Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Article

Deep Sequential Feature Learning in Clinical Image Classification of Infectious Keratitis

Yesheng Xu [a,#], Ming Kong [b,#], Wenjia Xie [a,#], Runping Duan [a], Zhengqing Fang [b], Yuxiao Lin [b], Qiang Zhu [b], Siliang Tang [b], Fei Wu [b,*], Yu-Feng Yao [a,*]

[a] Department of Ophthalmology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou 310016, China
[b] College of Computer Science and Technology, Zhejiang University, Hangzhou 31002, China

ARTICLE INFO

Article history:
Received 17 January 2020
Revised 18 March 2020
Accepted 20 April 2020
Available online 15 July 2020

Keywords:
Deep learning
Corneal disease
Sequential features
Machine learning
Long short-term memory

ABSTRACT

Infectious keratitis is the most common condition of corneal diseases in which a pathogen grows in the cornea leading to inflammation and destruction of the corneal tissues. Infectious keratitis is a medical emergency for which a rapid and accurate diagnosis is needed to ensure prompt and precise treatment to halt the disease progression and to limit the extent of corneal damage; otherwise, it may develop a sight-threatening and even eye-globe-threatening condition. In this paper, we propose a sequential-level deep model to effectively discriminate infectious corneal disease via the classification of clinical images. In this approach, we devise an appropriate mechanism to preserve the spatial structures of clinical images and disentangle the informative features for clinical image classification of infectious keratitis. In a comparison, the performance of the proposed sequential-level deep model achieved 80% diagnostic accuracy, far better than the 49.27% ± 11.5% diagnostic accuracy achieved by 421 ophthalmologists over 120 test images.

1. Introduction

Traditionally, triage and diagnosis of diseases are carried out by physicians through observation based upon experience and knowledge constructed by individuals. In recent years, deep learning algorithms using deep convolutional neural networks have been tested for medical imaging interpretation with significant advances. The application of algorithms for triage and diagnosis of diseases has been mainly tested in fields that widely apply medical imaging technologies, including computerized tomography, magnetic resonance imaging (MRI), fundus photography, optical coherence tomography (OCT), and pathologic images [1]. This is because medical imaging technology exports naturally rich image data, and commercialized medical imaging technologies create standardized and consistent medical images that can be collected in a short period of time in a single institution or from multiple medical centers.

The diagnosis for many clinical diseases does not need commercialized medical imaging technologies, with which imaging recording is not routinely carried out in medical practice in many medical institutions; therefore, the collection of a large amount of image data will be dependent on historical accumulation sporadically dispersed in different medical centers. However, the development of machine learning diagnostic systems for such diseases has at least equal importance. A study classifying skin lesions [2], offering malignant or benign judgment, is a pioneer attempt in the field of non-conventional medical imaging technologies. Corneal diseases may also be broadly classified in this category. Corneal diseases are a major cause of blindness worldwide [3,4]. There are an estimated 4.5 million individuals worldwide who suffer from moderate to severe vision impairment due to the loss of corneal clarity after contracting corneal diseases [4]. Infectious keratitis is the most common cause of corneal diseases [5]. The normal cornea possesses a unique characteristic of transparency. The most distinct feature of infectious keratitis is the pathogen growth in the cornea leading to focal mass cloudiness and the cornea roughness, inevitably bringing out the unique characteristics of each pathogenic microorganism for its growth in the tissue [6]. The diagnosis of infectious keratitis mostly depends on discriminatively identifying the visual features of the infectious lesion in the cornea by an ophthalmologist. Clinically, ophthalmologists routinely depend on slit lamp microscopes to

[*] Corresponding authors.
E-mail addresses: wufei@zju.edu.cn (F. Wu), yaoyf@zju.edu.cn (Y.-F. Yao).
[#] These authors contributed equally to this article.

https://doi.org/10.1016/j.eng.2020.04.012
2095-8099/

observe the normality or abnormality of the cornea and beyond. Apart from being an observational tool, the slit lamp microscope can also be used to take a photograph and record the existing status of the corneal manifestations for each patient simultaneously, contributing to the development of a well-annotated dataset for artificial intelligence (AI)-based infectious keratitis recognition and analysis.

Since 1998, we have developed a large, well-annotated slit lamp microscopic image dataset of 115 408 images in total from 10 609 corneal disease patients. The collected dataset enabled us to devise a deep learning based method to perform infectious keratitis diagnosis in an end-to-end manner. To intuitively mimic the way in which ophthalmologists diagnose infectious keratitis, we proposed a feature learning mechanism to identify the informative visual patterns via sequential-level feature learning, which means the sampled patches from the center to the edge of the infectious lesion area in the clinical picture are grouped into a sequential-ordered set (SOS) and fed into a neural network for feature learning. We argue that the proposed sequential-level feature learning mechanism can utilize the spatial relationship among patches from the infectious lesion area and can disentangle exploratory factors of variations underlying the data sample. In addition, it provides a potential strategy to achieve more reliable, effective, and accurate diagnosis.

Our model was evaluated using the dataset and achieved an accuracy of correct diagnosis higher than that of 400 ophthalmologists.

2. Related works

2.1. Medical data mining

Over the years, electronic medical records (EMRs) have accumulated large quantities of medical data, which has enabled researchers to discover underlying knowledge. Data mining methods have been widely used on medical data to discover hidden knowledge and to use the extracted knowledge to aid in the prediction, diagnosis, and treatment of various harmful diseases.

Disease prediction is significant in preventing the occurrence of disease and reducing harm. Yang et al. [7] used patients' health records to forecast potential diabetes complications as well as discover the underlying association between complications and laboratory test types. He et al. [8] predicted lung cancer postoperative complications using the EMR dataset and extracted crucial variables from the dataset simultaneously.

EMR with predicted diagnostic labels and medication information can help an automatic assistant to predict the disease diagnosis and provide a rapid diagnostic reference for doctors. Nee et al. [9] used a large EMR text dataset to model the context of EMR of each disease and performed an accurate disease diagnosis prediction in EMR. Wright et al. [10] used data mining methods to obtain useful relations and rule sets from the medical datasets to predict which medication is prescribed next.

2.2. Traditional shallow models in medical image application

The traditional method uses hand-craft features (in general shallow models) for medical image classification and segmentation. Scott et al. [11] used gradient orientation, corner, and edge strength to detect vertebrae in dual energy X-ray images in 2003. Region splitting and merging is a well-known technique in the region-based approach. Manousakas et al. [12] applied the splitting and merging technique in an attempt to overcome the difficulties encountered when using homogeneity measures on MRI. Zhao et al. [13] introduced basic mathematical morphological theory and operations, and they proposed that the novel mathematical morphological edge detection could distinguish the edge of lungs in computed tomography (CT) images with salt-and-pepper noise. The experimental result shows that the method proposed was more efficient for both medical image de-noising and edge detection than the best edge detection method in 2006. Kaus et al. [14] used K-means clustering to automatically perform segmentation of the left ventricle in cardiac MRI. Cordes et al. [15] had performed research by using hierarchical clustering to measure connectivity in functional MRI. This method could detect similarities of low-frequency fluctuations, and the results indicated that the patterns of functional connectivity can be obtained with hierarchical clustering that resembles known neuronal connections. In 2006, Pohl et al. [16] presented a method of embedding signed distance maps into the linear log odds space, which could solve the modeling problems. Although these methods focused on regions, edges, and clustering, they have limited performance on real-world data [17].

2.3. Deep learning methods in medical image application

In computer-aided diagnosis, deep learning is now widely used for medical image recognition [18,19]. The basic structure of deep learning is the convolutional neural network (CNN), which has three types of layers, namely convolution, pooling, and total connection. To develop a robust AI algorithm based on CNN, we usually require a large amount of annotated data.

The standardized collection of medical images is not as easy as collecting general natural images. However, nowadays, several public medical image databases and multicenter collections of data can help solve the problem. Some types of medical image data such as X-rays, CT, electrocardiographs, and pathology images can be collected in large quantities. By using these big data, CNN-based AI algorithms can perform anatomical structure segmentation on CT images [20], classify normal or abnormal findings of chest radiographs [21], perform screening for lung or breast cancer [22,23], detect critical findings in head CT scans [24], classify liver lesions using a generative adversarial network (GAN)-based model [25], perform screening for heart conditions [26,27], and detect lymph node metastases in pathology images [27,28].

In the field of ophthalmology, due to the easy collection of images from fundus photography and OCT, the major area in which CNN-based AI algorithms have been applied is detecting retinal diseases, such as diabetic retinopathy, age-related macular degeneration, and glaucoma [29–31].

Currently, AI-assisted medical diagnostic systems are mainly applied in the field of medical imaging. The diagnosis of disease, which relies on the use of natural observation, mainly depends on the personal experience of the doctor. One example is for skin lesions; the current AI algorithm can differentiate malignant melanoma from benign lesions on digital skin photographs [2]. Corneal disease is another example, where ophthalmologists may use a slit lamp microscope to obtain the right diagnosis. Thus far, there is no research that has utilized AI to improve diagnostic accuracy for corneal disease.

3. Methods

3.1. Image datasets

Upon an institutional review board approval, the image dataset for this study included 115 408 clinical digital images taken from 10 609 patients with 89 categories of corneal diseases by slit lamp microscopy during the time period of May of 1998 to 2018 in the Department of Ophthalmology, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University. The clinical images were taken by two types of slit lamp microscopes, that is, Zeiss slit lamp microscope SL 130 (Carl Zeiss Meditec AG, Germany), integrated with the SL Cam for imaging module, providing each image with a resolution of 1024 × 768 pixels; and Topcon slit lamp microscope (TOPCON Corporation, Japan), affiliated with digital camera Unit DC-1 offering an image resolution of 1740 × 1536 pixels or 2048 × 1536 pixels.

In the dataset, images taken from patients with corneal infection at the active stage, including bacterial keratitis (BK), fungal keratitis (FK), and herpes simplex virus stromal keratitis (HSK), were selected for the training or testing set for algorithmic classification into each infectious category. All the images from the patients with corneal infections were annotated with a definite clinical diagnosis that was corroborated by at least two pieces of the following evidence: ① the clinical manifestations of the corneal infection as shown in Fig. 1(a); ② the progression of the corneal infection was influenced and terminated by diagnostic pertinent single-drug or combined-drug therapy leading to its ultimate curing; ③ pathogen identification of the sample from the infection site: in bacterial and fungal infections, pathogenic diagnosis either confirmed by sample smear under microscopic examination or organism culture, and in viral infection, pathogenic diagnosis confirmed by polymerase chain reaction (PCR) evaluation of samples from the tear or corneal scraping tissues. In addition to the categories of the corneal infections, images taken from patients suffering from other corneal diseases with similar visual features were classified into the category of other diagnosis. This category includes varieties of corneal dystrophies, phlyctenular keratoconjunctivitis, various corneal tumors, corneal papilloma, corneal degeneration, and even acanthamoeba keratitis. Representative image series for each category are shown in Fig. 1(a).

The final dataset contained 2284 images from 867 patients for this study. The training set consisted of 387 randomly selected images of BK, 519 images of FK, 488 images of HSK, and 528 images of other corneal diseases, from 747 patients. The testing set consisted of 86 randomly selected images of BK, 97 images of FK, 51 images of HSK, and 128 images of other diagnosis, from 120 patients. To evaluate the ophthalmologists' classification performance, the first-time diagnosis images of each patient in the testing set were selected to construct a dataset to evaluate the ophthalmologists (i.e., a total of 120 images had been used to evaluate the performance of ophthalmologists).

3.2. Sequential-level feature learning-based diagnostic deep models

As aforementioned, we devised a sequential-level feature learning method for the classification of infectious keratitis. To demonstrate the superiority of the proposed method, we compared our proposed method with other models, namely image-level feature learning and the patch-level feature learning.

The image-level feature learning deep model uses a transfer learning technique to solve the problem of limited training data [32,33], in which original clinical images without annotation are applied directly to a CNN for diagnostic analysis and classification. In our experiments, we chose three classic architectures for image classification: visual geometry group network (VGG)-16 [34], GoogLeNet-v3 [35], and DenseNet [36].

In the patch-level feature learning deep model, the image of the anterior segment of an eye is initially annotated by manual drawing, dividing the image into four parts: the infectious lesion area of the cornea, the area beyond the lesion of the cornea, the injection of conjunctiva, and the exudation of the anterior chamber. There are three transfer learning architectures in this deep model, that is, VGG-16, GoogLeNet-v3, and DenseNet. After each patch is classified, a method of majority voting is implemented to predict the classification result of each clinical image.

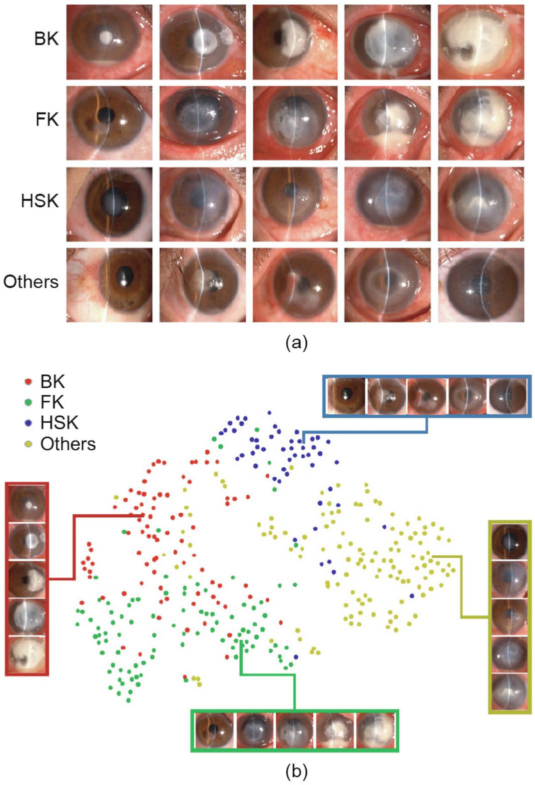


Fig. 1. Representative slit lamp microscopic images and the representations of t-distributed stochastic neighbor embedding (t-SNE) visualization of the embedding features in the proposed SOS model for the four classes of the corneal diseases. (a) The representative slit lamp microscopic images of BK, FK, HSK, and the others, including those apart from the three abovementioned categories of corneal diseases. They exhibit different visual features at different stages of the same disease or show a difference of visual features among categories. (b) The deep features learned by the proposed SOS model being embedded into a two-dimensional space via t-SNE for each category of the disease. t-SNE is utilized for visualizing the high-dimensional data that are the feature representation in the SOS model of the diagnosis-proven photographic test sets (362 images). Colored point clouds represent the different categories of the diseases, showing how the algorithm groups the diseases into different clusters. Insets show images corresponding to various points.

In the sequential-level feature learning model, for each image, the focus of attention is placed on the lesion, if there is any. The centroid of the lesion is annotated to build a minimum circumscribed area. The minimum circumscribed area is further divided into K circular rings scaling up around the center. The partitioning method is illustrated in Fig. 2. From the inner to the outer circular rings, the sampled patches inside the ith circular ring are used to build a set of patches denoted as S_i and a sequence of sets $\{S_1, S_2, \ldots, S_K\}$ following the order from the innermost to the outermost. To address the issue of limited annotated data, in the training process, a drop-out mechanism of randomly dropping out elements from each set is applied, which can generate more sequences of the sets, helping expand the data diversity and making the trained model more robust.

Each patch in a set is applied to a deep residual CNN (i.e., DenseNet) through sequential feature learning via an encoder–decoder framework [37–39]. The convolutional structured encoder can transform the jth patch in the ith set p_{ij} into a vectorial feature f_{ij} to describe its innate characteristics, represented as a set of patch-level features $\{F_1, F_2, \ldots, F_K\}$. For each set F_i, combined overall-patch features can be generated through a max pooling calculation, denoted as the set f_i, which represents the global characteristic over a given set. Since the sets from the innermost to the outermost rings of a lesion consist of a sequence of sets, a long

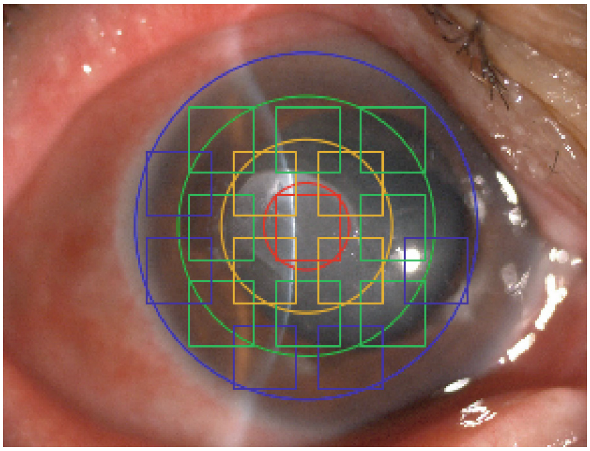


Fig. 2. Illustration of how patches are sampled and how they are divided into *K* sets. Circles represent the boundaries for each set and squares represent the sampled regions. Note that, to avoid excessive overlapping in the picture, only half of the patches are shown.

short-term memory (LSTM) [37], one of the classic models to learn sequential data in deep learning, can be used to transfer the set feature sequence $\{f_1, f_2, \ldots, f_K\}$ into a representation for the classification. The features for the images can be decoded by a fully connected network layer, and the probability of each category of corneal diseases is described by a softmax calculation for the learned features. Fig. 1(b) illustrates the embedding features of each lesion in a two-dimensional space. The working system is shown in Fig. 3. Comparing the results of the predicted probability with the ground-truth type of keratitis, the loss from the result is back-propagated to fine-tune parameters of the model [40,41].

3.3. Recruitment of ophthalmologists for image-based diagnostic analysis

Ophthalmologists were recruited from all over China to test their performances for image based diagnostic analysis as a comparison study with the developed deep learning methodology. The images presented to the recruited ophthalmologists and diagnosis-proven images of each patient in the testing set were randomly selected from the first visit (i.e., a total of 120 images). The recruited ophthalmologists varied in academic title (from residential ophthalmologists to senior ophthalmologists, all the way to full clinical professors in medical schools), affiliation (from teaching hospitals in university medical schools to public municipal hospitals to community clinics), and professional experience (categorized into 1–5 years, 6–10 years, 11–15 years, 16–20 years, and over 20 years). In total, we had recruited 421 ophthalmologists.

The ophthalmologist manual examination for image-based diagnostic analysis followed the two-step protocol. In the first step, an ophthalmologist conducted an image-only diagnosis. Images of four categories of corneal diseases from the first-time diagnosis images of each patient in the testing set, specifically BK, FK, HSK and other corneal diseases, were presented to the ophthalmologist, who made a diagnostic decision for each image through manual examination. Then in the second step, the ophthalmologist was provided with additional standardized and structured medical information affiliated with each image, including brief medical history, time of onset, the grade of pain and recurrence episodes if any, and history of drug use. The ophthalmologist was then asked to make a diagnostic decision for each image through manual examination and by considering the additional medical information. All ophthalmologists performed this procedure independently and without time limitations.

3.4. Statistical analysis

Given the different confidences resulting from the different academic titles, affiliations, and professional experiences, the statistical package for social sciences (SPSS version 18.0; Cary, USA) was used for statistical analysis of the ophthalmologist manual diagnosis data. The average performances denoted as the diagnostic accuracy achieved by the ophthalmologists were summarized and represented in terms of mean ± standard deviation in percentage. Data normality was initially verified using the Kolmogorov–Smirnov test. Differences in the diagnostic accuracy among different hospital levels and professional title groups were analyzed using one-way analysis of variance (ANOVA), in accordance with the data normality. The least significant difference was used for post-hoc analysis of the parametric variables. Correlation between the diagnostic accuracy and the years of professional experience was tested using Pearson's correlation coefficient. Multi-linear regression analysis with the stepwise method was employed to explore the influence of the demographic factors, in terms of academic titles, hospital levels, and years of professional experience. Paired *t*-test (for normally distributed variables) and Wilcoxon signed ranks test (for non-normally distributed variables) were performed to determine if there were any significant differences in diagnostic accuracy between the doctors' performances with and without additional medical information. The significance level for all the tests was set to 0.05.

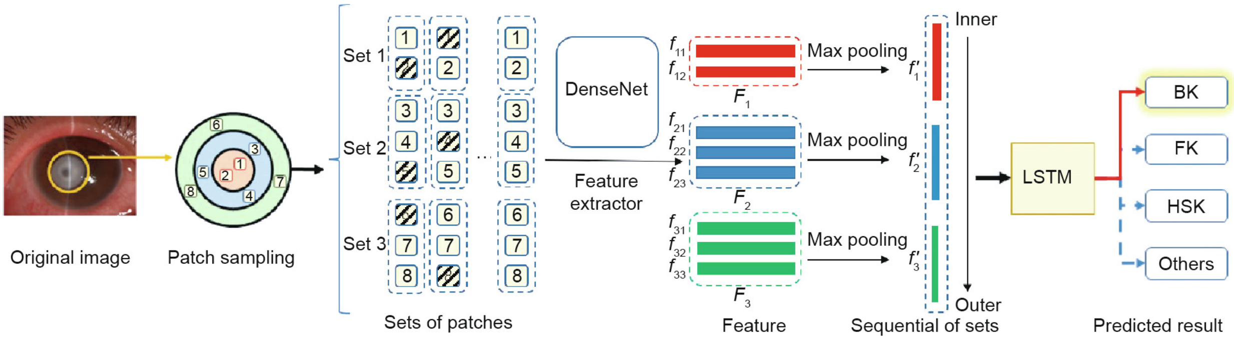


Fig. 3. The process of sequential deep feature learning for one lesion area. For each slit lamp microscopic image, the lesion area is divided into the minimum circumscribed circle to *K* circular ring parts (*K* = 3 here, only for intuitive clarification). From the innermost to the outermost circular rings, we sample patches from each circular ring, and the sampled patches are used to generate sequence of sets. The sequential features can be learned via max-pooling and LSTM.

4. Results

4.1. Performances of the different deep models

Image-level deep model is currently popular for clinical image diagnosis in which original clinical images are directly applied to CNNs. Three classic deep architectures, VGG-16, GoogLeNet-v3, and DenseNet, are used in this study to report the diagnostic performances of this model for BK, FK, and HSK, respectively, documented in Table 1. Considering the fact that the whole image directly applied to CNN in the training process may contain irrelevant information, we thereafter developed a patch-level deep model [42,43] using the VGG-16, GoogLeNet-v3, and DenseNet architectures. In the patch-level deep model, instead of using the whole image, patches including infectious lesion of the cornea, beyond infectious lesion of the cornea, the injection of conjunctiva, and the exudation of the anterior chamber are initially annotated by manually segmentation. We have found that the three patch-level deep models can achieve an accuracy of 49.62%, 51.52%, and 60%, respectively, for patch classifications (i.e., the classification of each patch into a corresponding infectious keratitis). After each patch is classified, majority voting is implemented to perform clinical image classification. The patch-level deep models with voting have respectively achieved an accuracy of 52.50%, 55.52%, and 66.30%, as documented in Table 1.

Finally, we applied sequential-level deep models, which are considered to have the ability to preserve the subtle spatial structures of clinical images. As aforementioned, the sequential-level features are learned in an inner–outer sequential order (referred to as the SOS), and we have achieved 78.73% classification accuracy with SOS features. Instead of generating a sequence of sets in an inner–outer sequential order, we can also generate a sequence in terms of random-ordered patches (ROPs) and sequential-ordered patches (SOPs). ROP generates a sequence of patches via a random order and SOP generates a sequence of patches via an inner–outer order (but without the utilization of a set structure to group each patch into different sets). The final evaluation shows that ROP features yielded an accuracy of 74.23% (75.29% for BK, 68.04% for FK, and 82.35% for HSK) and SOP features yielded an accuracy of 75.14%. This evaluation study has demonstrated that the sequential-level deep models are the best models for automatic, image-only diagnosis for corneal diseases.

4.2. Comparison study on ophthalmologists' diagnosis

We evaluated all the algorithms that we considered in this paper using the dataset to compare the performances between each algorithm and the ophthalmologist. Table 2 lists the accuracies of all the algorithms and the average performance of ophthalmologists on this dataset (120 images). The performances of ophthalmologists in the diagnosis of clinical images are listed in Table 3. There were 421 ophthalmologists recruited from all over China participating in this study. The average accuracy performance all ophthalmologists without additional medical information was 49.27% ± 11.5% (range: 20.00%–86.67%), which was far lower than that achieved by AI deep learning models. For example,

Table 1
Performance of classification accuracy among different deep learning models on the test dataset.

Level	Algorithm	Test dataset (%)				
		Acc	BK	FK	HSK	Others
Image-level	VGG-16 (image)	55.24	48.84	52.57	62.74	53.90
	GoogLeNet-v3 (image)	57.73	53.49	55.67	66.67	58.59
	DenseNet (image)	61.04	60.46	56.70	80.39	57.03
Patch-level	VGG-16 (voting)	52.50	45.34	54.64	56.00	54.68
	GoogLeNet-v3 (voting)	55.52	44.19	51.55	74.51	58.59
	DenseNet (voting)	66.30	59.30	68.04	58.82	72.66
Sequential-level	Random-ordered patches (ROPs)	74.23	75.29	68.04	82.35	75.00
	Sequential-ordered patches (SOPs)	75.14	66.28	86.60	84.31	68.75
	SOSs	78.73	65.12	83.51	90.20	79.70

The image-level, patch-level, and sequential-level features are learned from the whole image, the lesion area, and the sequence of patch sets, respectively. The test dataset contains 362 images, including 86 BK, 97 FK, 51 HSK, and 128 others from 120 patients. Acc indicates the overall accuracy of each model, and columns BK, FK, HSK, and others show the recall for each corresponding category.

Table 2
Deep learning models competing with ophthalmologists using a dataset of 120 images in total.

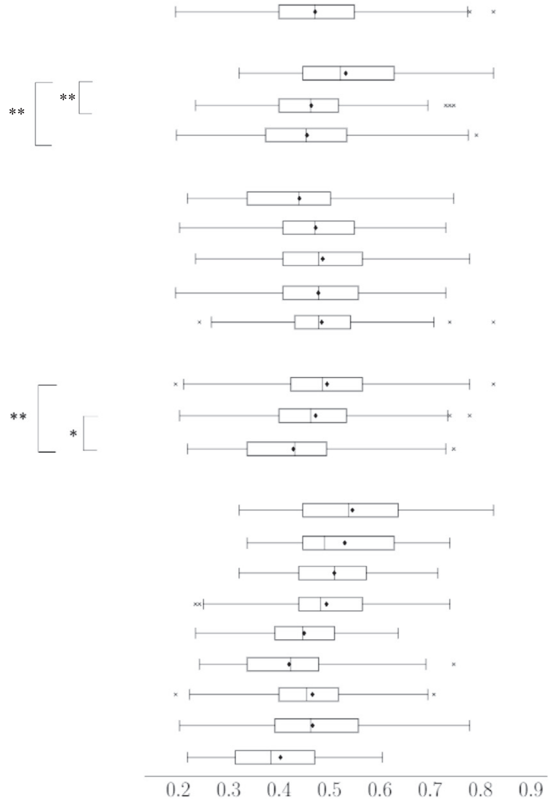
Level	Algorithm	Dataset for evaluation of ophthalmologists (%) (%)				
		Acc	BK	FK	HSK	Others
Image-level	VGG-16 (image)	50.83	46.67	43.33	73.33	40.00
	GoogLeNet-v3 (image)	55.83	50.00	63.33	70.00	40.00
	DenseNet (image)	64.17	56.67	63.33	80.00	56.67
Patch-level	VGG-16 (voting)	51.67	23.33	43.33	76.67	63.33
	GoogLeNet-v3 (voting)	54.17	26.67	73.33	80.00	36.67
	DenseNet (voting)	71.67	46.67	86.67	73.33	80.00
Sequential-level	ROPs	77.50	66.67	70.00	93.33	80.00
	SOPs	79.17	73.33	70.00	96.67	76.67
	SOSs	80.00	53.33	83.33	93.33	90.00
Human-level	Average performance of ophthalmologists provided with image only	49.27	46.55	45.56	65.01	39.95
	Average performance of ophthalmologists provided with image together with medical history	57.16[a]	55.55[a]	56.28[a]	73.25[a]	43.56[a]

The first-visit and diagnosis-proven images of patients with four categories of the corneal diseases were selected from the testing set to construct a dataset for evaluation and comparison of deep learning models with ophthalmologists. The dataset included 120 clinical images.

[a] $P < 0.001$ compared to the average performance of ophthalmologists provided with image only.

Table 3
Average classification accuracy performance according to the hospital level, years of employment, and professional titles of the ophthalmologists.

Dr. Group	Participant number	Boxplot	Mean ± STD (%)	Range (%)
Total	421		49.27 ± 11.85	[20.00, 86.67]
Hospital RK				
Teaching	84		55.69 ± 12.19	[33.33, 86.67]
City	171		48.46 ± 10.70	[24.17, 78.33]
Community	166		46.84 ± 11.63	[20.00, 81.67]
Year of employments				
1–5	89		45.96 ± 13.22	[22.50, 78.33]
6–10	117		49.39 ± 11.80	[20.83, 76.67]
11–15	69		50.85 ± 12.03	[24.17, 81.67]
16–20	41		49.94 ± 12.30	[20.00, 76.67]
> 20	105		50.64 ± 9.66	[25.00, 86.67]
Physician RK				
Attending	173		51.76 ± 11.94	[20.00, 86.67]
Fellow	150		49.38 ± 11.18	[20.83, 81.67]
Resident	98		44.69 ± 11.66	[22.50, 78.33]
Hospital RK and physician RK				
Attending in teaching	36		57.08 ± 12.02	[33.33, 86.67]
Fellow in teaching	30		55.47 ± 12.21	[35.00, 77.50]
Resident in teaching	18		53.29 ± 12.21	[33.33, 75.00]
Attending in city	78		51.63 ± 10.45	[24.17, 77.50]
Fellow in city	59		46.96 ± 9.07	[24.17, 66.67]
Resident in city	34		43.80 ± 11.57	[25.00, 78.33]
Attending in community	59		48.69 ± 10.36	[20.00, 74.17]
Fellow in community	61		48.72 ± 12.53	[20.83, 81.67]
Resident in community	46		41.99 ± 10.51	[22.50, 63.33]



RK: rank; STD: standard deviation; * $P = 0.003$; ** $P < 0.001$.

the SOS algorithm achieved a diagnostic accuracy of 80%, including accuracies of 53.33%, 83.33%, and 93.33% for BK, FK, and HSK, respectively (Table 2). Fig. 4 depicts the receiver operating characteristic (ROC) curve, the confusion matrix of SOS model, and the performance of ophthalmologists. The ROC curve is a visualization method for classification models. The area under the curve (AUC) is a measure of performance, with a maximum value of 1. The model achieves superior performance over an ophthalmologist if the sensitivity–specificity point of the ophthalmologist lies below the curve of the classification model.

The effect of location of work on the ophthalmologists' performance was revealed in this study, wherein those from teaching hospitals demonstrated a far better performance than those from city hospitals and community clinics (both probability value $P < 0.001$), whereas no significant difference was found between city hospitals and community clinics ($P = 0.226$). The ophthalmologists with higher professional ranks appear to have a better performance in diagnosing clinical images with a better accuracy, such as the attending ophthalmologists and fellows who performed better than residents ($P < 0.001$ and $P = 0.003$, respectively), but no significant difference was found between the groups of attending and fellow ophthalmologists ($P = 0.071$). No significant correlation was found between the duration of employment and diagnostic accuracy ($P = 0.084$).

When the factors of hospital ranking and doctor's ranks were considered together, better performance was found in the group of ophthalmologists with the attending title from teaching hospitals (accuracy of 57.08% ± 12.02%, range: 33.33%–86.67%) than the group of ophthalmologists with resident ranks from community clinics (accuracy of 41.99% ± 10.51%, range: 22.50%–63.33%). The stepwise multiple regression analysis resulted in three models that affected diagnostic accuracy. Model 1 (coefficient of determination $R^2 = 0.062$) had only the factor of hospital levels (beta error $\beta = 0.254$, $P < 0.001$); Model 2 ($R^2 = 0.100$) had the factors of hospital levels ($\beta = 0.239$, $P < 0.001$) and professional titles ($\beta = 0.200$, $P < 0.001$); Model 3 ($R^2 = 0.109$) had all three factors of hospital levels ($\beta = 0.227$, $P < 0.001$), professional titles ($\beta = 0.326$, $P < 0.001$), and years of employment ($\beta = -0.164$, $P = 0.024$).

When the ophthalmologists were further provided with additional medical information affiliated to each image, including brief medical history, time of onset, the grade of pain and recurrence episodes if any, and history of drug use, the mean total diagnostic accuracy increased from 49.27% to 57.16%, resulting in a statistically significant difference (Wilcoxon signed ranks test, $P < 0.001$). In detail, the accuracy increased from 46.55% to 55.55% ($P < 0.001$) for BK, from 45.56% to 56.28% ($P < 0.001$) for FK, and from 65.01% to 73.25% ($P < 0.001$) for HSK. With additional medical information, the mean total accuracy of 404 doctors increased by 8.28%, the accuracy of nine doctors decreased by 2.13%, and the accuracy of the other eight doctors remained unchanged.

5. Discussion

In general, fact judgment by humans is achieved through vision, audio, touch, taste, and smell, which enables a person to classify things into appropriate categories [44]. Visual perception plays the most important role for this purpose [45] and visual knowledge

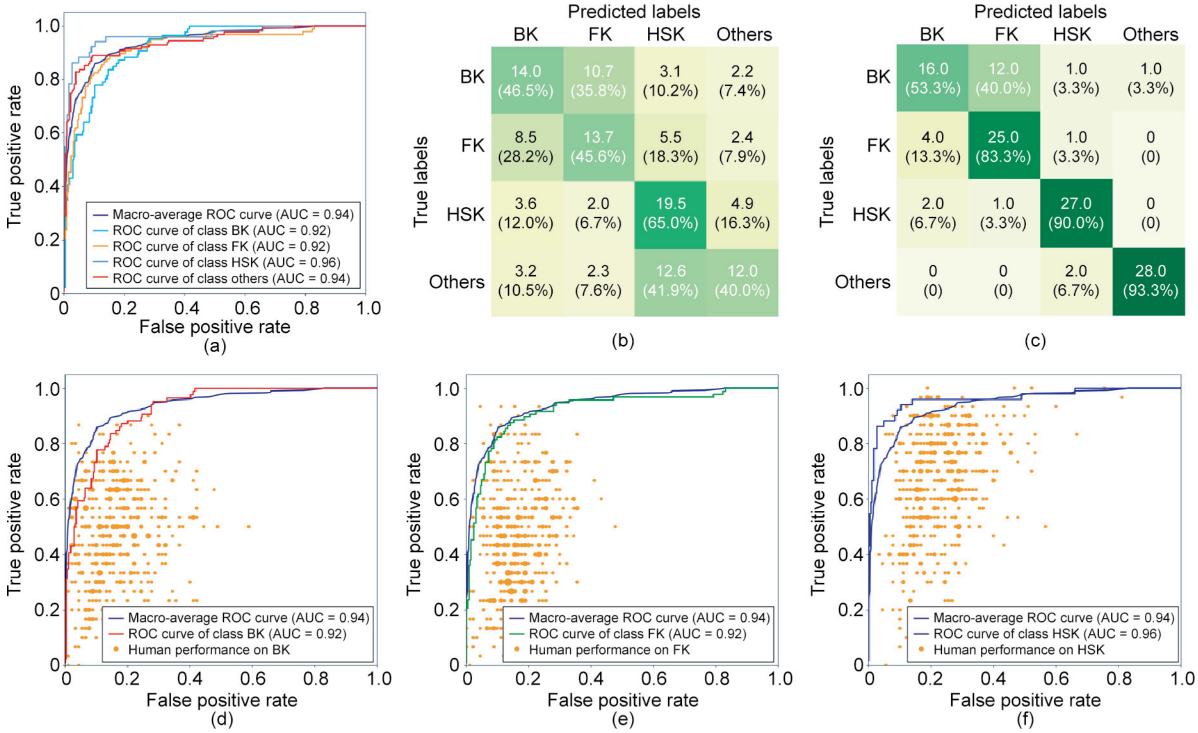


Fig. 4. ROC curve and confusion matrix of SOS model and the performance of ophthalmologists. (a) ROC curve of SOS model; (b, c) confusion matrices of the ophthalmologists and the SOS model on the dataset for evaluation of ophthalmologists; (d–f) ROC curves for the disease categories of BK, FK, and HSK, respectively. AUC: the area under the curve.

can describe the relation between spatial shapes, sizes and correlation, as well as colors and textures [46]. Physicians making diagnoses for diseases primarily depend on observation and reasoning. Among all human diseases, corneal diseases have the most direct and most significant displays of changes in visual perception, because the healthy cornea of an eye has a unique characteristic of complete transparency, which is in sharp contrast to the pathological conditions that always manifested as image changes in and beyond the cornea. Diagnostic decision-making of corneal disease by human professionals is carried out through image understanding and analysis, which is likely the most appropriate task for an AI to provide assistance to humans.

Generally speaking, deep learning is driven by a large amount of annotated data [47,48]. However, it is not clear how much training data of clinical images is sufficient for developing an AI system for diagnosing clinical diseases. Our center has been collecting and documenting corneal disease cases with clinical images for 20 years, but when all the images are annotated according to each disease category, there can be thousands of clinical images in the most common disease categories, whereas there are only a few dozen clinical images in some rare disease categories. The imbalance of the annotated data in each corneal disease category leads us to focus on the most common diseases, such as infectious corneal diseases, to develop the first stage of AI diagnostic system in this study.

In this study, we have demonstrated that deep learning through CNNs can be applied to clinical diagnosis for corneal infectious diseases using clinical images taken via slit lamp microscopy. We have evaluated three sets of nine deep learning architectures in total in an effort to develop an image-only diagnostic system for corneal infectious diseases. From the results of image-level and patch-level deep models, we can say that despite only having four categories, this is a hard problem, especially for VGG-16 and GoogLeNet-v3. These two structures achieved poor performance in patch classification, and as a result, voting among patches did not improve their performance significantly. In contrast, DenseNet reached 60% in patch classification and achieved 66.3% after voting. This shows that focusing on patches from the infectious lesion area can yield higher performance than seeing the whole picture, provided that the model performs well enough in patch classification. The ROP method can be viewed as another way to combine patch features besides voting. Its result shows that even without spatial information, patch-level deep model can be further improved given the appropriate combining method. We have found that overall, SOS is the most promising method for image-only diagnosis for corneal disease images. A possible reason for why SOS was better than the other methods is how an appropriate utilization of spatial structures of clinical images is directly implemented into this model of deep learning. SOP did not perform as well because it did not consider the circular structure of the lesion area. To the best of our knowledge, this is the first study that presents a deep learning model to perform corneal disease classification with higher accuracy than that of human ophthalmologists in image-only diagnosis. It was noted in this study that general professional human performance in image-only corneal disease diagnosis was worse than that of an AI system. There is no doubt that incorrect diagnosis can lead to prolonged use of inappropriate medications that cause the identifying features to be obscured [6], making human decision-making for diagnosis more difficult. The multiple regression analysis in our study demonstrated that the three demographic factors, in terms of academic ranks, affiliations, and professional service duration, had influences on the diagnostic performance, whereas the coefficients of determination were low in the three models. This indicates that the above factors may not truthfully and comprehensively determine the diagnostic accuracy of corneal diseases in ophthalmologists, or the factors affecting the diagnostic performance may be very complicated and may not be accurately summarized simply by the above three factors. Therefore, if AI can help clinicians improve their ability significantly with a higher diagnostic accuracy, this will greatly benefit patients suffering from corneal diseases, save medical resources, and reduce societal burden. There is still a large population of 4.5 million individuals who are now suffering from moderate to severe vision impairment due to the loss of corneal clarity caused by corneal diseases worldwide [4], especially in developing countries. There are two ways of raising diagnostic accuracy. One is to improve the physician training system and to strengthen the professional education and training for physicians; the other is to develop a practical AI system to assist in diagnosis. Our current study demonstrates that it is realistically achievable to develop an AI system by using clinical images to improve the diagnostic accuracy for corneal diseases. In examining the ophthalmologists' performance, we found that when the medical professionals were provided with images together with medical history, the diagnostic accuracy increased to a certain extent (from 49.27% to 57.16%, $P < 0.001$) as compared to the accuracy when the professionals were provided with images only. This result indicates that, while additional information can help to further improve the performance. This may also be true to AI diagnostic systems. Researches show that integrating data-driven machine learning with human knowledge can effectively lead to explainable, robust, and general AI [49]; and information like medical history may contain humanlike common sense which can enable models to solve many different tasks with limited training data [50]. for improving our AI diagnostic system to raise the diagnostic accuracy, a multi-modal learning model (i.e., the effective combination of visual and non-visual information) or a more suitable sequential learning model may need to be devised in future work.

It is undeniable that our AI diagnostic accuracy at this stage is only confirmed by the limited image data we have collected, through a comparison study with the performances of ophthalmologists using the same clinical images. A real-world application of such an AI system in assisting physicians in clinical practice requires further and more extensive clinical evaluations on a larger scale [51].

6. Conclusions

Infectious keratitis is the most common ophthalmological disease that may cause blindness. Ophthalmologists observe and diagnose diseases by observing slit lamp images, facilitating diagnosis using computer-aided image analysis algorithms. In this work, we propose a sequential-level deep model for end-to-end diagnosis of infectious keratitis. Specifically, relying on the excellent feature extraction performance of deep convolutional networks, we first extract the detailed patterns of the corneal region and then group the local features into an ordered set that conforms to the spatial structure to learn the global representation of the corneal image and perform diagnosis. We collected over 110 000 images from more than 10 000 patients. On that basis, sufficient experimental comparison results proved that our model is a more feasible structure and has achieved better diagnostic performance than those conventional CNNs. In addition, through a comparison with more than 400 professional ophthalmologists, we found that our model can greatly exceed the average level of professionals and reach the level performance of top ophthalmologists. To the best of our knowledge, this is the first study on the diagnosis of infectious keratitis, and our research has strongly demonstrated the potential of using AI to perform clinically assisted diagnosis of these types of diseases.

Acknowledgements

This research is supported by the Health Commission of Zhejiang Province (WKJ-ZJ-1905 and 2018ZD007), the Key Research and Development Projects of Zhejiang Province (2018C03082), and the National Natural Science Foundation of China (61625107). We thank Zhongfei Zhang for discussions and comments.

Compliance with ethics guidelines

Yesheng Xu, Ming Kong, Wenjia Xie, Runping Duan, Zhengqing Fang, Yuxiao Lin, Qiang Zhu, Siliang Tang, Fei Wu, and Yu-Feng Yao declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Bejnordi BE, Veta M, van Diest PJ, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318 (22):2199–210.

[2] 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 (7639):115–8.

[3] Sommer A, Taylor HR, Ravilla TD, West S, Lietman TM, Keenan JD, et al. Challenges of ophthalmic care in the developing world. JAMA Ophthalmol 2014;132(5):640–4.

[4] Pascolini D, Mariotti SP. Global estimates of visual impairment: 2010. Br J Ophthalmol 2012;96(5):614–8.

[5] Clemens LE, Jaynes JM, Lim E, Kolar SS, Reins RY, Baidouri H, et al. Designed host defense peptides for the treatment of bacterial keratitis. Investig Ophthalmol Vis Sci 2017;58(14):6273–81.

[6] Gopinathan U, Garg P, Fernandes M, Sharma S, Athmanathan S, Rao GN. The epidemiological features and laboratory results of fungal keratitis: a 10-year review at a referral eye care center in South India. Cornea 2002;21(6):555–9.

[7] Yang Y, Luyten W, Liu L, Moens MF, Tang J, Li J. Forecasting potential diabetes complications. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence; 2014 Jul 27–31; Quebec City, QC, Canada; 2014. p. 313–9.

[8] He T, Guo J, Chen N, Xu X, Wang Z, Fu K, et al. MediMLP: using Grad-CAM to extract principal variables for lung cancer postoperative complication prediction. IEEE J Biomed Health Inf 2020;24(6):1762–71.

[9] Nee P, Li Y, Zhu J, Peng J, Dai Z, Li G, et al. Disease diagnosis prediction of EMR based on BiGRU–Att–CapsNetwork model. In: Proceedings of 2019 IEEE International Conference on Big Data (Big Data); 2019 Feb 27–Mar 2; Los Angeles, CA, USA; 2019. p. 6166–8.

[10] Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential patternmining to predict next prescribed medications. J Biomed Inf 2015;53:73–80.

[11] Scott IM, Cootes TF, Taylor CJ. Improving appearance model matching using local image structure. In: Proceedings of Biennial International Conference on Information Processing in Medical Imaging; 2003 Jul 20–25; Ambleside, UK; 2003. p. 258–69.

[12] Manousakas IN, Undrill PE, Cameron GG, Redpath TW. Split-and-merge segmentation of magnetic resonance medical images: performance evaluation and extension to three dimensions. Comput Biomed Res 1998;31 (6):393–412.

[13] Zhao Y, Gui W, Chen Z, Tang J, Li L. Medical images edge detection based on mathematical morphology. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference; 2006 Jan 17–18; Shanghai, China; 2006. p. 6492–5.

[14] Kaus MR, von Berg J, Weese J, Niessen W, Pekar V. Automated segmentation of the left ventricle in cardiac MRI. Med Image Anal 2004;8(3):245–54.

[15] Cordes D, Haughton V, Carew JD, Arfanakis K, Maravilla K. Hierarchical clustering to measure connectivity in fMRI resting-state data. Magn Reson Imaging 2002;20(4):305–17.

[16] Pohl KM, Fisher J, Shenton M, McCarley RW, Grimson WEL, Kikinis R, et al. Logarithm odds maps for shape representation. In: Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention; 2006 Oct 1–6; Copenhagen, Denmark; 2006. p. 955–63.

[17] Lee LK, Liew SC, Thong WJ. A review of image segmentation methodologies in medical image. In: Sulaiman HA, Othman MA, Othman MFI, Rahim YA, Pee NC, editors. Advanced computer and communication engineering technology. Cham: Springer; 2011. p. 1069–80.

[18] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48.

[19] 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 (9):60–88.

[20] Zhou X, Takayama R, Wang S, Hara T, Fujita H. Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. Med Phys 2017;44(10):5221–33.

[21] Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 2019;290(2):537–44.

[22] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–61.

[23] Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020;39(4):1184–94.

[24] Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal V, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392(10162):2388–96.

[25] Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 2018;321:321–31.

[26] Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewing DJ, Satam GP, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25(1):70–4.

[27] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65–9.

[28] Bejnordi BE, Veta M, Van Diest PJ, van Ginneken B, Karssemeijer N, Litjen G. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–210.

[29] 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 (22):2402–10.

[30] Ting DS, Cheung CY, Lim G, Tan GS, 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 populations with diabetes. JAMA 2017;318(22):2211–23.

[31] Kim SJ, Cho KJ, Oh S. Development of machine learning models for diagnosis of glaucoma. PLoS ONE 2017;12(5):e177726.

[32] Kermany DS, 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(5):1122–31.

[33] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22(10):1345–59.

[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.

[35] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 26–Jul 1; Las Vegas, NV, USA; 2016. p. 2818–26.

[36] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honoululu, HI, USA; 2017. p. 4700–8.

[37] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9 (8):1735–80.

[38] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. 2014. arXiv:1406.1078.

[39] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. 2014. arXiv:1409.3215.

[40] Schmidhuber J, Wierstra D, Gagliolo M, Gomez F. Training recurrent networks by evolino. Neural Comput 2007;19(3):757–79.

[41] Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Sys 2016;28 (10):2222–32.

[42] Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. 2016. arXiv:1606.05718.

[43] Lam C, Yu C, Huang L, Rubin D. Retinal lesion detection with deep learning using image patches. Invest Ophthalmol Vis Sci 2018;59(1):590–6.

[44] Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. Science 1959;130(3366):9–21.

[45] Claus-Christian C. Understanding human perception by human-made illusions. Front Hum Neurosci 2014;8:566.

[46] Pan Y. On visual knowledge. Front Inf Technol Electron Eng 2019;20 (8):1021–5.

[47] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.

[48] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504–7.

[49] Zhuang YT, Wu F, Chen C, Pan YH. Challenges and opportunities: from big data to knowledge in AI 2.0. Front Inf Technol Electron Eng 2017;18 (1):3–14.

[50] Zhu Y, Gao T, Fan L, Huang S, Edmonds M, Liu H, et al. Dark, beyond deep: a paradigm shift to cognitive AI with humanlike common sense. Engineering 2020;6(3):310–45.

[51] Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019;1 (1):20–3.