The Visual Computer (2021) 37:2407–2417
https://doi.org/10.1007/s00371-020-01994-3

ORIGINAL ARTICLE

Fundus image-based cataract classification using a hybrid convolutional and recurrent neural network

Azhar Imran[1] · Jianqiang Li[1,2] · Yan Pei[3] · Faheem Akhtar[1,4] · Tariq Mahmood[1,5] · Li Zhang[6]

Accepted: 7 October 2020 / Published online: 8 November 2020

Abstract

Cataract is the most prevailing reason for blindness across the globe, which occupies about 4.2% population of the world. Even with the developments in visual sciences, fundus image-based diagnosis is deemed as a gold standard for cataract detection and grading. Though the increase in the workload of ophthalmologists and complexity of fundus images, the results may be subject to intelligence. Therefore, the development of an automatic method for cataract detection is necessary to prevent visual impairment and save medical resources. This paper aims to provide a novel hybrid convolutional and recurrent neural network (CRNN) for fundus image-based cataract classification. The proposed CRNN fuses the advantages of convolution neural network and recurrent neural network to preserve long- and short-term spatial correlation between the patches. Coupled with transfer learning, we adopt AlexNet, GoogLeNet, ResNet and VGGNet to extract multilevel feature representation and to analyse how well these models perform cataract classification. The results demonstrate that the proposed method outperforms state-of-the-art methods with an average accuracy of 0.9739 for four-class cataract classification and provides a compelling reason to be applied for other retinal diseases.

Keywords Cataract detection · Fundus images · CNN · Retinal diseases · Transfer learning

1 Introduction

Cataract is the clouding of the lens which develops slowly and leads to vision impairment [1]. It is a prevalent foundation of blindness across the globe, which has affected about 314 million people [25]. According to the World Health Organization (WHO) report, cataract accounts for 39% of visual impairment and 51% of blindness in which 90% of these people are from underdeveloped countries [24]. It is accepted worldwide that the early diagnosis of cataract can reduce the risks of permanent visual loss, but it necessitates professional ophthalmologists and other eye care services that are not frequently available in remote areas. Besides the other imaging modalities such as slit-lamp images, retro-illumination images, and ultrasonic Nakagami images, the analysis of retinal fundus images is the most widely used method for cataract diagnosis [11].

There are several challenges associated with fundus images for cataract diagnosis. First, many previous studies of cataract detection and grading are based on manual feature extraction [7,10,20,26,37], which is a complicated and time-taking effort. Second, retinal datasets are very small and highly unbalanced. Third, professional ophthalmologists with profound experience are needed to assess the reliability of the fundus image diagnosis. Therefore, an automatic method for fundus image analysis is required to alleviate the above-mentioned problems.

Recently, deep learning (DL) has evolved as an emerging field of computer vision and image processing [16]. The

This work was supported by the Key Program of National Natural Science of China (Grant No. 71432004).

✉ Jianqiang Li
lijianqiang@bjut.edu.cn

[1] School of Software Engineering, Beijing University of Technology, Beijing 100124, China

[2] Beijing Engineering Research Center for IoT Software and Systems, Beijing 100124, China

[3] Computer Science Division, University of Aizu, Aizuwakamatsu, Fukushima 965-8580, Japan

[4] Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan

[5] Division of Science and Technology, University of Education, Lahore 54000, Pakistan

[6] Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China

convolutional neural network (CNN) is the main type of deep neural network, which is used for analysing visual imagery [30]. CNN does not require human intervention for feature extraction, unlike the traditional ML methods. Nevertheless, the challenges remain in the use of CNN, such as the availability of labelled data, retinal features such as vessels are so complex that may affect the other vessels and dependent on image quality, and the manually labelled fundus images are prone to subjectivity. Therefore, transfer learning (TL) is used to alleviate these challenges. TL has made substantial progress and gained strong attention over the last few years. TL methods are initially trained on the larger dataset (ImageNet) and can be reused for any other image classification task by slightly adjusting the hyperparameters [23]. The fundus images are characterized by high resolution (3888*2592), unlike the natural images. It is very difficult to input the whole fundus image for cataract classification into CNN because of the limited memory of GPUs. Moreover, it is also unrealistic to directly resize the high-resolution images to low-resolution images, because the substantial information of fundus images like blood vessels and optic disk can be lost.

The preceding methods segment the whole image into small patches and then apply CNN to excerpt feature representation of every patch and finally combine the patches to complete the classification results by using majority voting or traditional ML methods, e.g. support vector machine (SVM) [32]. Although the mainstream methods have achieved optimal classification accuracy on normal images, they generally face three main challenges on high-resolution images. First, the current patch-based method does not combine the patches effectively as they only integrate the short-distance spatial dependencies to make an image-based classification. These methods neglect the long-distance dependencies, which is also very essential to understand the contextual information of the complete image. Next, the feature representation of fundus images is not very richer, and most of the considerable information vanishes before the image-wise classification. Third, some challenges are associated with the size of the dataset. Even some larger datasets such as ImageNet [16] are available publically with millions of images, but the medical imaging datasets are not very large enough. Some researchers have released labelled retinal datasets such as DRIVE [35] and STARE [14], but they are relatively very small for deep learning-based image classification.

In this paper, we propose a novel CRNN method that combines the features of CNN and RNN to preserve the short- and long-term spatial correlation amongst the patches. First, the whole image is split into smaller patches. Then, the pre-trained models of transfer learning: AlexNet, GoogLeNet, ResNet and VGGNet, are used to extract the feature representation of these smaller patches. All the feature vectors extracted individually from its corresponding CNN architecture are incorporated into a fully connected layer using global average pooling (GAP) and termed as ’multilevel feature set’. Finally, the multilevel feature set is used as an input of bidirectional long- and short-term memory RNN (BLSTM-RNN) to make an image-wise four-grade cataract classification. The proposed CRNN has attained an average accuracy of 97.39% and outperforms the baseline methods.

2 Related works

In the past few years, the development of automated methods for cataract detection and grading has made substantial progress. There are various studies conducted to analyse the fundus image for the classification of retinal diseases like glaucoma [21], keratoconus [34], macular degeneration [11], and diabetic retinopathy (DR) [6]. The fundus image analysis for cataract diagnosis is the least focused area of the research.

2.1 Traditional methods for cataract detection and grading

The fundus image unveils the everlasting prospects for the detection and classification of the cataract. Most of the traditional methods of cataract diagnosis are based on human-engineered features with optical coherence tomography [28]. Fan et al. [10] extracted wavelet- and sketch-based features and reduced the dimensionality of features using principal component analysis (PCA) and applied machine learning (ML) methods for cataract grading and classification. Manchalwar et al. [20] used a histogram of oriented gradient (HOG) feature with minimal distance classifier for the detection of cataract and conjunctivitis. Another approach for fundus image analysis was proposed by Qiao et al. [26] in which three manual feature sets (colour, texture, wavelet) were extracted and weighted by genetic algorithm and SVM was used for classification. Xiong et al. [37] classified five grades of cataract blurriness with vitreous opacity in which the features of standard deviation, mean and pixel number of the visible structure were extracted and fed into the decision tree method. The cataract classification based on vascular information was proposed by Dong et al. [7] in which Kirsh template filter was used to obtain vessel information and wavelet and texture features were extracted followed by SVM for cataract grading. These methods extract the features in a single direction, which is insufficient to capture all the details of the fundus image.

Cao et al. [5] presented an improved Haar wavelet feature and added the detail component for presenting the texture features in horizontal, vertical and diagonal directions. The cataract classification method was based on majority voting, which has shown higher accuracy for cataract detection and grading. The number of annotated fundus images is usually limited, and to overcome this issue, a semi-supervised fundus image analysis was proposed by Song et al. [33]; the wavelet, sketch and textures were used for multiclassifier-based cataract classification. Yang et al. [40] proposed an ensemble learning-based method for cataract classification; three feature sets are texture, sketch, and wavelet; two classifiers are SVM and BPNN; and ensemble methods of stacking and majority voting were examined for image classification. A pattern recognition-based fundus image analysis was performed by Zheng et al. [42] in which the spectrum features were calculated by using 2-D discrete Fourier transform and linear discriminant analysis followed by AdaBoost algorithm was used for cataract classification. Xiong et al. [38] presented a new approach based on a multifeature set that combines texture features from the grey-level co-occurrence matrix (GLCM) and higher-level features from pre-trained ResNet model, and these fused features are fed into SVM for automatic cataract classification.

These cataract classification techniques with manual feature extraction have three main limitations. First, the extraction of features must be verified by the retinal experts, which is time-consuming and increases the burden on ophthalmologists. Second, the robustness and generalization of these studies are limited because the results are evaluated on smaller datasets. Third, the clinical symptoms of a cataract are not very clear, and the size of a vessel is unsatisfactory for professional graders. Thus, it is difficult to discern vessel features accurately and perform cataract classification. Therefore, the goal of this study is to process automatic feature engineering by applying DL concepts.

2.2 Deep learning for cataract detection and grading

DL networks are generally used for medical image analysis, and CNN provides great support to address the aforesaid problems. The CNN model extracts the local features directly from the fundus image without human intervention. Dong et al. [8] proposed a DL-based 4-level cataract classification. A deep learning framework Caffe and the softmax function were used for the grading of cataract. Zhang et al. [41] presented a deep convolutional neural network (DCNN) with 8 layers (5 convolutions and 3 fully connected) and used a softmax function to produce four-level cataract classification. As the deeper network discerns more features, a hybrid method based on DCNN (with 17 layers) and random forests (RF) was presented by Ran et al. [27]. Firstly, DCNN used the residual network to learn the abstract representation of the fundus image. Then, these features were fed into RF for six-level cataract grading. Li et al. [18] proposed an 18-layer deep learning model that inputs a G-channel of the retinal image and outputs the cataract classes with heatmaps based on global features. A convolutional and deconvolution network (DN)-based cataract classification was performed by Xu et al. [39] in which CNN was employed to learn features from the fundus image, and DN was used to examine how CNN performs layer-by-layer cataract classification; this model gives the detailed insight of global-local feature representation. The pre-trained models such as AlexNet and GoogLeNet were modified with class activation pooling (CAM) by Li et al. [17] in which the last two fully connected layers were replaced with the global pooling layer to obtain better accuracy of cataract classification.

Although recently DL methods combine the results of image patches (patch-wise) to produce image-wise classification either combining short-distance dependencies or using SVM and majority voting. Most of these methods simply ignore the long-distance spatial dependencies of the fundus image. Furthermore, current methods average pool the last fully connected layer of CNN into 1-D feature vector and use as a feature illustration of image patches, which is not appropriate for patch-wise synthesis.

3 Dataset

3.1 Materials

Our dataset comprises 8030 high-resolution (3888*2592 pixels) RGB fundus images. The images are captured from high-resolution fundus camera Canon-EOS-40D with additional settings such as 72 DPI resolution, no-flash, manual exposure, and auto-white balance. The dataset is collected from Tongren Hospital, China, which is graded into four classes (0–3) by professional ophthalmologists as given in Fig. 1 and Table 1. These images are labelled by two professional graders according to the severity of cataract from class 0–3 for normal (no-cataract), mild, moderate and severe, respectively.

3.2 Image processing

The characteristics of cataract may be overlapped with other eye diseases due to the complexity of retina. Moreover, the fundus images contain a range of imaging noise, like low contrast, dark space on either side of the image, and ambiguity between the vessel and background. Therefore, some pre-processing steps are required to improve the quality of the fundus images. Patch is the subsection of an input image. It describes that the kernel looks at one patch of an image at a time, and then, it moves to next patch, and so on. The patch is the input to the kernel. CNN kernels/filters only process one patch at a time, rather than the whole image. This is because we want filters to process small pieces of the image in order to detect features. First, we devised an algorithm to remove the dark spaces on either side of the image by cropping a fixed

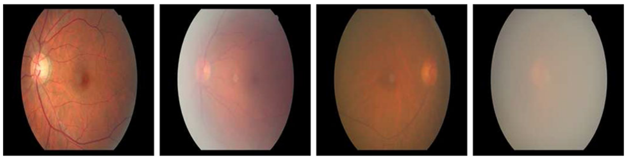

Fig. 1 Various levels of cataract severity based on the vessel details (both small and large) and optic disk. Both the small and large vessels and optic disk are visible in normal image, while nothing is apparent in severe image. **a** Normal, **b**, mild **c** moderate, **d** severe

Table 1 Dataset classification

Class	Severity	Numbers
0	Normal	4671
1	Mild	2283
2	Moderate	675
3	Severe	401

amount of pixels. This algorithm takes image as an input with size 3888*2592 pixels and removes dark spaces of 50 pixels from all the sides of the image. Then, we employed a baseline normalization scheme by subtracting the mean of all pixels and dividing by the variance, as in Eq. 1:

$$y = \frac{(x - \bar{x})}{\sigma} \tag{1}$$

where $\bar{x}$ represents mean value and σ is the variance. The 'mean' value represents the individual pixel intensity of the entire image, while variance is used to distinguish each pixel from neighbouring pixels. The mean value is calculated by adding all the pixel values and then dividing by a total number of pixels $\bar{x} = \text{sum}(x)/\text{length}(x)$. The variance is calculated by simple averaging the numbers, subtracting the mean and computing the average of squared difference $\sigma = \Sigma(x - \bar{x})^2/(\text{length}(x) - 1)$. Finally, we used non-local means denoising (NLMD) [4] method to remove the noisy data. The denoising of an image x at channel i and pixel j is given in Eq. 2:

$$\hat{x}_i(j) = 1/c_j \sum_{k\epsilon\beta(j,r)} x_i(j)w(j,k), C(j) = \sum_{k\epsilon\beta(j,r)} w(j,k) \tag{2}$$

where $i = 1, 2, 3, \ldots, C_j$ is the normalization factor, w is the weight function and $\beta(j, r)$ represents the neighbouring centre at pixel j. These estimates can be finally averaged at each pixel location to produce a final denoised image. The resultant pre-processed image is acquired after applying these pre-processing steps. The pre-processing results of normal, mild, moderate and severe cataract classes corresponding to the original image are shown in Fig. 2.

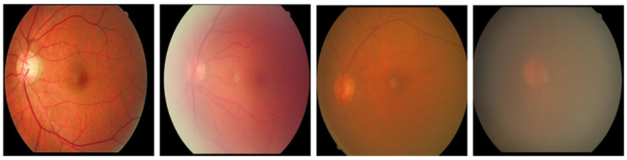

Fig. 2 The result of pre-processed images of cataract severity corresponding to the original image as shown in Fig. 1

3.3 Data augmentation

It is eminent that the performance of DL methods is directly related to the size of the training dataset. Thus, a larger dataset for complex network structure is required to evade generalizing and overfitting problems. Generally, the size of the medical imaging dataset is very small and to address this problem; we apply various data augmentation settings such as rotating, shifting, flipping and cropping of fundus images. First, the retinal images are cropped from the centre and the corner to remove extra areas, then varying the degree of rotation (0°–180°.) and shifting within a given frame of reference which is performed. Finally, flipping is used to augment the training dataset.

4 Methods

The high-resolution fundus images are used as an input of the system to categorize the image into one of four classes: normal, mild, moderate and severe, as given in Fig. 1. We proposed a novel hybrid method, namely CRNN, based on convolutional neural network (CNN) and recurrent neural network (RNN) for automatic cataract classification. The schematic of our model is presented in Fig. 3.

In the training phase, the fundus images are pre-processed before applying the classifier to increase the quality of the images. After pre-processing, the pre-trained models AlexNet [16], GoogLeNet [36], VGGNet [31] and ResNet [13] are fine-tuned. For every image, the trained patch-wise models of the corresponding CNN architecture are employed to excerpt feature representation of 12 patches. Further, these feature vectors are utilized to train image-wise LSTM. In the testing phase, the single fundus image is distributed into twelve smaller patches. Next, the pre-trained models with hyper-tuning adjustments are employed to extract patch-wise features of the fundus image (12*1*2048). The length vector of 2048 feature vector indicates that the corresponding 2048-dimensional feature vector is computed by applying feedforward (1024) and backward (1024) propagation to a trained CRNN model. Then, all the feature vectors extracted individually from the corresponding CNN models are incorporated into the final fully connected layer by using global average pooling (GAP). Finally, the combined feature set

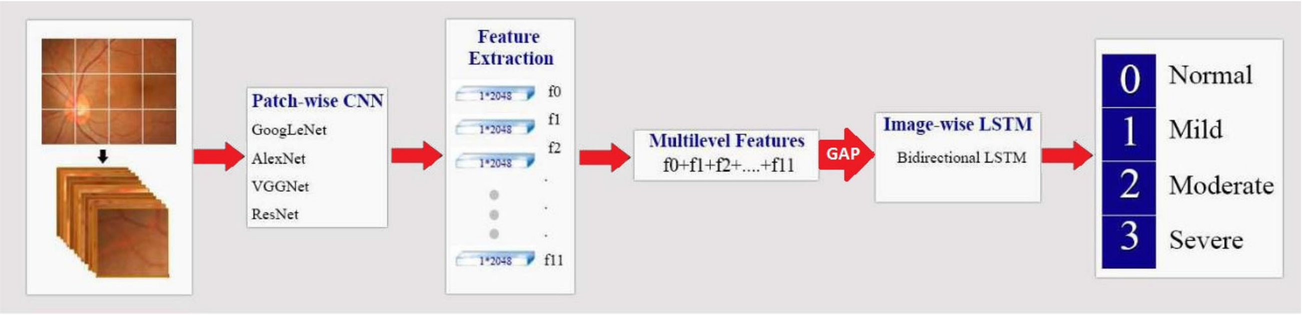

Fig. 3 An architecture of the proposed CRNN hybrid model. First, the fundus image is distributed into 12 patches. Then, each image patch is fed to pre-trained CNN models to extract features. The feature extracted against the corresponding model are combined using global average pooling and termed as 'multilevel features'. Finally, the multilevel feature set is fed into bidirectional LSTM to make an image-wise classification

termed as 'multilevel feature set' is fed into a bidirectional LSTM (BLSTM) to make an image-wise classification. As the proposed CRNN assimilates the benefits of CNN and RNN models, the patches between the long-term and short-term spatial associations are preserved. The details of the proposed methods are given in the subsequent section.

4.1 Patch-wise method

The performance of the CNN models is heavily dependent on the size of the dataset, but it is very problematic to manage a larger dataset in some cases. To overcome such issues, the concepts of transfer learning can be used, which is pre-trained on larger datasets such as ImageNet [16]. Transfer learning-based methods are commonly used methods of ML, which are designed for a specific problem and can be reused it on other relevant problem by fine-tuning the hyperparameters [23]. In transfer learning, the main network is reserved, and the pre-trained weights are utilized to adjust the network. The initialized weights of the network are modified constantly to extract task-related features. The current studies have demonstrated that the transfer learning methods with fine-tuning are efficient for various medical imaging classification tasks like DR detection [12] and skin cancer classification [9]. In this paper, we employed the four most common CNN architectures such as AlexNet, GoogLeNet, VGGNet and ResNet for classifying fundus images. These architectures are discussed in the following subsections.

4.1.1 AlexNet

AlexNet [16] was designed by Alex Krizhevsky and won the completion of ILSVRC (ImageNet Large Scale Visual Recognition Competition) in 2012 with a top-5 error rate of 15.3%. It is comprised of five convolution layers and three fully connected layers along with the Relu activation function that is applied after every convolution and fully connected layers. The dropout value of 0.5 is used before the first two fully connected layers. The network is trained on ImageNet with 100 different categories (1000-way softmax). The underlying AlexNet architecture is given in Fig. 4.

4.1.2 GoogLeNet

GoogLeNet [36] is also known as Inception v1 architecture. It is from Google, inspired by LeNet and implemented the inception module. It is the winner of ILSVRC-2014. GoogLeNet is a deep network comprised of 22 layers, which is pre-trained on the ImageNet database with 1000 object categories. The topmost error rate of 6.67% is obtained on GoogLeNet, which is very near to human-level performance (5.1%). The network is comprised of convolution layers, pooling layers, rectified linear (Relu) layers and fully connected layers. The basic architecture of GoogLeNet is given in Fig. 5.

4.1.3 VGGNet

VGGNet [31] is the project of the visual geometry group (VGG) at Oxford University, which was developed by Simonyan and Zisserman. It was the runner up at the ILSVRC-2014 competition. VGG16 contains 16 convolutional layers with multiple of 3*3 kernel-size filters unlike AlexNet with large kernel size filters (11*11 and 5*5). VGG contains 138 million parameters, which is quite large and slow to train the network. The top-5 error of 6.8% is achieved by the ensemble of the two methods. The VGGNet architecture is shown in Fig. 6.

4.1.4 ResNet

The residual neural network (ResNet) [13] was designed by He et al. and won the ILSVRC-2015 competition. This novel architecture is based on skip connections (grated recurrent

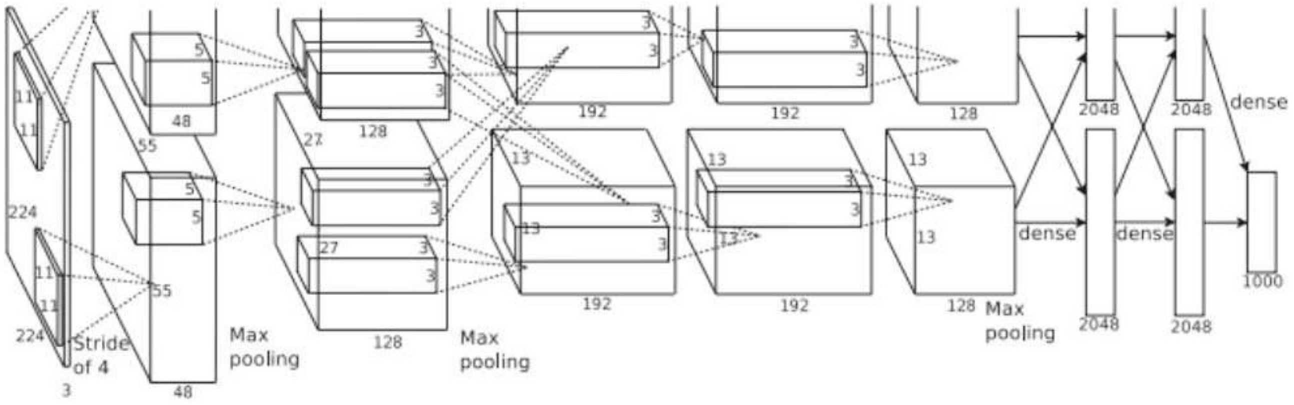

Fig. 4 The basic architecture of AlexNet

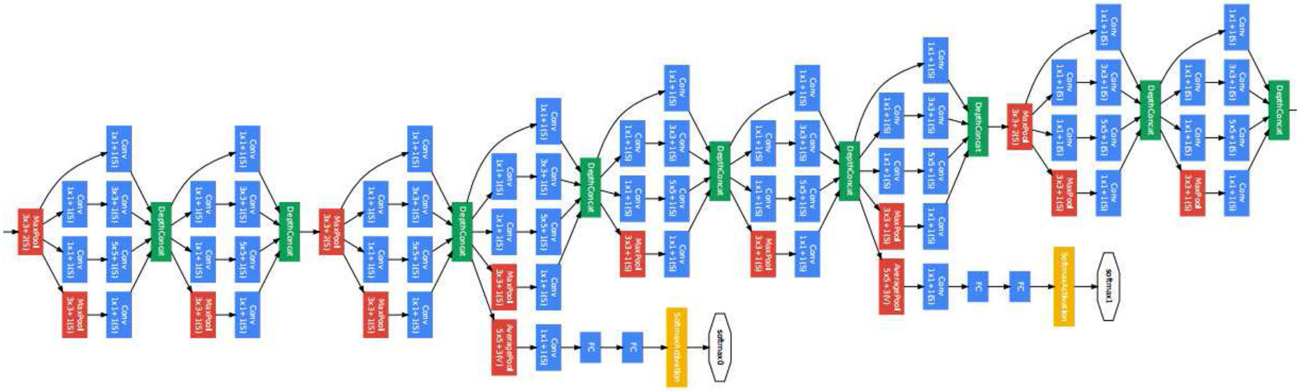

Fig. 5 The basic architecture of GoogLeNet

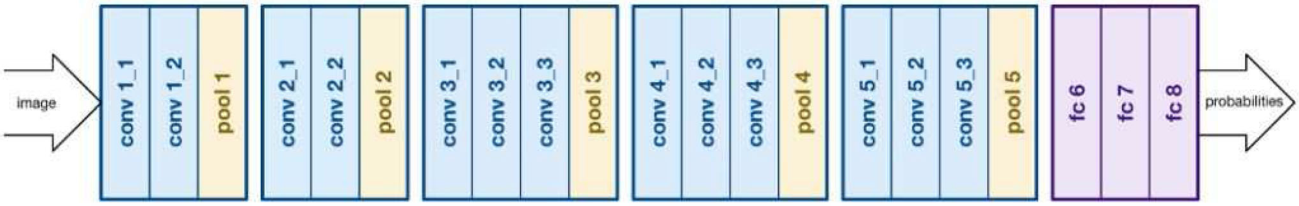

Fig. 6 The basic architecture of VGGNet

units) and batch normalization, which is quite similar to current elements used in RNNs. ResNet is composed of 152 layers and achieved a top-5 error of 3.57% by surpassing human-level performance. The basic ResNet architecture is given in Fig. 7.

4.2 Image-wise method

The medical images are mostly high-quality images, which preserve special information. These images are fed as a whole into DL models for an end-to-end training. Since the image size is very large, the original image is distributed into many smaller patches. The main issue arises when integrating the results of every small patch and then performs image-wise classification. SVM and majority voting are the common techniques associated with this task, which are simple and direct and achieved significant results. However, this method loses contextual information of the image fed as a whole into the network. To preserve the contextual information is the present research focus, and the few methods such as context-aware learning [3] and patch probability fusion methods [2,22] have been proposed to retain the contextual information. The context-aware learning method joints four feature vectors acquired from a patch-wise method and then flattened into a single vector, which has difficulty in bound spatial features. In the patch probability fusion method, the

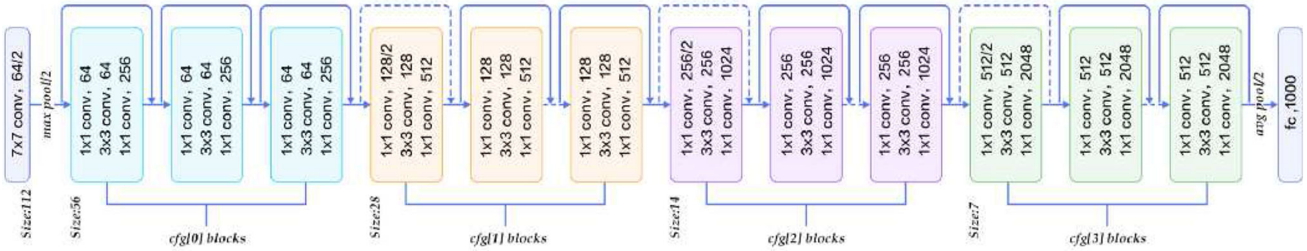

Fig. 7 The basic architecture of ResNet

first patch-wise network is used to extract spatial features and then the image-wise network performs classification. The main drawback of this technique is that the remote contextual information is not preserved.

To solve the aforementioned problems, we used a recurrent neural network for image-wise classification on the top of a CNN to retain the contextual information of features [15]. In our proposed method, the different patch-wise features are extracted by using separate CNN methods, whereas the RNN is used to fuse contextual information by capturing the LSTM dependencies. Moreover, we employed bidirectional long- and short-term memory (BLSTM), which integrates the output of two LSTMs in both the forward and backward directions [19]. BLSTM is the modification of LSTM, which gives the output layer with all the past and future contextual information for every location in the input sequence [29]. This structure proves our stance that there is no difference between left-right or up-down positions for every patch in the fundus image. In our paper, different CNN models are used to extract 12 feature vectors from the fundus image, which are combined in a fully connected layer and then fed to BLSTM. Finally, the image-wise classification is performed for four cataract classes: normal, mild, moderate and severe as shown in Fig. 8.

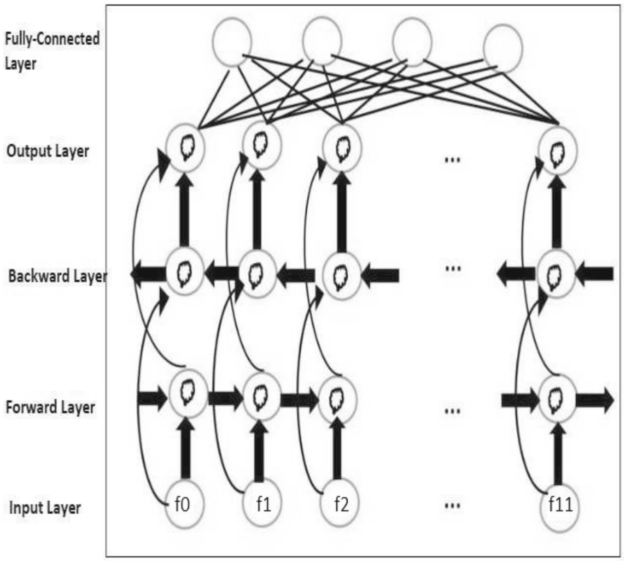

Fig. 8 An overview of bidirectional LSTM: the input layer contains 12 nodes in which data flow in both the forward and backward layers. A fully connected layer with softmax is equal to four output classes. The arrow in this figure represents the data flow, and circles are nodes of the neuron

Table 2 Class distribution of the cataract classification system

Class	Training	Testing
Normal	3737	934
Mild	1826	457
Moderate	540	135
Severe	321	80
Total	6424	1606

5 Results and discussion

5.1 Configuration

All the experimentation was performed on Quadro K620 GPU with 8GB memory, Ubuntu 16 operating system using Keras (http://keras.io/). The retinal dataset was distributed into 80% for training and 20% for testing as shown in Table 2. The transfer learning-based models were fine-tuned with hyperparameters presented in Table 3.

5.2 Metrics

The metrics such as accuracy (ACC), specificity (SP) and sensitivity (SE) are used to measure the performance of the proposed CRNN method, and their mathematical expressions are given in Eqs. 3, 4 and 5, respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

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

Table 3 Hyperparameters configuration

Configuration	Value
Learning rate	0.001
Activation function	Softmax
Epochs	100
Batch size	64
Momentum	0.9
Optimization function	Adam
Dropout	0.5

where TP, FP, TN, FN are the true positive, false positive, true negative and false negative, respectively.

5.3 Result analysis

We selected the four most common pre-trained models (AlexNet, GoogLeNet, ResNet and VGGNet) to reduce the computational cost, which already represents the best learning abilities, whereas combining the feature set extracted from the respective CNN model, we anticipated the classification method to attain optimal performance.

We used different combinations of methods to perform comparative analysis, which is listed in Table 4. First, we applied all these four patch-wise CNN models independently, then ensembled these four methods and applied the two image-wise mainstream methods: majority voting and SVM. Finally, we used the pair combination of ensemble methods and bidirectional LSTM with 4 layers. The experimental results indicate that the proposed method achieves better results in terms of model accuracy and loss as shown in Fig. 9.

The confusion matrix as shown in Fig. 10a displays better performance for the normal, mild and moderate classes while slightly worsening the results in severe class. The average accuracy of the single DL method is 0.944, and the ensemble of patch-wise and traditional image-wise is 0.967, while the proposed hybrid method yields 0.973 of accuracy. The results suggest that the ensemble methods are relatively correct for fundus image classification as compared to single DL methods. It can also be seen in Fig. 10b that the proposed method displays the mean area under the curve (AUC) value of 97%, corresponding to 98%, 98%, 97% and 96% for the four cataract classes based on receiver operator characteristic (ROC) curve analysis. The ROC curve indicates that the results of moderate and severe classes are relatively

Table 4 Performance analysis of different combination of methods

Model	SE	SP	ACC	AUC
1:AlexNet+Fine-tuned+Softmax	0.9332	0.9351	0.9342	0.9402
2:GoogLeNet+Fine-tuned+Softmax	0.9321	0.9782	0.9491	0.9551
3:ResNet+Fine-tuned+Softmax	0.9343	0.9789	0.9521	0.9570
4:VGGNet+Fine-tuned+Softmax	0.9125	**0.9828**	0.9435	0.9476
Ensemble(1,2,3,4)+SVM	0.9574	0.9770	0.9662	0.9672
Ensemble(1,2,3,4)+Majority Voting	0.9644	0.9800	0.9689	0.9701
Ensemble(1,2,3,4)+BLSTM	**0.9764**	0.9770	**0.9739**	**0.9756**

Bold values indicate the outstanding results achieved with the comparison of previous methods

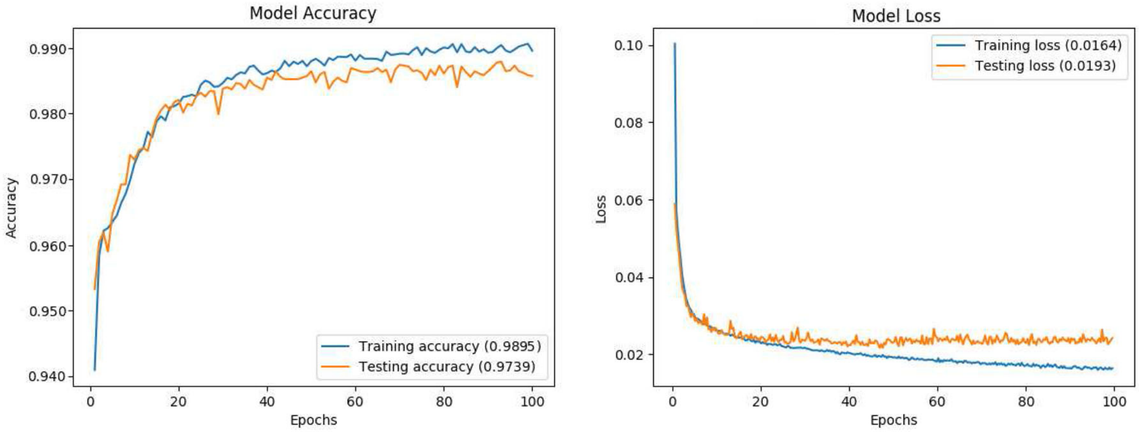

Fig. 9 The accuracy and loss curve indicate the optimal performance and minimal loss of the proposed CRNN model evaluated on 100 epochs. **a** Accuracy curve. **b** Loss curve

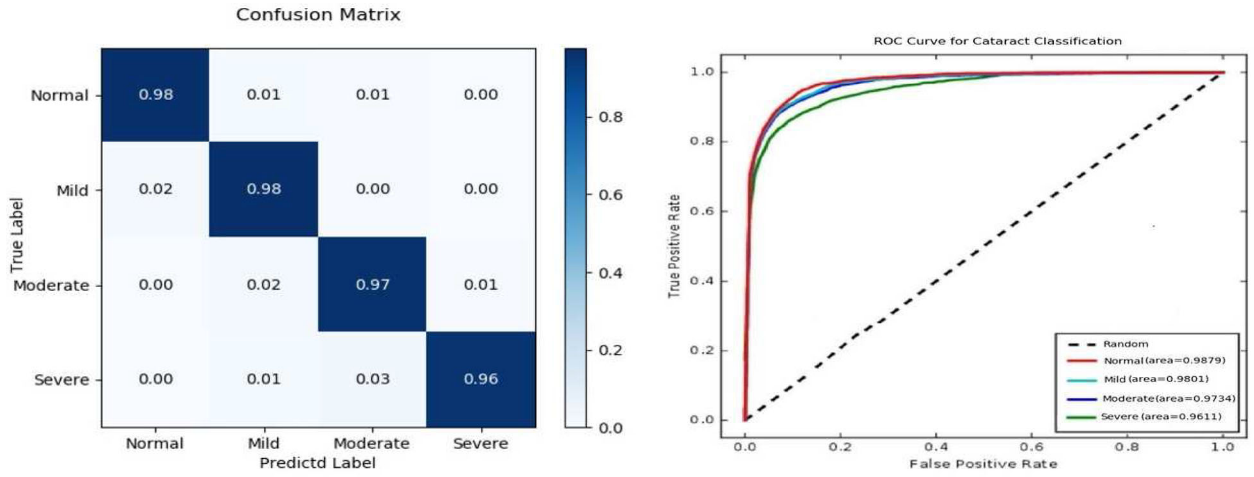

Fig. 10 The confusion matrix of four cataract classes in which diagonal elements are correctly identified labels and the ROC curve of normal, mild, moderate and cataract classes. **a** Confusion matrix. **b** ROC curve

Table 5 Comparative analysis with other methods

Method	Accuracy
Ran [27]	0.9069
Li [18]	0.8770
Xu [39]	0.8624
Li [17]	0.9493
Proposed method	0.9739

low because of the limited number of fundus images of these categories. However, these results are still promising as compared to mainstream methods.

5.4 Comparative analysis with other methods

The results of the proposed method were compared with four well-known methods to relate its strength. It can be observed from Table 5 that the baseline methods [17,18,27,39] give an average accuracy of 91%, 88%, 86% and 95%, respectively, whereas the proposed method yields 97% of accuracy, which is higher than the mainstream methods. These results also demonstrate that our method outperforms others in terms of accuracy.

6 Conclusion

In this paper, we proposed a fundus image-based cataract classification using a hybrid convolutional and recurrent neural network (CRNN) method. The fundus images are pre-processed by using NLMD and normalization techniques and augmented by applying geometric transformations. The four pre-trained CNN models: AlexNet, GoogLeNet, ResNet and VGGNet are used to extract features representation of patches. The feature vectors are extracted from the respective models and integrated into the last fully connected layer using global average pooling and dubbed as a multilevel feature set. This feature set is fed into bidirectional long- and short-term memory RNN to produce image-wise cataract diagnosis. The short-term and long-term spatial correlation among patches are preserved through RNN. The experimental results have verified that our method outperforms the baseline methods in terms of performance. In future, we are intended to improve classification accuracy by evaluating the proposed method on a larger and diverse dataset.

Acknowledgements The authors would like to thank the local hospital (Tongren) for providing data. Further, all procedures performed in studies involving human participants were under the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants involved in the study. The study protocol and consent have been reviewed by the School of Software Engineering, Beijing University of Technology.

Funding The funding was provided by Beijing Municipal Science and Technology (Grand No. KM201910005028)

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abràmoff, M.D., Garvin, M.K., Sonka, M.: Retinal imaging and image analysis. IEEE Rev. Biomed. Eng. **3**, 169–208 (2010)
2. An, F., Liu, Z.: Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM. Vis. Comput. **36**(3), 483–498 (2020)

3. Awan, R., Koohbanani, N.A., Shaban, M., Lisowska, A., Rajpoot, N.: Context-aware learning using transferable features for classification of breast cancer histology images. In: International Conference Image Analysis and Recognition, Springer, pp. 788–795 (2018)
4. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, IEEE, pp. 60–65 (2005)
5. Cao, L., Li, H., Zhang, Y., Zhang, L., Xu, L.: Hierarchical method for cataract grading based on retinal images using improved haar wavelet. Inf. Fusion **53**, 196–208 (2020)
6. Chorage, S., Khot, S.S.: Detection of diabetic retinopathy and cataract by vessel extraction from fundus images. In: 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 1, IEEE, pp. 638–641 (2017)
7. Dong, Y., Wang, Q., Zhang, Q., Yang, J.: Classification of cataract fundus image based on retinal vascular information. In: International Conference on Smart Health, Springer, pp. 166–173 (2016)
8. Dong, Y., Zhang, Q., Qiao, Z., Yang, J.J.: Classification of cataract fundus image based on deep learning. In: 2017 IEEE International Conference on Imaging Systems and Techniques (IST), IEEE, pp. 1–5 (2017)
9. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
10. Fan, W., Shen, R., Zhang, Q., Yang, J.J., Li, J.: Principal component analysis based cataract grading and classification. In: 2015 17th International Conference on E-health Networking, Application and Services (HealthCom), IEEE, pp. 459–462 (2015)
11. Güven, A.: Automatic detection of age-related macular degeneration pathologies in retinal fundus images. Comput. Methods Biomech. Biomed. Eng. **16**(4), 425–434 (2013)
12. Hagos, M.T., Kant, S.: Transfer learning based detection of diabetic retinopathy from small dataset. arXiv:1905.07203 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piece-wise threshold probing of a matched filter response. In: Proceedings of the AMIA Symposium, American Medical Informatics Association, p. 931 (1998)
15. Kabbai, L., Abdellaoui, M., Douik, A.: Image classification by combining local and global features. Vis. Comput. **35**(5), 679–693 (2019)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
17. Li, J., Xie, L., Zhang, L., Liu, L., Li, P., Yang, J.j., Wang, Q.: Interpretable learning: a result-oriented explanation for automatic cataract detection. In: International Conference on Frontier Computing, Springer, pp. 296–306 (2018)
18. Li, J., Xu, X., Guan, Y., Imran, A., Liu, B., Zhang, L., Yang, J.J., Wang, Q., Xie, L.: Automatic cataract diagnosis by image-based interpretability. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, pp. 3964–3969 (2018)
19. Liang, D., Liang, H., Yu, Z., Zhang, Y.: Deep convolutional bilstm fusion network for facial expression recognition. Vis. Comput. **36**(3), 499–508 (2020)
20. Manchalwar, M., Warhade, K.: Detection of cataract and conjunctivitis disease using histogram of oriented gradient. Int. J. Eng. Technol. (IJET) (2017)
21. Nayak, J., Acharya, R., Bhat, P.S., Shetty, N., Lim, T.C.: Automated diagnosis of glaucoma using digital fundus images. J. Med. Syst. **33**(5), 337 (2009)
22. Nazeri, K., Aminpour, A., Ebrahimi, M.: Two-stage convolutional neural network for breast cancer histology image classification. In: International Conference Image Analysis and Recognition, Springer, pp. 717–726 (2018)
23. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
24. Pascolini, D., Mariotti, S.P.: Global estimates of visual impairment: 2010. Br. J. Ophthalmol. **96**(5), 614–618 (2012)
25. Pizzarello, L., Abiose, A., Ffytche, T., Duerksen, R., Thulasiraj, R., Taylor, H., Faal, H., Rao, G., Kocur, I., Resnikoff, S.: Vision 2020: The right to sight: a global initiative to eliminate avoidable blindness. Arch. Ophthalmol. **122**(4), 615–620 (2004)
26. Qiao, Z., Zhang, Q., Dong, Y., Yang, J.J.: Application of SVM based on genetic algorithm in classification of cataract fundus images. In: 2017 IEEE International Conference on Imaging Systems and Techniques (IST), IEEE, pp. 1–5 (2017)
27. Ran, J., Niu, K., He, Z., Zhang, H., Song, H.: Cataract detection and grading based on combination of deep convolutional neural network and random forests. In: 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), IEEE, pp. 155–159 (2018)
28. Röhlig, M., Schmidt, C., Prakasam, R.K., Rosenthal, P., Schumann, H., Stachs, O.: Visual analysis of retinal changes with optical coherence tomography. Vis. Comput. **34**(9), 1209–1224 (2018)
29. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
30. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. **19**, 221–248 (2017)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
32. Somasundaram, S., Alli, P.: A machine learning ensemble classifier for early prediction of diabetic retinopathy. J. Med. Syst. **41**(12), 201 (2017)
33. Song, W., Cao, Y., Qiao, Z., Wang, Q., Yang, J.J.: An improved semi-supervised learning method on cataract fundus image classification. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), vol. 2, IEEE, pp. 362–367 (2019)
34. Souza, M.B., Medeiros, F.W., Souza, D.B., Garcia, R., Alves, M.R.: Evaluation of machine learning classifiers in keratoconus detection from orbscan ii examinations. Clinics **65**(12), 1223–1228 (2010)
35. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
36. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
37. Xiong, L., Li, H., Xu, L.: An approach to evaluate blurriness in retinal images with vitreous opacity for cataract diagnosis. J. Healthc. Eng. **2017** (2017)
38. Xiong, Y., He, Z., Niu, K., Zhang, H., Song, H.: Automatic cataract classification based on multi-feature fusion and SVM. In: 2018 IEEE 4th International Conference on Computer and Communications (ICCC), IEEE, pp. 1557–1561 (2018)
39. Xu, X., Zhang, L., Li, J., Guan, Y., Zhang, L.: A hybrid global-local representation CNN model for automatic cataract grading. IEEE J. Biomed. Health Inform. (2019)
40. Yang, J.J., Li, J., Shen, R., Zeng, Y., He, J., Bi, J., Li, Y., Zhang, Q., Peng, L., Wang, Q.: Exploiting ensemble learning for automatic cataract detection and grading. Comput. Methods Programs Biomed. **124**, 45–57 (2016)
41. Zhang, L., Li, J., Han, H., Liu, B., Yang, J., Wang, Q., et al.: Automatic cataract detection and grading using deep convolutional neural network. In: 2017 IEEE 14th International Conference

on Networking, Sensing and Control (ICNSC), IEEE, pp. 60–65 (2017)

42. Zheng, J., Guo, L., Peng, L., Li, J., Yang, J., Liang, Q.: Fundus image based cataract classification. In: 2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings, IEEE, pp. 90–94 (2014)

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

Azhar Imran is a PhD student at Beijing University of Technology, Beijing, China. He received his M.S. degree in Computer Science and B.S. degree in Software Engineering from University of Sargodha, Pakistan, in 2016 and 2012, respectively. From 2012 to 2017, he worked as a lecturer with the Department of Computer Science, University of Sargodha, Pakistan. His research interest includes machine learning, image processing, medical imaging and data mining. Mr. Imran awards and honours include the China Scholarship Council (CSC, China), and the star contribution award for best researcher (Beijing University of Technology, China).

Jianqiang Li received his B.S. degree in Mechatronics from Beijing Institute of Technology, Beijing, China in 1996, and M.S. and PhD degrees in Control Science and Engineering from Tsinghua University, Beijing, China in 2001 and 2004, respectively. He joined Beijing University of Technology, Beijing, China, in 2013 as Beijing Distinguished Professor. His research interests are in data mining, information retrieval and big data. He has over 40 publications including 1 book, 10+ journal papers and 37 international patent applications. He served as Guest Editor to organize a special issue on information technology for enhanced healthcare service in Computer in Industry.

Yan Pei obtained Doctor of Engineering at Kyushu University, Fukuoka, Japan. He received the B.Eng. and M.Eng. Degree from Northeastern University, Shenyang, China. He is currently working at the University of Aizu as an Associate Professor. His research interests include evolutionary computation, machine learning and software engineering.

Faheem Akhtar received the M.S. degree in computer science from the National University of Computing and Emerging Science, NUCES FAST Karachi, Pakistan. He is currently an Assistant Professor with the Department of Computer Science, Sukkur IBA University. Meanwhile, he is on study leave from Sukkur IBA to pursue his PhD degree with the School of Software Engineering, Beijing University of Technology, Beijing, China, from 2016 to 2020. He is the author of various SCI and EI journals. His research interests include data mining, machine learning, information retrieval, privacy protection, Internet security, the Internet of Things and big data.

Tariq Mahmood received his master's degree from the Department of Computer Science, University of Lahore, Lahore, Pakistan. Currently, he is pursuing his PhD at The Beijing University of Technology, China, under the supervision of Prof. Dr. Li Jianqiang. His research interests include software engineering, machine learning, deep learning, data mining, big data, pattern recognition, wireless sensor network and cloud computing

Li Zhang received her B.S. degree in Clinical Medicine Science from Tianjin Medical University, Tianjin China in 1997. Then, she worked in the Beijing Moslem People's Hospital as a Physician until now. Her research interests include respiratory, gastroenterology and cardiology medicines.