glaucoma, 0.98 for vitreoretinal disease, and 0.92 for normal fundus. Meanwhile, the CNN model performed moderately well on all four ocular surface diseases, with an accuracy of 0.91 for cataract, 0.90 for surface disease, 0.90 for neoplasma, and 0.81 for normal surface images. This may be because fundus images contain more discriminative features than do ocular surface images. The model confused normal surface cases with cataract (12.0%) and confused cataract with

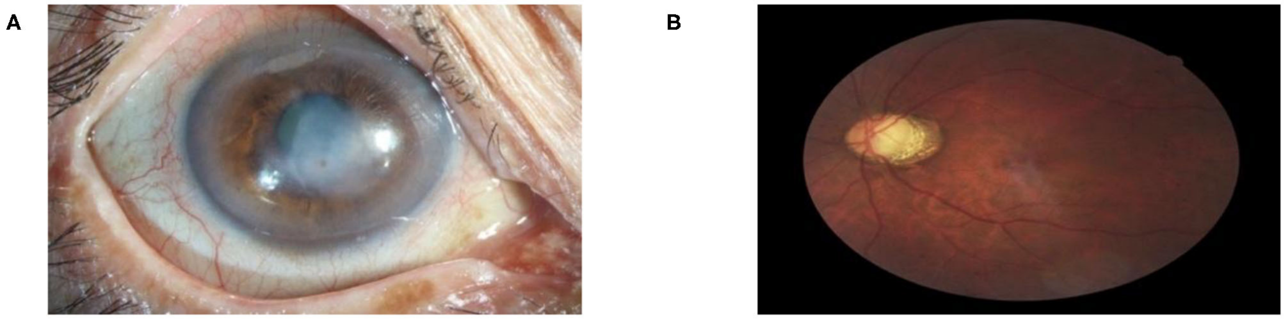
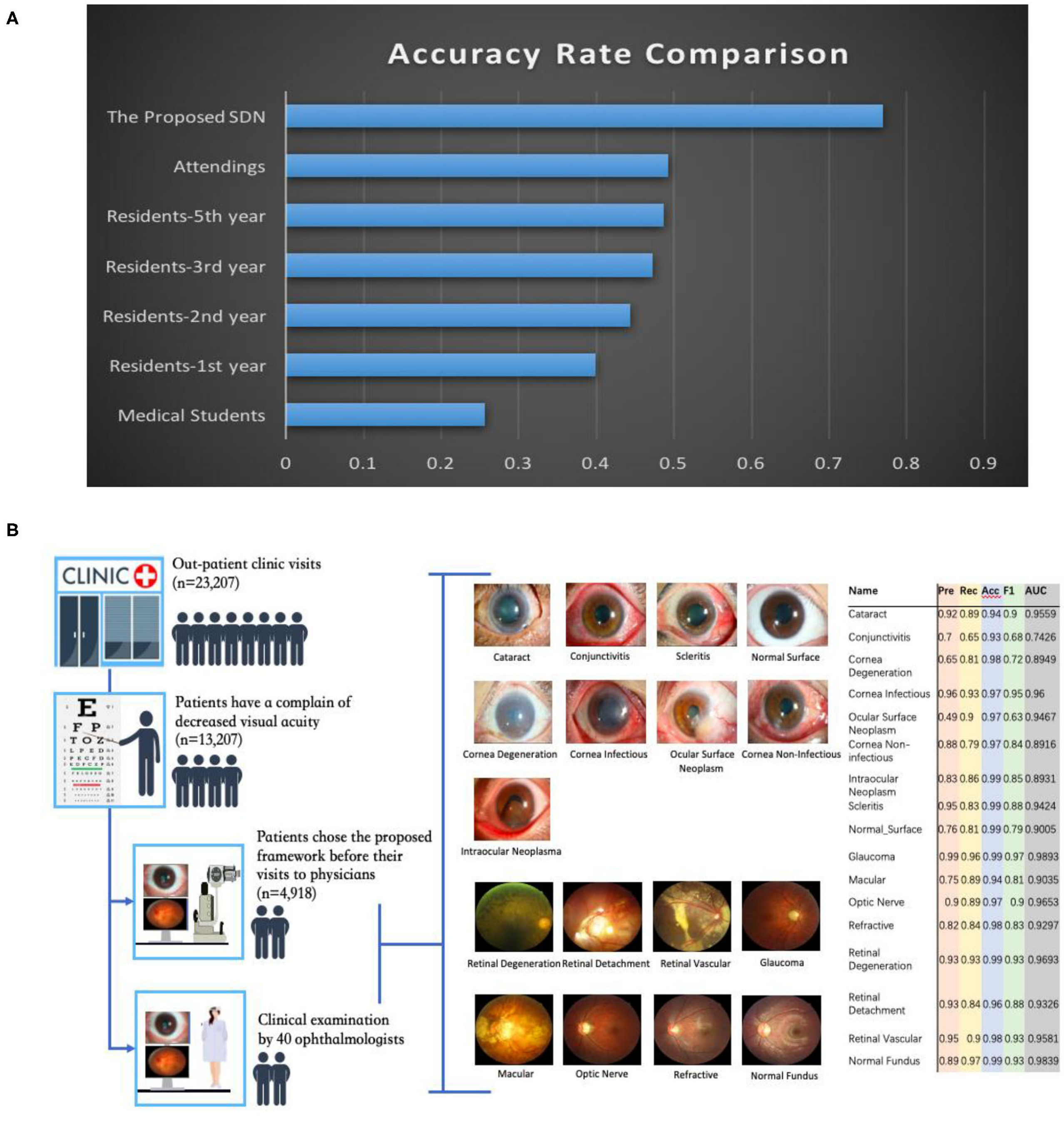


FIGURE 5 | Multi-label diagnostic results. The proposed hierarchical deep learning framework is capable of detecting multiple diseases simultaneously on the same patient: **(A)** cataract with 76.74% and corneal disease with 75.94% confidence and **(B)** glaucoma with 79.04% and retinopathy with 51.01% confidence.

surface disease (5.0%), neoplasm (2.0%), and normal surface images (2.0%). From these results, we can conclude that it is easy to confuse the normal surface with cataract because of appearance similarities, whereas cataract has more appearance diversity, which can also be confused with other ocular surface diseases and neoplasms. Similar results can be found in level 2 confusion matrices.

Because of the multi-task & multi-label property of the proposed framework, the trained model is capable of detecting multiple diseases simultaneously on the same patient, reflecting



Name	Pre	Rec	Acc	F1	AUC
Cataract	0.92	0.89	0.94	0.9	0.9559
Conjunctivitis	0.7	0.65	0.93	0.68	0.7426
Cornea Degeneration	0.65	0.81	0.98	0.72	0.8949
Cornea Infectious	0.96	0.93	0.97	0.95	0.96
Ocular Surface Neoplasm	0.49	0.9	0.97	0.63	0.9467
Cornea Non-infectious	0.88	0.79	0.97	0.84	0.8916
Intraocular Neoplasm	0.83	0.86	0.99	0.85	0.8931
Scleritis	0.95	0.83	0.99	0.88	0.9424
Normal_Surface	0.76	0.81	0.99	0.79	0.9005
Glaucoma	0.99	0.96	0.99	0.97	0.9893
Macular	0.75	0.89	0.94	0.81	0.9035
Optic Nerve	0.9	0.89	0.97	0.9	0.9653
Refractive	0.82	0.84	0.98	0.83	0.9297
Retinal Degeneration	0.93	0.93	0.99	0.93	0.9693
Retinal Detachment	0.93	0.84	0.96	0.88	0.9326
Retinal Vascular	0.95	0.9	0.98	0.93	0.9581
Normal Fundus	0.89	0.97	0.99	0.93	0.9839

FIGURE 6 | Eye disease classification performance of the proposed hierarchical deep learning framework and ophthalmologists. **(A)** The proposed hierarchical deep learning framework was tested against 40 board-certified ophthalmologists in diagnosing the clinical cases of 13 patients in a real-world setting. For each image, the ophthalmologists were asked to make three diagnoses. The proposed hierarchical deep learning framework outperformed all levels of board-certified ophthalmologists for all cases. **(B)** Clinical application of the proposed hierarchical deep learning framework for visual impairment diseases in a tertiary eye center. Discrepancies between manual grades and the proposed hierarchical deep learning framework results were sent to an independent panel of senior specialists for arbitration.

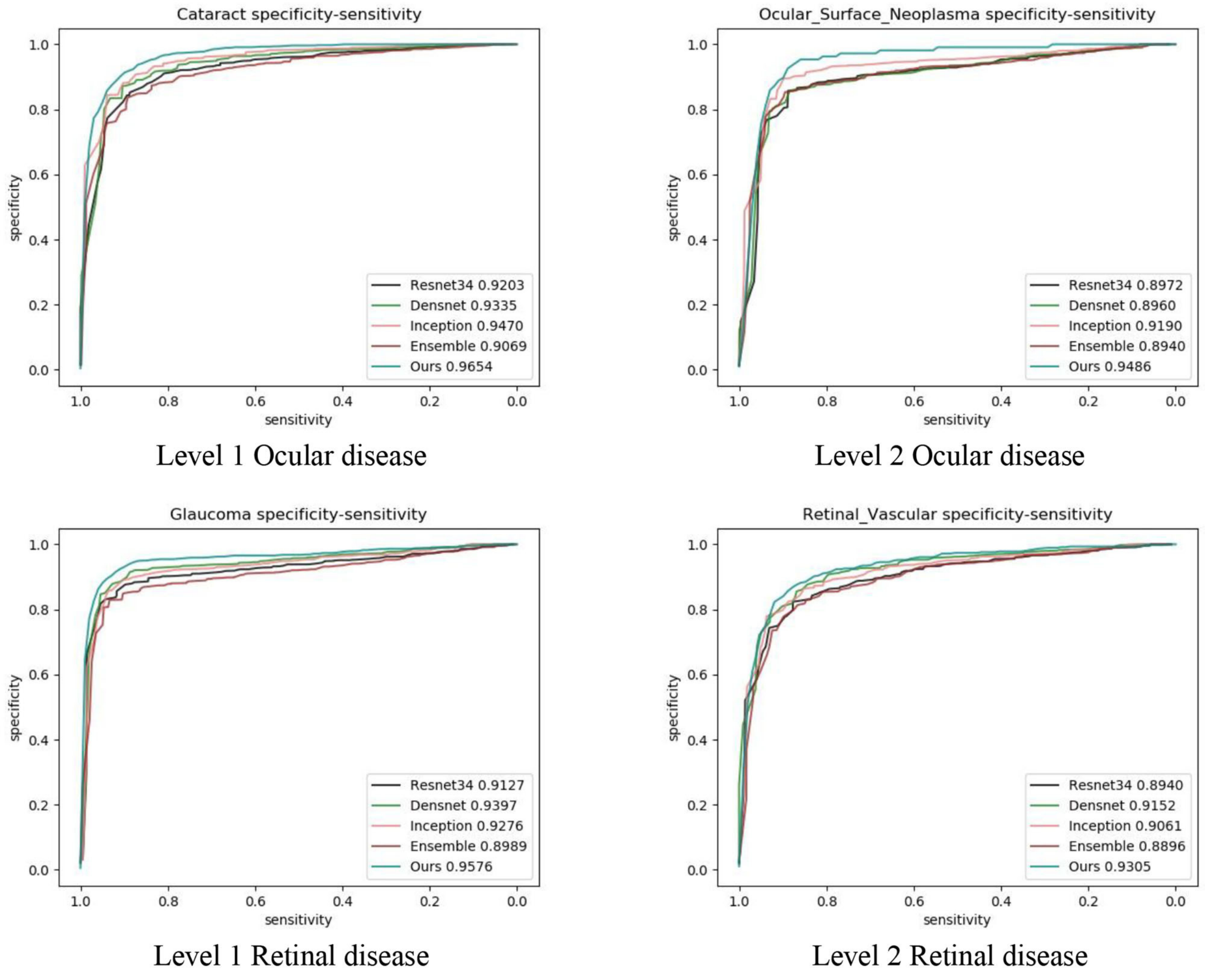
true clinical cases. As illustrated in **Figure 5**, both cataract and corneal disease were detected simultaneously within a single ocular surface image with 76.74 and 75.94% confidence, respectively. Similarly, both glaucoma and retinopathy were also detected within one retinal image with 79.04 and 51.01% confidence, respectively. It needs to be mentioned here that in this study, if the prediction score was > 50%, the system considered the screening output of the patient with the corresponding disease. In a real-world setting, if the screening output of the patient has one of the diseases listed above, the patient would be referred to a specialist for further diagnosis. Physicians need to consider not only the screening result but also the diagnostic severity of the disease to make clinical decisions for a patient. This was beyond the scope of our study. Our goal was to provide a fast and cost-effective screening tool for patients with visual impairment.

Comparison Tests

To both quantitatively and qualitatively demonstrate the effectiveness of the proposed framework, we also compared it with 40 board-certified ophthalmologists in diagnosing clinical cases. The comparison tests used 20 images from 13 patients.

TABLE 1 | Computational cost comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

Computational cost	**Ours**	**Inception-v3**	**ResNet34**	**DenseNet101**	**Ensemble**
Training (hours)	12.5	11.2	10.0	11.4	11.0
Inference (seconds)	0.097	0.083	0.069	0.075	0.106



Level 1 Ocular disease

Level 2 Ocular disease

Level 1 Retinal disease

Level 2 Retinal disease

FIGURE 7 | Performance comparison with four deep learning frameworks.

TABLE 2 | AUC comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	Inception-v3	ResNet34	DenseNet101	Ensemble
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.96**	0.94	0.92	0.93	0.91
Ocular surface (*n* = 2,018)	**0.95**	0.94	0.91	0.93	0.90
Ocular neoplasm (*n* = 251)	**0.93**	0.89	0.89	0.91	0.88
Normal surface (*n* = 205)	**0.93**	**0.93**	**0.93**	0.92	0.90
Weighted average	**0.95**	0.93	0.91	0.93	0.90
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.97**	0.95	0.92	0.93	0.91
Conjunctivitis (*n* = 372)	**0.83**	0.82	0.81	**0.83**	0.81
Cornea degeneration (*n* = 137)	**0.89**	0.86	0.85	**0.89**	0.83
Cornea infectious (*n* = 1,098)	**0.96**	0.95	0.93	0.94	0.91
Intraocular neoplasma (*n* = 107)	**0.95**	0.92	0.90	0.90	0.89
Cornea non-infectious (*n* = 297)	0.91	0.89	**0.93**	0.88	0.86
Ocular surface neoplasm (*n* = 144)	**0.90**	0.88	0.86	0.87	0.85
Scleritis (*n* = 114)	**0.94**	0.93	0.93	0.93	0.90
Normal surface (*n* = 205)	**0.94**	0.93	**0.94**	0.93	0.91
Weighted average	**0.94**	0.92	0.91	0.91	0.89
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.96**	0.94	0.91	0.92	0.90
Vitreoretinal disease (*n* = 2,283)	**0.97**	0.95	0.93	0.94	0.92
Normal fundus (*n* = 323)	**0.96**	**0.96**	0.94	0.94	0.92
Weighted average	**0.97**	0.95	0.93	0.94	0.91
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.96**	0.94	0.91	0.93	0.90
Macular disease (*n* = 480)	**0.89**	0.88	0.85	0.86	0.85
Optic nerve disease (*n* = 467)	**0.94**	**0.94**	0.90	0.91	0.89
Refractive error (*n* = 156)	**0.91**	0.90	0.89	0.89	0.89
Retinal degeneration (*n* = 138)	0.96	**0.97**	0.93	0.96	0.92
Retinal detachment (*n* = 584)	**0.90**	0.89	0.87	0.88	0.85
Retinal vascular disease (*n* = 458)	**0.93**	0.92	0.89	0.91	0.89
Normal fundus (*n* = 323)	0.96	**0.97**	0.93	0.94	0.92
Weighted average	**0.93**	0.92	0.89	0.91	0.88

Bold value means "Best performance".

The tested diseases include allergic conjunctivitis, dry eye, bacterial conjunctivitis, Mooren's corneal ulcer, keratoconus, fungal keratitis, viral keratitis, scleritis, age-related macular degeneration, cataract, primary angle closure glaucoma, myopia, diabetic retinopathy, and retinal detachment. For this study, each ophthalmologist was asked for the three most likely diagnoses of the patient. This choice of question reflects the actual in-clinic task in which ophthalmologists would decide whether or not to request further examinations. For a fair comparison, the proposed hierarchical deep learning framework also outputs the top three diagnoses with probability/confidence scores. The outcome was considered "correct" when one of the three diagnoses made by the proposed hierarchical deep learning framework or an ophthalmologist included the real diagnosis for the case. Remarkably, the proposed hierarchical deep learning framework outperformed all levels of board-certified ophthalmologists in every case, as shown in **Figure 6A** ($P < 0.05$ in t-test).

In addition, we performed an observational diagnostic assessment comparison between the proposed framework and human graders in a tertiary eye center to determine whether or not the proposed framework can be introduced into visual impairment disease screening. As demonstrated in **Figure 6B**, 4,670 consecutive patients visiting the Shanghai Eye and ENT Hospital were invited to get their slit-lamp photographs taken before they were checked by their physicians. Discrepancies between manual grades and the proposed hierarchical deep learning framework results were sent to a panel of senior

TABLE 3 | Accuracy comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	Inception-v3	ResNet34	DenseNet101	Ensemble
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.93**	0.92	0.9	0.91	0.89
Ocular surface (*n* = 2,018)	**0.92**	0.9	0.88	0.89	0.87
Ocular neoplasm (*n* = 251)	0.96	**0.97**	0.96	0.96	0.95
Normal surface (*n* = 205)	0.98	0.98	**0.99**	0.98	0.98
weighted average	**0.93**	0.92	0.90	0.91	0.89
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.94**	0.93	0.91	0.92	0.9
Conjunctivitis (*n* = 372)	0.93	0.93	0.93	**0.94**	0.92
Cornea degeneration (*n* = 137)	0.97	0.97	0.97	**0.98**	0.97
Cornea infectious (*n* = 1,098)	**0.97**	0.96	0.94	0.95	0.93
Intraocular neoplasma (*n* = 107)	**0.99**	0.98	0.98	0.98	0.98
Cornea non-infectious (*n* = 297)	0.97	**0.98**	**0.98**	0.97	0.97
Ocular surface neoplasm (*n* = 144)	0.98	**0.99**	0.98	0.98	0.98
Scleritis (*n* = 114)	**0.99**	0.98	0.98	0.98	0.07
Normal surface (*n* = 205)	0.98	0.98	0.98	0.98	**0.99**
Weighted average	**0.96**	0.95	0.94	0.95	**0.90**
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.96**	0.95	0.93	0.94	0.93
Vitreoretinal disease (*n* = 2,283)	**0.97**	0.96	0.93	0.95	0.92
Normal fundus (*n* = 323)	0.97	**0.98**	0.97	0.97	0.97
Weighted average	**0.97**	0.96	0.93	0.95	0.93
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.97**	0.96	0.94	0.94	0.93
Macular disease (*n* = 480)	**0.93**	0.92	0.91	0.91	0.9
Optic nerve disease (*n* = 467)	**0.96**	**0.96**	0.95	0.95	0.95
Refractive error (*n* = 156)	0.98	**0.99**	0.98	0.98	0.98
Retinal degeneration (*n* = 138)	**0.99**	**0.99**	0.98	0.98	0.98
Retinal detachment (*n* = 584)	**0.96**	0.95	0.94	0.94	0.93
Retinal vascular disease (*n* = 458)	**0.97**	0.96	0.96	0.96	0.96
Normal fundus (*n* = 323)	**0.99**	0.98	0.98	0.98	0.98
Weighted average	**0.97**	0.96	0.95	0.95	0.94

Bold value means "Best performance".

ophthalmologists for arbitration. Our data showed that the proposed hierarchical deep learning framework achieved an acceptable detection accuracy rate for visual impairment disease screening when compared with that of human graders in a clinical setting. The detection AUC of the proposed hierarchical deep learning framework for 17 subclasses in level 2 of visual impairment diseases ranged from 0.743 to 0.989.

We also compared our algorithm performance with four previously reported methods, namely Inception-v3 (8), ResNet (26), DenseNet (27), and Ensemble (28). The Ensemble model combined all backbone features extracted from the other three models and applied a tree-based classifier for the final classification. To have a fair comparison, all the networks above were also trained as multi-task & multi-label networks but without the proposed hierarchical architecture. To be more specific, the last layers of these networks were replaced with a set of binary classifiers with a flat architecture for each level of the disease classification. As shown in **Table 1**, the computational costs for both the training and the inference stage were comparable for all models. However, with the proposed hierarchical architecture, our algorithm outperformed all four existing methods in most of the diseases. For example, as shown in **Figure 7**, for level 1 disease identification—such as glaucoma—our framework achieved AUC 0.958, whereas ResNet, DenseNet, Inception-v3, and Ensemble methods achieved AUC 0.913, 0.940, 0.928, and 0.899, respectively. Similarly, for level 2 disease identification, such as ocular surface neoplasm, our framework achieved

TABLE 4 | Recall comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	**Inception-v3**	**ResNet34**	**DenseNet101**	**Ensemble**
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.91**	0.9	0.88	0.89	0.86
Ocular surface (*n* = 2,018)	**0.9**	0.89	0.86	0.88	0.85
Ocular neoplasm (*n* = 251)	**0.86**	0.82	0.8	0.84	0.8
Normal surface (*n* = 205)	**0.81**	0.75	0.8	0.75	0.8
Weighted average	**0.90**	0.88	0.86	0.87	0.85
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.9**	0.89	0.86	0.88	0.85
Conjunctivitis (*n* = 372)	**0.75**	0.73	0.73	0.74	0.73
Cornea degeneration (*n* = 137)	0.78	0.78	0.78	**0.79**	0.77
Cornea infectious (*n* = 1,098)	**0.94**	0.92	0.89	0.9	0.86
Intraocular neoplasma (*n* = 107)	**0.95**	0.9	0.82	0.82	0.86
Cornea non-infectious (*n* = 297)	0.78	0.8	**0.82**	0.8	0.76
Ocular surface neoplasm (*n* = 144)	**0.86**	0.79	0.73	0.76	0.76
Scleritis (*n* = 114)	**0.91**	0.87	0.87	0.87	0.87
Normal surface (*n* = 205)	0.81	0.8	0.81	0.8	**0.82**
Weighted average	**0.88**	0.86	0.84	0.85	0.83
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.91**	0.89	0.87	0.88	0.86
Vitreoretinal disease (*n* = 2,283)	**0.98**	0.97	0.95	0.96	0.95
Normal fundus (*n* = 323)	0.91	**0.92**	0.88	0.89	0.86
Weighted average	**0.96**	0.94	0.92	0.93	0.92
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.92**	0.9	0.87	0.88	0.84
Macular disease (*n* = 480)	**0.85**	0.82	0.8	0.79	0.79
Optic nerve disease (*n* = 467)	0.86	**0.88**	0.84	0.84	0.84
Refractive error (*n* = 156)	**0.83**	0.81	0.8	0.77	0.81
Retinal degeneration (*n* = 138)	0.82	**0.86**	0.79	0.82	0.79
Retinal detachment (*n* = 584)	**0.83**	0.79	0.77	0.78	0.75
Retinal vascular disease (*n* = 458)	**0.86**	0.84	0.82	0.84	0.8
Normal fundus (*n* = 323)	0.89	**0.91**	0.88	0.89	0.86
Weighted average	**0.87**	0.86	0.83	0.83	0.81

Bold value means "Best performance".

AUC 0.949, whereas ResNet, DenseNet, Inception-v3, and Ensemble methods achieved AUC 0.897, 0.896, 0.919, and 0.894, respectively. More detailed comparison results can be found in **Tables 2–5**.

Saliency Maps

To show the interpretation of the proposed framework, we also created heatmaps via the gradient-weighted class activation mapping (Grad-CAM) algorithm (29), which can produce visual explanations for CNN-based deep learning models. Grad-CAM uses the gradient information flowing into the last convolutional layer to understand the importance of each neuron for a decision of interest, thereby highlighting the important regions in the image for prediction. It first computes the gradient of the score for a given class with respect to feature maps of a convolutional layer. Then, these gradients are average-pooled to obtain the neuron importance weights. Finally, the coarse heatmap for a given class is generated via a weighted combination of forward activation maps followed by a ReLU function. As illustrated in **Figure 8**, the generated heatmaps helped indicate the potential corneal lesion regions for further examination, thereby establishing prediction trust and interpretation for physicians.

DISCUSSION

In this study, we demonstrated the effectiveness of the proposed hierarchical deep learning framework in

TABLE 5 | Precision comparison between the proposed hierarchical deep learning framework and existing deep learning frameworks.

	Ours	**Inception-v3**	**ResNet34**	**DenseNet101**	**Ensemble**
Level 1 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.88**	0.85	0.81	0.84	0.8
Ocular surface (*n* = 2,018)	**0.96**	0.94	0.93	0.93	0.92
Ocular neoplasm (*n* = 251)	**0.7**	0.68	0.67	**0.7**	0.63
Normal surface (*n* = 205)	0.59	**0.75**	0.71	**0.75**	0.71
Weighted average	**0.90**	0.88	0.86	0.88	0.85
Level 2 anterior segment (*n* = no. of images)					
Cataract (*n* = 1,120)	**0.91**	0.89	0.86	0.87	0.85
Conjunctivitis (*n* = 372)	**0.67**	0.66	0.63	0.65	0.61
Cornea degeneration (*n* = 137)	0.64	0.64	0.6	**0.7**	0.61
Cornea infectious (*n* = 1,098)	**0.95**	0.94	0.92	0.93	0.92
Intraocular neoplasma (*n* = 107)	**0.68**	0.62	0.62	0.62	0.58
Cornea non-infectious (*n* = 297)	0.88	0.9	**0.91**	0.89	0.87
Ocular surface neoplasm (*n* = 144)	**0.99**	0.98	0.98	**0.99**	0.97
Scleritis (*n* = 114)	**0.98**	0.97	0.96	**0.98**	0.95
Normal surface (*n* = 205)	0.87	0.92	**0.93**	0.92	**0.93**
Weighted average	**0.88**	0.87	0.85	0.86	0.84
Level 1 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.94**	0.91	0.86	0.88	0.86
Vitreoretinal disease (*n* = 2,283)	**0.98**	0.96	0.95	0.96	0.94
Normal fundus (*n* = 323)	0.91	0.87	0.93	0.91	**0.95**
Weighted average	**0.96**	0.94	0.93	0.93	0.92
Level 2 retinal disease (*n* = no. of images)					
Glaucoma (*n* = 901)	**0.95**	0.93	0.89	0.9	0.88
Macular disease (*n* = 480)	**0.7**	0.66	0.62	0.64	0.62
Optic nerve disease (*n* = 467)	0.82	**0.85**	0.8	0.81	0.79
Refractive error (*n* = 156)	0.96	0.96	**0.99**	0.96	0.96
Retinal degeneration (*n* = 138)	0.82	**0.86**	0.79	0.79	0.76
Retinal detachment (*n* = 584)	**0.9**	0.87	0.84	0.87	0.81
Retinal vascular disease (*n* = 458)	**0.93**	0.89	0.89	0.89	0.86
Normal fundus (*n* = 323)	0.91	0.92	**0.93**	0.92	0.92
Weighted average	**0.88**	0.86	0.84	0.85	0.82

Bold value means "Best performance".

identifying most causes of visual impairment diseases worldwide. Training the proposed hierarchical deep learning framework on eye images captured using commonly available equipment, we outperformed the performance of 40 board-certified ophthalmologists on 13 clinical cases. Further assessment of 4,670 cases in a tertiary eye center also demonstrated that the proposed framework achieved a high identification accuracy rate for different visual impairment diseases compared with that of human graders in a clinical setting.

Although we acknowledge that the clinical impression and diagnosis by an ophthalmologist are based on contextual factors beyond the visual inspection of the eye, the ability to classify eye images with the accuracy of a board-certified ophthalmologist has the potential to profoundly expand access to vital medical care. It has the potential to aid the delivery of eye disease screening in developed and developing countries in a manner that is inexpensive, efficient, and easily accessible. It can also be used to provide eye care guiding services in communities and assist doctors in diagnosing visual impairment diseases.

To validate this technique across the full distribution and spectrum of visual impairment diseases encountered in a clinical setting, further research is necessary to evaluate performance in a large community screening setting. This method is primarily constrained by data and can be validated for more visual conditions if sufficient training examples are provided.

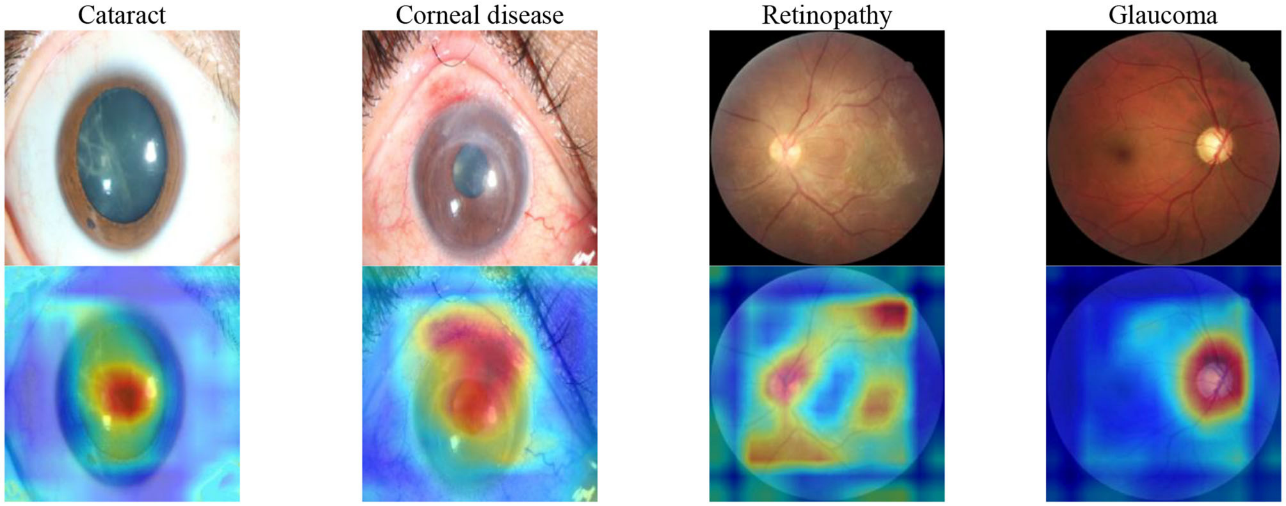


FIGURE 8 | Saliency maps for images with various common visual impairment diseases. These visualizations are generated automatically, locating regions for closer examination after a patient is seen by a consultant ophthalmologist. The bluer the color, the lower the attention of the model; the redder the color, the higher the attention of the model. Visualization maps are generated from deep learning features.

In this study, we applied multiple train–test splits via a 5-fold cross-validation where we randomly divided the entire image dataset into five subsets. Splitting data with respect to patients instead of images is indeed a better strategy; however, the dataset we had did not contain user identification information after data anonymization. We added this as a limitation of our study and would maybe explore it as future work. We would also conduct further experiments with publicly available datasets (such as EyePACS; Kaggle) as one of the future works. In the future, it may also be important to investigate different types of common patient metadata, such as genetic factors, patient history, and other clinical data that may influence a patient's risk of visual impairment diseases. Adding this information to the classification model may yield insightful information outside of strictly imaging information, potentially enhancing the diagnostic accuracy.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of the Shanghai Eye and EENT Hospital (EENTIRB20170607). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XL and JH: conception and design. XL, JH, and JC: administrative support. JH, LG, JX, YL, and XS: provision of study materials or patients. JH, XL, HG, ZY, and JL: collection and assembly of data. YG: data analysis and interpretation. All authors: manuscript writing and final approval of manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (81970766 and 8217040684), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Shanghai Innovation Development Program (2020-RGZN-02033), the Shanghai Key Clinical Research Program (SHDC2020CR3052B); LG was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), EXC 2026, Cardio-Pulmonary Institute (CPI), Project ID 390649896; and Guizhou Science and Technology Program (GZWJKJ2018-1-003).

ACKNOWLEDGMENTS

We thank the residents and faculties of the Department of Ophthalmology at the Shanghai Eye and ENT Hospital and the Affiliated Hospital of Guizhou Medical University for participating in our tests. We also thank Drs. Xiaobo Yu, Rui Jiang, Jingyi Cheng, and Lijia Tian from Fudan University for their help with project coordination. We thank the support from China National GeneBank.

REFERENCES

1. Bourne RRA, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *Lancet Glob Health*. (2017) 5:e888–97. doi: 10.1016/S2214-109X(17)30293-0
2. Xu L, Wang Y, Li Y, Wang Y, Cui T, Li J, et al. Causes of blindness and visual impairment in urban and rural areas in Beijing: the Beijing eye study. *Ophthalmology*. (2006) 113:1134.e1-11. doi: 10.1016/j.ophtha.2006.01.035
3. Zhao J, Xu X, Ellwein LB, Cai N, Guan H, He M, et al. Prevalence of vision impairment in older adults in rural china in 2014 and comparisons with the 2006 china nine-province survey. *Am J Ophthalmol*. (2018) 185:81–93. doi: 10.1016/j.ajo.2017.10.016
4. Rosenblatt TR, Vail D, Saroj N, Boucher N, Moshfeghi DM, Moshfeghi AA. Increasing incidence and prevalence of common retinal diseases in retina practices across the United States. *Ophthalmic Surg Lasers Imaging Retina*. (2021) 52:29–36. doi: 10.3928/23258160-20201223-06
5. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
6. Kermany D S, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. (2018) 172:1122–31.e9. doi: 10.1016/j.cell.2018.02.010
7. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056
8. Szegedy C, Vanhoucke V, Loffe S, Shlens J,Wojna Z. Rethinking the inception architecture for computer vision, 2016. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV (2016).
9. Menegola A, Fornaciali M, Pires R, Avila S, Valle E, et al. Towards automated melanoma screening: exploring transfer learning schemes. *arXiv[Prerpint].arXiv:1609.01228*. (2016). Available online at: https://www.researchgate.net/publication/307636270_Towards_Automated_Melanoma_Screening_Exploring_Transfer_Learning_Schemes
10. Setio AA, Ciompi F, Litjens G, Gerke P, Jacobs C, van Riel SJ, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Trans Med Imaging*. (2016) 35:1160–9. doi: 10.1109/TMI.2016.2536809
11. Nie D, Zhang H, Adeli E, Liu L, Shen D. 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. *Med Image Comput Comput Assist Interv*. (2016) 9901:212–20. doi: 10.1007/978-3-319-46723-8_25
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. (2016) 316:2402–10. doi: 10.1001/jama.2016.17216
13. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*. (2017) 124:962–9. doi: 10.1016/j.ophtha.2017.02.008
14. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. (2018) 125:1199–206. doi: 10.1016/j.ophtha.2018.01.023
15. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol*. (2017) 135:11706. doi: 10.1001/jamaophthalmol.2017.3782
16. Schlegl T, Waldstein SM, Bogunovic H, Endstraßer F, Sadeghipour A, Philip AM, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology*. (2018) 125:549–58. doi: 10.1016/j.ophtha.2017.10.031
17. Treder M, Lauermann JL, Eter N. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. *Graefes Arch Clin Exp Ophthalmol*. (2018) 256:259–65. doi: 10.1007/s00417-017-3850-3
18. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat Biomed Eng*. (2017) 1:24. doi: 10.1038/s41551-016-0024
19. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic population with diabetes. *JAMA*. (2017) 318:2211–23. doi: 10.1001/jama.2017.18152
20. Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. (2018) 24:1342–50. doi: 10.1038/s41591-018-0107-6
21. Li W, Yang Y, Zhang K, Long E, He L, Zhang L, et al. Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders. *Nature Bio Eng*. (2020) 4:767–77. doi: 10.1038/s41551-020-0577-y
22. Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res*. (2008) 9:2579–605. Available online at: https://search.ebscohost.com/login.aspx?direct=true&db=asr&AN=36099312&lang=zh-cn&site=ehost-live
23. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. (2020) 42:318–27. doi: 10.1109/TPAMI.2018.2858826
24. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. *IJCV*. (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y
25. Paszke A, Gross S, Chintala S, Chanan G, Yang E, Devito Z, et al. Automatic differeçntiation in PyTorch. *NIPS*. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach (2017). Available online at: https://openreview.net/forum?id=BJJsrmfCZ
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV (2016). p. 770–8. doi: 10.1109/CVPR.2016.90
27. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI (2017). doi: 10.1109/CVPR.2017.243
28. van Veen HJ, Nguyen L, Dat T, Segnini A. *Kaggle Ensembling Guide*. (2015). Available online at: https://mlwave.com/kaggle-ensembling-guide/ (accessed February 6, 2018).
29. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *International Conference on Computer Vision (ICCV)*. Venice (2017). p. 618–26. doi: 10.1109/ICCV.2017.74

Conflict of Interest: BP and JC are employed by the company Complete Genomics Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.