Methods xxx (xxxx) xxx

Contents lists available at ScienceDirect

Methods

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

Glaucoma screening using an attention-guided stereo ensemble network

Yuan Liu [a], Leonard Wei Leon Yip [b], Yuanjin Zheng [a], Lipo Wang [a,*]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[b] Department of Ophthalmology, Tan Tock Seng Hospital, Singapore

ARTICLE INFO

Keywords:
Computer-aided screening and diagnosis
Neural network
Glaucoma
Deep learning
Stereoscopy

ABSTRACT

Glaucoma is a chronic eye disease, which causes gradual vision loss and eventually blindness. Accurate glaucoma screening at early stage is critical to mitigate its aggravation. Extracting high-quality features are critical in training of classification models. In this paper, we propose a deep ensemble network with attention mechanism that detects glaucoma using optic nerve head stereo images. The network consists of two main sub-components, a deep Convolutional Neural Network that obtains global information and an Attention-Guided Network that localizes optic disc while maintaining beneficial information from other image regions. Both images in a stereo pair are fed into these sub-components, the outputs are fused together to generate the final prediction result. Abundant image features from different views and regions are being extracted, providing compensation when one of the stereo images is of poor quality. The attention-based localization method is trained in a weakly-supervised manner and only image-level annotation is required, which avoids expensive segmentation labelling. Results from real patient images show that our approach increases recall (sensitivity) from the state-of-the-art 88.89% to 95.48%, while maintaining precision and performance stability. The marked reduction in false-negative rate can significantly enhance the chance of successful early diagnosis of glaucoma.

1. Introduction

Glaucoma is one of the leading causes of blindness [1]. It gradually damages the optic nerve, which results in vision loss. The major risk factor for glaucoma is elevation of the Intraocular Pressure (IOP) due to blockage of the eye drainage channels [2]. Although existing vision loss cannot be restored, further vision loss can be effectively prevented if proper treatment is applied at an early stage. Therefore, it is of great significance to detect glaucoma early and take necessary measures to mitigate the disease. For IOP measurement, a tonometer [3] is used to detect if the pressure is above a normal range, suggesting a risk of developing or having glaucoma. However, it is not uncommon for a glaucoma patient to also have a normal IOP in normotensive glaucoma. Fundus imaging is widely used in glaucoma diagnosis. Clinicians typically perform optic nerve head analysis on the fundus image, calculating measurement metrics to determine the glaucoma cases. The vertical cup to disc ratio (CDR) [4] is a very popular metric for glaucoma screening, which is computed by dividing the vertical cup diameter by the vertical disc diameter. However, according to Jonas et al. [5], the CDR is affected by inter-individual variabilities. Therefore, eyes with a high CDR should not be considered glaucomatous without further studying other features, such as the population and disc sizes. As shown by Hagiwara et al. [6], computer-aided detection systems are effective in glaucoma screening and can alleviate clinician's workload. In addition, algorithm-based feature engineering is normally able to extract more features from an image apart from CDR itself. Hence, machine learning feature extraction and classifiers have attracted significant research interest.

As the optic disc region is commonly believed to be crucial to the glaucoma screening task, many methods have been proposed to localize the disc and perform prediction based on disc features. Typically, these methods identify the region of interest, especially optic disc and cup region, and calculate related clinical measures like CDR to detect glaucoma. An example is the method proposed by Poshtyar et al. [7], where masks and thresholds were used to locate the optic disc and cup, followed by CDR calculations. In a paper by Atheesan and Yashothara [8], the cup and disc were extracted using average and maximum grey level pixels with histograms, the CDR was then calculated, and blood vessel information was utilized to aid diagnosis. In order to alleviate the limitation of CDR, Cheng et al. [9] cropped optic disc region and extracted various features with CNN. Ferreira et al. [10] extracted optic disc features with segmentation. However, it is costly to obtain pixel-level

* Corresponding author.
E-mail addresses: yliu050@e.ntu.edu.sg (Y. Liu), leonard_yip@ttsh.com.sg (L.W.L. Yip), yjzheng@ntu.edu.sg (Y. Zheng), elpwang@ntu.edu.sg (L. Wang).

https://doi.org/10.1016/j.ymeth.2021.06.010
Received 4 April 2021; Received in revised form 9 June 2021; Accepted 16 June 2021
Available online 19 June 2021

segmentation annotations. Furthermore, as shown by Jonas et al. [4], completely ignoring features from regions outside the optic disc has negative effects on robustness of the model. In recent years, attention mechanisms have been introduced to locate regions of interest in images, while preserving global information in other image regions. Using attention in glaucoma screening intuitively satisfies the need to focus on the optic disc region while keeping tissue and vessel features. Oktay et al. [11] applied visual attention mechanism on CT images. Li et al. [12] applied attention mechanisms to glaucoma diagnosis, however, human labelled ground truth of attention masks required in [12] is costly and subject to annotator bias. Hence, automatically learning attention with only image-level labels is of great importance and beneficial to glaucoma screening.

In [15] by Ruengkitpinyo et al, the optic disc was located using eclipse fitting, then the rim width was obtained using the INST Rule, and the CDR was calculated and used as a feature together with the rim width as input for a support vector machine. In [16] by Agarwal et al, a histogram with adaptive thresholding was used to segment the disc and cup and to calculate CDR for glaucoma screening. Jose and Balakrishnan [17] took advantage of morphological operations and blood vessel information in the optic disc and cup segmentation, before the CDR was calculated. In [18] by Alghmdi et al, super-pixels classification was performed before utilizing morphological operations to finalize cup and disc boundary. In [19] by Cheng et al, hand crafted features were fed into a super-pixel-based classifier to segment the optic disc and cup regions. The segmented region was then used for CDR calculation to detect glaucoma. Dey et al. [20] used the Harris Corner detector to draw circles over optic disk and cup regions to aid segmentation. Kande et al. [21] used color morphology with active contours. In [22] by de Carvalho et al, Otsu and K-means were employed to localize disc regions and phylogenetic diversity indices were then utilized for feature extraction. Aquino et al. [23] also used morphology with the Hough Transform for edge screening and disk boundary extraction.

On the other hand, methods that automatically extract image features and perform classification based on these features are also widely studied. David and Jayachandran [25] used hybrid color and structure descriptors to perform feature selection [58–63] for classification. Araújo et al. [26] utilized diversity indexes as texture descriptors for screening. Salam et al. [27] adopted texture and intensity features to detect glaucoma. Acharya et al. [28] applied the Gabor transform to fundus images. Noronha et al. [29] extracted higher order spectra cumulants by the Radon transform. In [30] by Dua et al, energy features were extracted using 2D discrete wavelet filters. In [31] by Sousa et al, binary patterns were used to represent disc regions and geostatistical functions were applied to represent texture patterns, and the features were then used by a support vector machine for glaucoma screening. Xiong et al. [32] applied principal component analysis to obtain eigenvector spaces for normal and glaucoma sets, and then projected test images to these spaces to detect glaucoma. In [33] by Bai et al, the optimal error-correcting output code matrix was learned to screen glaucoma through ensembles. Nayak et al. [36] extracted various features, such as CDR and ratio of the distance between optic disc center and optic nerve head to diameter of the optic disc, and performed screening with a neural network classifier. A number of researchers have studied detection and grading of diabetic retinopathy in retinal images using various machine learning techniques [53–57].

With the advancement in Deep Learning, specially-designed or hand-crafted features are generally less robust compared to features extracted from convolutional neural networks (CNNs). The top performing learning-based models are mostly based on Deep Learning or CNNs, which extract highly discriminative features. According to Ker et al. [34,51,52], CNNs are well-suited for medical image classification. Fu et al. [9] and Edupuganti et al. [24] used CNN to segment optic discs and cups jointly at the same time. Ferreira et al. [10] employed CNN to perform disc segmentation, then extracted features from disc regions with vessels removed, before feeding the features to a CNN-based classifier. Chen et al. [35] utilized CNN with fully connected dense layers for both feature extraction and glaucoma screening. Orlando et al. [37] used image pre-processing and pre-trained CNN for glaucoma screening. Pal et al. [38] used autoencoders to compress CNN features for screening. Al-Bander et al. [39] employed CNN features in a support vector machine for classification. Li et al. considered [40] the region of interest together with grid patches. Lima et al. [41] compared mainstream CNN architectures as feature extractors. Norouzifard et al. [42] applied transfer learning and pre-trained CNN models to perform end-to-end glaucoma screening. In [43] by Li et al, features extracted by multiple CNNs were combined to boost performance. However, these CNN methods generally considered all image regions equally, instead of prioritizing regions of interests and suppressing noises in regions outside of the optic disc. Fu et al. [44] considered both local and global image information, together with polar transformation of the local region to improve performance through ensembles. Chai et al. [45] proposed a multi-branch neural network which extracts global and disc features simultaneously. However, these methods focused on only single 2D fundus images, which might not contain enough information to further improve glaucoma screening performance.

Stereo images, taken as image pairs by two cameras from different viewpoints, have been introduced in medical applications. A pair of stereo images contain more information compared to a single image. Hence, utilizing both left and right images in a stereo pair to perform glaucoma screening can potentially improve the robustness of the screening model, especially when one image is of poor quality, the other image can usually compensate the loss and maintain a reliable performance. Nakagawa et al. [13] obtained depth in stereo images by calculating location differences between corresponding points; however, they [13] did not use the depth information in glaucoma screening, since the depth obtained heavily depends on the selection accuracy of the corresponding points. Clinically, judgement of the optic disc is important in the diagnosis of glaucoma. In general, ophthalmologists classify stereoscopic optic disc photographs moderately well for glaucoma. However, there can be large variability in diagnosis accuracy based on clinician judgement [14]. The need for an automatic technique for glaucoma screening was therefore raised to address this issue. Corona et al. [46] developed an algorithm to generate abundant glaucoma measures from stereo images that can guide clinical judgement. Norouzifard et al. [47] generated disparity maps using stereo images and performed segmentation on the disparity maps to help clinicians analyze abnormalities. While these methods suggested potential applications of stereo images in glaucoma CAD, they mainly facilitated types of analytics other than prediction and they did not provide algorithms or experiments on automatic glaucoma screening.

In this paper, we propose a novel attention-guided stereo ensemble network (AGSE-Net) using stereo fundus image pairs. It automatically locates optic disc regions through an attention module without the need of ground-truth disc region labels (or attention mask) and can be trained in an end-to-end manner.

The main contributions of this paper are as follows:

1. We propose a novel technique for glaucoma optic disc region localization through a visual attention mechanism, which focuses on the disc region while keeping global information (including regions outside the optic disc).
2. We show the effectiveness and advantages of using stereo images for glaucoma screening. The two images in a stereo pair compensate each other, especially when one of the images is of poor quality.
3. We design a fully automatic end-to-end network that performs localization and screening at the same time, which is supervised by disease labels only and avoids the additional effort for disc region/attention mask labelling.

2. Methods

The overall structure of our proposed network is shown in Fig. 1, which consists of two types of sub-components, namely the Deep CNN and the Attention-Guided network. The Deep CNN captures image features at a global scale to obtain holistic information. The Attention-Guided network localizes the optic disc while maintaining beneficial information in other regions. Both the stereo images are fed into the two modules to produce abundant features for classification. We fuse the information from the original images and the attention-guided local regions by combining the outputs from the all sub-networks during inference.

2.1. The deep CNN

Several types of CNNs have been proposed and applied with varying degrees of success [64–68]. We select the ResNet-34 [48] as our CNN feature extractor in this paper. One of the common issues of training a deep neural network is the vanishing gradient problem. With the increasing number of layers, the gradient tends to zero when the back-propagation process goes to earlier layers. The ResNet structure provides a feasible solution to this issue. It consists of many residual blocks, which are composed of a few stacked convolutional layers with a short cut connection from the first layer to the last layer. Through this process, the gradient can "by-pass" a few layers and directly reach those earlier layers to push parameter learning. Lima et al. [41] compared the performance of ResNet-50 with other CNN models and showed advantages of the ResNet architecture in glaucoma screening. We adopt the light-weight ResNet-34 instead of ResNet-50 in order to reduce computational resources needed during clinical applications, while maintaining a high screening accuracy. The ResNet-34 structure as the backbone feature extractor is connected to a fully connected (FC) classifier. In order to reduce overfitting, two dropout layers are inserted before each FC layers in the classifier to regularize the model.

2.2. Attention network

The signals of glaucoma-affected eyes are mainly contained in the optic disc region. Therefore, specifically extracting features from this region can reduce image noise and potentially improve network performance. However, many state-of-the-art methods for disc region segmentation require ground-truth label for the optic disc region, which is costly to obtain. Inspired by the work in [49] that utilizes deep learning based attention mechanism to automatically focus on important feature region, we apply a visual attention method with convolutional layer to automatically locate the disc regions and highlight the glaucoma signals. We design and insert our attention module on the output from the third residual block, which is named as the pre-attention layer, of the ResNet-34 feature extractor. The structure of the attention network is shown in Fig. 2.

The attention module consists of a convolutional layer with kernel size of 1×1 and the number of channels 1 and a sigmoid operation. By applying convolutional operation with the attention module onto the pre-attention layer, a 2D output of the same height and width as the pre-attention layer can be obtained. We then apply a sigmoid layer onto the 2D output to highlight the important features and compress the irrelevant features. Lastly, we perform an element-wise multiplication between the 2D mask and the pre-attention layer on all pre-attention layer channels. As the disc region features contribute more to the glaucoma detection performance, the attention mask will automatically learn to focus on the disc features during back-propagation supervised by the classification loss. Some visualizations of the learned attention masks are presented in later sections.

The following equations show the mathematic operations behind the attention module.

$$F' = F \odot \{S(A(F, \theta))\} \tag{1}$$

$$S(x) = \frac{e^x}{e^x + 1} \tag{2}$$

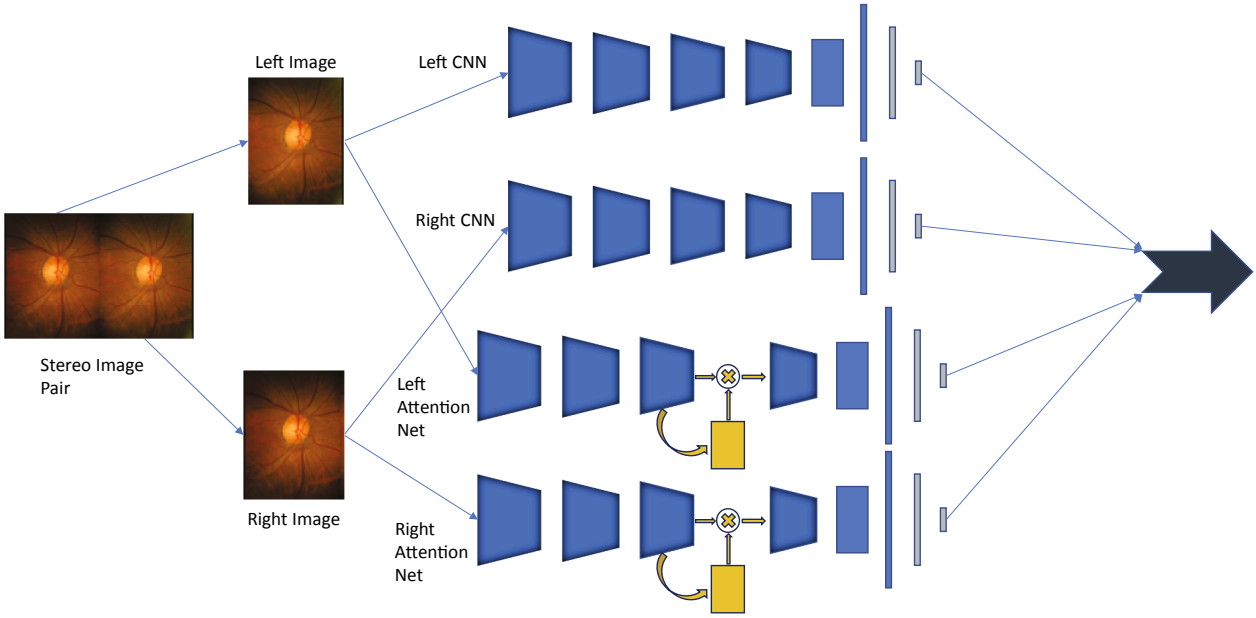

Fig. 1. An overview of our attention-guided stereo ensemble network (AGSE-Net). The elements in blue represents ResNet CNN backbone for feature extraction. The elements in grey represent the classifier modules in the CNNs. The elements in yellow represent visual attention localization modules. The large arrow at the right end of the figure decides the overall output of the ensemble based on the output of each of the four branch networks. In the present work, the output of each branch represents the likelihood for positive or negative in glaucoma screening and the highest value determines the screening outcome. The input stereo image pair is first split into left and right images, each of which is fed into the respective CNN and attention net, and the results are then combined as an ensemble at the end to calculate prediction scores. For both the CNN and the Attention Network, the Global Average Pooling layer after the convolution layers are maintained. ResNet-34 [48] is selected as the CNN in this paper, due to its effective error-propagation, excellent performance, and yet reasonable computational load. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

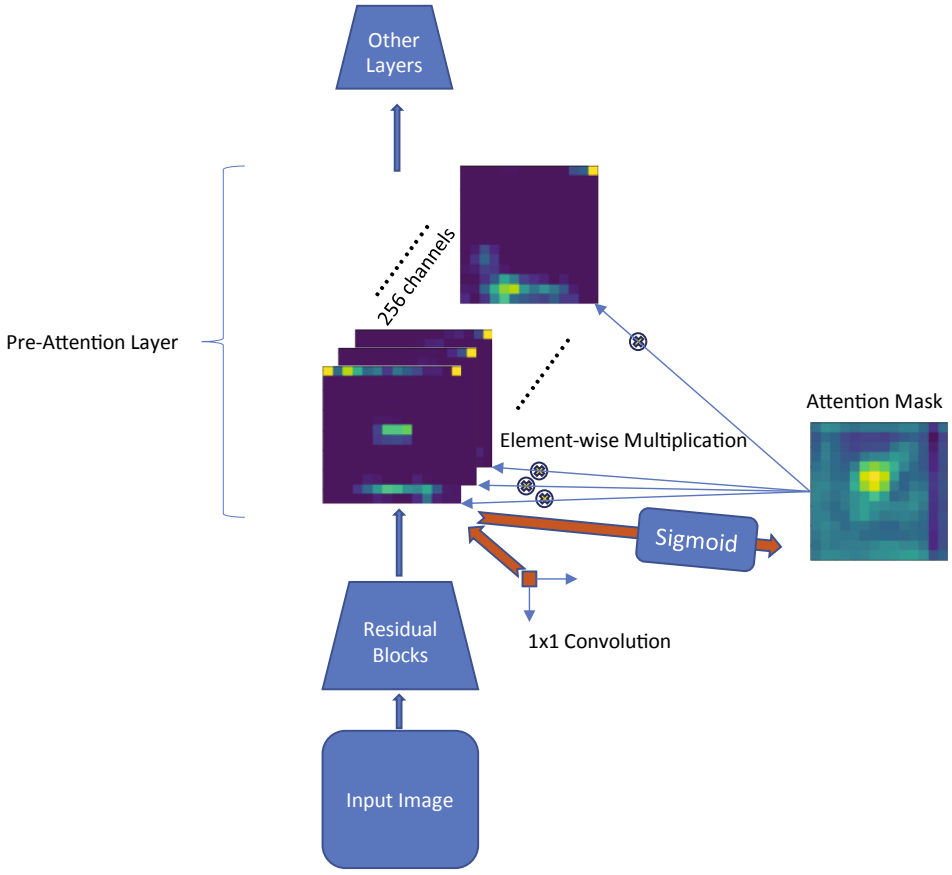

Fig. 2. The structure of the attention network.

$$A(x) = \theta_B + \sum_{j=1}^{C} \theta_W * x \tag{3}$$

where$A$ represents the attention module's convolutional layer with parameters θ, S represents the sigmoid operation, $\odot$ represents the element-wise multiplication and $F \in R^{H\times W\times C}$represents the feature map (pre-attention layer) output from the first three residual blocks of the ResNet-34 with shape $H \times W \times C$. $F' \in R^{H\times W\times C}$ is the disc highlighted features generated by the attention module and pre-attention layer, which is then fed into the rest of the layers in our network.

As optic disc region features are beneficial for improving the glaucoma screening accuracy and reducing training loss, the attention kernel automatically learns to focus on the optic disc region without ground-truth of that region (e.g., bounding boxes and bounding curves).

In addition, the kernel assumes relatively small, instead of completely zero, values for non-optic disc regions. As shown in [4], keeping global information is beneficial to improve model performance. Hence, by focusing on the disc region while keeping global information, the attention kernel ensures the stability of the model and improves its robustness. Although it is plausible to consider that the local features extracted by the attention net are already included in the global features extracted by the CNN, the local features could focus or enhance weightage on some important local features, thereby offering potential for improved classification accuracy. This conjecture will be shown to be valid by the subsequent experiments.

2.3. Model ensemble

We aggregate the results from all our network modules together, combining information from original images and attention-guided information, as well as fusing information from left and right images in a stereo pair. In our network, the output of each branch represents the likelihood for positive or negative in glaucoma screening. Finally, the highest value determines the screening outcome.

3. Experiment details & results

3.1. Evaluation criteria

As the dataset is biased towards the normal case (without glaucoma or negative case), directly applying the naïve accuracy (the percentage of correct predictions) as an evaluation metric may not be the most sensible. Hence, we adopt precision, recall, and the average precision (AP) as the main performance metrics, with the naïve accuracy as reference. Precision (also called positive predictive value or PPV) is the fraction of true positive instances (correctly classified as positive) among all instances classified as positive, whereas recall (also known as sensitivity, hit rate, true positive rate or TPR) is the fraction of true positive instances (correctly classified as positive) among all positive instances in the dataset.

$$Precison = \frac{TP}{TP + FP} \tag{4}$$

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

$$Avergae\ Precision = \sum_{n}(R_n - R_{n-1})P_n, n \in [1, k] \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP is the number of true positives (correctly classified positives),

FP is the number of false positives (incorrectly classified positives), *TN* is the number of true negatives (correctly classified negatives), and *FN* is the number of false negatives (incorrectly classified negatives). In equation (6), R_n represents the recall for the n^{th} sample, P_n represents the precision for the n^{th} sample, and k is the number of samples. Thus, average precision is the weighted average of precision at every decision, with the increment of recall from the last decision as the weight.

We regard naïve accuracy given in equation (7) as the least important metric and list it only for reference purpose, as accuracy may not be significant for a biased dataset (imbalanced in glaucoma and normal subjects).

There exist other evaluation metrics. For example, specificity (also called selectivity, true negative rate or TNR:

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

But we are less concerned with negative results than positive results in screening.

An alternative representation of the results would be a confusion matrix:

TP	*FP*
FN	*TN*

This representation is quite intuitive if the computer experiment is run only once. If the computer experiments are run multiple times, we could still enter the averages and the standard deviations in the confusion matrix, if the data split is done only once and fixed in all experiments. However, if the data split is done randomly multiple times, as in this paper, the averages in the confusion matrix may not make sense, since we are not sure what the correct values would be. Hence, we will not show the confusion matrix in this paper.

3.2. Data preprocessing & augmentation

The stereo glaucoma image dataset is provided by Tan Tock Seng Hospital, Singapore. The data are annotated by a glaucoma fellowship trained ophthalmologist according to the international gold standard. The dataset contains a total of 282 images with 70 glaucoma cases and 212 normal cases. The stereo images are taken during patient examinations, i.e., each left and right images are taken separately and are not the same.

Fig. 3 is a sample stereo image pair in the dataset. The original sizes of the left and right images are about 1000 × 1400 pixels. All the images are then resized into 224 × 224 for training and testing.

We randomly split the entire data set to 70% as training data and 30% as test data. We split for 3 times. We train 10 models for each training/test set and record the averages and standard deviations for AP, precision, recall, and accuracy.

In order to improve model performance and robustness, data augmentation is used to generate more training samples. We apply rotation (90°/270°) and horizontal flip to all training images. There are two main reasons for us to adopt rotation augmentation though it may seem counter-intuitive in this scenario. First, it has been shown that features other than CDR can also contribute to model learning. Second, we hope to encourage the deep learning model to explore rotational invariant features in the image. Third, although the optical discs are almost circular, they are not exactly circular. In addition, the entire images may not be centered at the centers of the optical discs. Therefore, rotations will generate different images. Hence, by applying these rotations, the model has to learn robust features instead of depending only on features like vertical CDR.

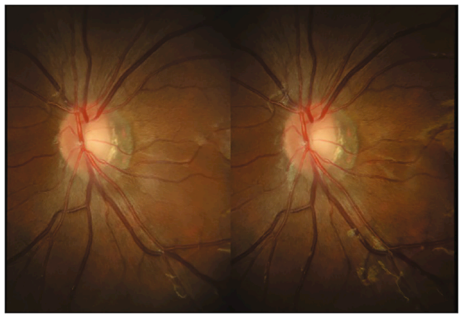

Fig. 3. A stereo image pair.

3.3. Implementation details

The network shown in Fig. 1 is implemented using Python with PyTorch. During training, the Adam optimizer is utilized to update model parameters. We initialize the model to the ImageNet pre-trained weights. A warmup of 5 epochs is applied to stabilize the training process. We adopt a learning rate decay scheme of reducing the learning rate by half after every 10 epochs. The base learning rate is set to 0.001.

3.4. Results

The experimental results are described in two parts. We first compare our method to the re-implementations of other state-of-the-art learning-base methods. We then perform ablation studies to show the effectiveness of using attention mechanism as well as utilizing stereo images. We highlight the top 1st and 2nd performance on all the metrics.

4. Comparisons with other methods

We select five state-of-the-art learning-based methods for fair comparisons. As the original models were trained on other datasets, we implement these methods and apply them to our data. The methods to be compared are:

- A CNN [35] with overlapping pooling.
- An 18-layer deep CNN [50].
- CNN-SVM [39] which extracts features using pre-trained CNN (ResNet) and perform screening using the Support Vector Machine.
- Inception-ResNet-V2 [42] which leverages on transfer learning and deep CNN architecture to perform end-to-end glaucoma screening.
- VGG-19 [42] which is similar to the Inception-ResNet-V2 method but with a different model architecture.

As shown in Table 1, our proposed method yields the highest precision and recall, while maintaining top-2 smallest standard deviations (indicating stable performance). In particular, we increased the recall with a significant margin from 88.89 by Inception-ResNet-V2 to 95.48 using the proposed AGSE-Net. Our proposed method archives the

Table 1
Comparisons with Other Methods (top results are high-lighted in bold).

Method	Precision (Mean/ Standard Deviation)	Recall (Mean/ Standard Deviation)	Average Precision (Mean/ Standard Deviation)	Accuracy (Mean/ Standard Deviation)
CNN [35]	85.30/9.80	69.23/14.28	86.48/3.95	93.20/1.65
DCNN [49]	82.27/15.67	80.34/8.36	92.45/4.75	92.94/5.02
Inception-ResNet-V2 [42]	93.74/**0.35**	88.89/**1.48**	94.51/**0.46**	**97.52/0.23**
VGG-19 [42]	91.41/4.70	78.63/12.71	94.57/2.56	95.42/2.12
CNN-SVM [39]	79.45/12.17	86.60/6.20	92.72/5.50	91.33/**0.74**
AGSE Net (Ours)	**94.72/1.52**	**95.48/3.66**	**95.23/1.18**	**97.12/0.96**

highest AP score, while maintaining top-2 smallest standard deviation, in comparisons with other state-of-the-art approaches. Excellent and stable performance is especially important in medical diagnosis, and our proposed method satisfies this requirement.

4.1. Ablation study

In order to show the improvement from the attention mechanism and stereo images separately, we also perform ablation studies with different combinations of the four network branches in Fig. 1.

4.2. Effects of Attention:

In Fig. 1, the difference between the Left CNN branch and the Left Attention Net branch lies in the added attention module in the Left Attention Net (and similarly for the Right CNN branch and the Right Attention Net branch). In Table 2, when we compare *Left CNN Only* vs. *Left CNN + Left Attention Net*, and *Right CNN Only* vs. *Right CNN + Right Attention Net*, we see that all four performance metrics increase, while the standard deviations decrease with the added attention modules. It is notable that *Attention Net* alone does not provide performance improvement over *CNN Net* for both single branch and left–right ensemble scenario. This is reasonable as the attention mechanism adopted in our network mainly provides fine-grained focus not emphasized by the usual CNN branch. Hence, its benefit may be significant only when fusing together with the CNN branch.

We visualize some attention mask in Fig. 4, in which the optic disc region and nearby related tissues are highlighted by the learned attention masks. Even for fundus images with very poor lighting and resolution, the attention mask is able to locate the region of interest, as shown in Fig. 4.

4.3. Effects of stereo Images:

Table 2 shows that the combination of *Left CNN Only* and *Right CNN Only* outperforms its two sub-components by a significant margin. Similarly, the combination of *Left Attention Net* and *Right Attention Net* significantly outperforms its two sub-components, while the combination of *Left CNN + Left Attention Net* and *Right CNN + Right Attention Net* also outperforms its two sub-components. This indicates the complementary effects between the images in a stereo pair. Fig. 5 shows a sample stereo image pair in which the left images is of poor quality (also appears as the top second image in Fig. 4). In this case, the right image can compensate the incompleteness of the left image and yield a stable result through the ensemble.

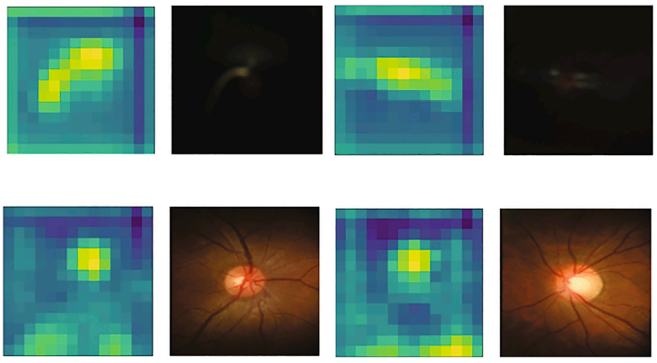

Fig. 4. Examples of the learned attention masks and the corresponding original images. The first row includes two images of poor quality together with their attention masks. The second row includes two images of normal quality together with their attention masks.

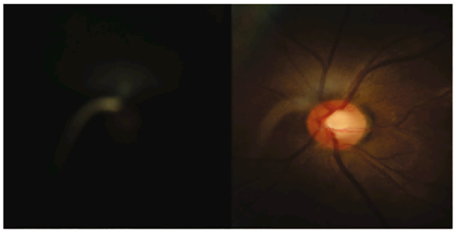

Fig. 5. A sample stereo image pair with poor quality on one side.

5. Conclusion

In this paper, we have introduced a novel approach to glaucoma screening, i.e., the Attention-Guided Stereo Ensemble Networks (AGSE-Net). The network considers abundant information from stereo image pairs to maintain reliability and robustness even when one image in the stereo pair is of poor quality, as well as incorporate attention mechanisms for regions of interest while preserving global features.

The network is very efficient in terms of the training labels needed, as both the attention localization and glaucoma diagnosis are fully-automatic and can be trained in an end-to-end manner with image-level annotation only. Hence, significant labor and time can be saved by avoiding manually labelling optic disc bounding-boxes or segmentation masks.

By performing ablation studies on the network performance with and

Table 2
Ablation Studies on Effects of Attention and Stereo Images (top results are high-lighted in bold). Comparisons between the results show that all four performance metrics of "Left CNN + Left Attention Net" are higher compared to those of "Left CNN Only", while the standard deviations of "Left CNN + Left Attention Net" are lower compared to those of "Left CNN Only". The same can be observed for the right-hand-side counterparts. This indicates that the Attention Net leads to more accurate and yet stabler performance compared to a CNN without attention. Our results show that the Attention Net alone does not out-perform a CNN without attention. We also observe that "Left CNN + Right CNN" outperforms "Left CNN Only" or "Right CNN Only", "Left Attention Net + Right Attention Net" outperforms individual "Left Attention Net" or "Right Attention Net", and "Left CNN + Left Attention Net + Right CNN + Right Attention Net" outperforms individual "Left CNN + Left Attention Net" or "Right CNN + Right Attention Net". This shows the importance of stereo pairs.

Network Branches	Attention	Stereo Image	Precision (Mean/Standard Deviation)	Recall (Mean/Standard Deviation)	Average Precision (Mean/Standard Deviation)	Accuracy (Mean/Standard Deviation)
Left CNN Only	No	No	82.25/6.11	87.77/6.04	91.01/3.49	92.55/1.74
Right CNN Only	No	No	83.15/4.74	87.00/7.90	92.46/2.70	93.02/1.50
Left Attention Net Only	Yes	No	82.10/6.56	84.70/7.27	89.20/4.53	91.81/1.80
Right Attention Net Only	Yes	No	82.99//6.09	85.95/6.31	92.22/2.89	92.43/1.97
Left CNN + Left Attention Net	Yes	No	85.47/5.29	87.80/4.52	91.43/2.79	93.50/1.41
Right CNN + Right Attention Net	Yes	No	86.25/3.87	87.57/5.09	93.31/2.36	93.85/1.54
Left CNN + Right CNN	No	Yes	89.81/2.86	91.04/4.91	94.89/1.50	95.46/1.11
Left Attention Net + Right Attention Net	Yes	Yes	88.38/4.76	88.77/6.74	94.41/1.85	94.52/1.79
Left CNN + Left Attention Net + Right CNN + Right Attention Net (as in Table 1, last row)	Yes	Yes	**94.72/1.52**	**95.48/3.66**	**95.23/1.18**	**97.12/0.96**

without attention, we validate the effectiveness of focusing on disc regions in a cost-efficient manner. As the attention module is implemented simply with a convolutional layer inserted at a single backbone location, there is a potential to apply the same attention module to multiple layers in the backbone CNN feature extractor layers to detect regions of interest at different scale, thereby achieving better results. By comparing results with and without stereo image ensembles, we have shown that the left and right images can compensate each other in terms of image quality and glaucoma information to generate more robust results.

In general, ensemble methods show their effectiveness when the individual components of the ensembles capture different information. As we are using the same backbone CNN feature extractor (ResNet-34) for all individual components, the improvement of the model performance after ensemble can be deemed as contributions from different features extracted by each network components. Hence, we can conclude that the attention module and the stereo image adopted are capable of providing distinct yet important features to jointly improve prediction performance.

In future work, we plan to implement other attention-based techniques, including multi-scale attention and channel-wise attention on a wider range of image types. Methods to efficiently supervise the attention training process is also an important topic to improve the accuracy and effectiveness of the attention mechanism that we may further investigate.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank the Editor-in-Chief and the reviewers for their constructive comments and suggestions that have helped to significantly improve this paper.

References

[1] Y.-C. Tham, X. Li, T.-Y. Wong, H. Quigley, T. Aung, C.-Y. Cheng, Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis, Ophthalmology 121 (2014) 2081–2090.

[2] J.K. Virk, M. Singh, M. Singh, Cup-to-disk ratio (cdr) determination for glaucoma screening, in: 2015 1st International Conference on Next Generation Computing Technologies (NGCT), 2015, pp. 504–507.

[3] M. Singh, Introduction to biomedical instrumentation, 2nd ed., PHI Learning Pvt. Ltd., 2014.

[4] J.B. Jonas, A. Bergua, P. Schmitz-Valckenberg, K.I. Papastathopoulos, W.M. Budde, Ranking of Optic Disc Variables for Detection of Glaucomatous Optic Nerve Damage, Investigative Ophthalmology Visual Science 41 (7) (2000) 1764–1773.

[5] Jost B. Jonas, Wido M. Budde, Songhomitra Panda-Jonas, Ophthalmoscopic evaluation of the optic nerve head, Survey of Ophthalmology 43 (4) (1999) 293–320.

[6] Yuki Hagiwara, Joel En Wei Koh, Jen Hong Tan, Sulatha V. Bhandary, Augustinus Laude, Edward J. Ciaccio, Louis Tong, U. Rajendra Acharya, Computer-aided diagnosis of glaucoma using fundus images: A review, Computer Methods and Programs in Biomedicine 165 (2018) 1–12.

[7] A. Poshtyar, J. Shanbehzadeh, H. Ahmadieh, Automatic measurement of cup to disc ratio for diagnosis of glaucoma on retinal fundus images, in: 2013 6th International Conference on Biomedical Engineering and Informatics, 2013, pp. 24–27.

[8] S. Atheesan, S. Yashothara, Automatic glaucoma detection by using funduscopic images, in: 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 813–817.

[9] H. Fu, J. Cheng, Y. Xu, D.W.K. Wong, J. Liu, X. Cao, Joint optic disc and cup segmentation based on multi-label deep network and polar transformation, IEEE Transactions on Medical Imaging 37 (7) (2018) 1597–1605.

[10] Marcos Vinícius dos Santos Ferreira, Antonio Oseas de Carvalho Filho, Alcilene Dalília de Sousa, Aristófanes Corrêa Silva, Marcelo Gattass, Convolutional neural network and texture descriptor-based automatic detection and diagnosis of glaucoma, Expert Systems with Applications 110 (2018) 250–263.

[11] O. Oktay, J. Schlemper, L.L. Folgoc, M.C.H. Lee, M.P. Heinrich, K. Misawa, K. Mori, S.G. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention u-net: Learning where to look for the pancreas, ArXiv vol. abs/1804.03999 (2018).

[12] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: A large-scale database and CNN model," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.10 563–10 572, 2019.

[13] T. Nakagawa, Y. Hayashi, Y. Hatanaka, A. Aoyama, T. Hara, A. Fujita, M. Kakogawa, H. Fujita, T. Yamamoto, Three-dimensional reconstruction of optic nerve head from stereo fundus images and its quantitative estimation, IEEE Engineering in Medicine and Biology Society Conference 2007 (2007) 6748–6751.

[14] Nicolaas J. Reus, Hans G. Lemij, David F. Garway-Heath, P. Juhani Airaksinen, Alfonso Anton, Alain M. Bron, Christoph Faschinger, Gábor Holló, Michele Iester, Jost B. Jonas, Andrea Mistlberger, Fotis Topouzis, Thierry G. Zeyen, Clinical assessment of stereoscopic optic disc photographs for glaucoma: The European optic disc assessment trial, Ophthalmology 117 (4) (2010) 717–723.

[15] W. Ruengkitpinyo, P. Vejjanugraha, W. Kongprawechnon, T. Kondo, P. Bunnun, H. Kaneko, An automatic glaucoma screening algorithm using cup-to-disc ratio and isnt rule with support vector machine, in: IECON 2015–41st Annual Conference of the IEEE Industrial Electronics Society, 2015, pp. 517–521.

[16] A. Agarwal, S. Gulia, S. Chaudhary, M. K. Dutta, R. Burget, andK. Riha, "Automatic glaucoma detection using adaptive threshold based technique in fundus image," 2015 38th International Conference on Telecommunications and Signal Processing (TSP), 2015, pp. 416–420.

[17] A. M. Jose and A. A. Balakrishnan, "A novel method for glaucoma detection using optic disc and cup segmentation in digital retinal fundus images," 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], 2015, pp. 1–5.

[18] H. Alghmdi, Hongying Lilian Tang, M. Hansen, A. O'Shea, L. AlTurk, and T. Peto, "Measurement of optical cup-to-disc ratio in fundus images for glaucoma screening," 2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM), 2015, pp. 1–5.

[19] J. Cheng, J. Liu, Y. Xu, F. Yin, D.W.K. Wong, N. Tan, D. Tao, C. Cheng, T. Aung, T. Y. Wong, Superpixel classification based optic disc and optic cup segmentation for glaucoma screening, IEEE Transactions on Medical Imaging 32 (6) (2013) 1019–1032.

[20] N. Dey, A.B. Roy, A. Das, S.S. Chaudhuri, Optical cup to disc ratio measurement for glaucoma diagnosis using harris corner, in: 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12), 2012, pp. 1–5.

[21] G. Kande, P. Venkata Subbaiah, T. Savithri, Feature extraction in digital fundus images, Journal of Medical and Biological Engineering 29 (2009) 122–130.

[22] Antonio Sousa Vieira de Carvalho Junior, Edson Damasceno Carvalho, Antonio Oseas de Carvalho Filho, Alcilene Dalília de Sousa, Aristófanes Corrêa Silva, Marcelo Gattass, "Automatic methods for diagnosis of glaucoma using texture descriptors based on phylogenetic diversity", Computers, Electrical Engineering 71 (2018) 102–114.

[23] A. Aquino, M.E. Gegundez-Arias, D. Marin, Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques, IEEE Transactions on Medical Imaging 29 (11) (2010) 1860–1869.

[24] V.G. Edupuganti, A. Chawla, A. Kale, Automatic optic disk and cup segmentation of fundus images using deep learning, in: 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 2227–2231.

[25] D. Stalin David, A. Jayachandran, A new expert system based on hybrid colour and structure descriptor and machine learning algorithms for early glaucoma diagnosis, Multimedia Tools and Applications 79 (2020) 5213–5224.

[26] José Denes Lima Araújo, Johnatan Carvalho Souza, Otilio Paulo Silva Neto, Jefferson Alves de Sousa, João Dallyson Sousa de Almeida, Anselmo Cardoso de Paiva, Aristófanes Corrêa Silva, Geraldo Braz Junior & Marcelo Gattass, "Glaucoma diagnosis in fundus eye images using diversity indexes" Multimedia Tools and Applications volume 78, pp.12987–13004 (2019).

[27] A. A. Salam, T. Khalil, M. U. Akram, A. Jameel, and I. Basit, "Automated detection of glaucoma using structural and non structural features," Springer Plus, vol. 5, no. 1, doi: 10.1186/s40064-016-3175-4, 2016.

[28] U. Rajendra Acharya, E.Y.K. Ng, Lim Wei Jie Eugene, Kevin P. Noronha, Lim Choo Min, K. Prabhakar Nayak, Sulatha V. Bhandary, Decision support system for the glaucoma using gabor transformation, Biomedical Signal Processing and Control 15 (2015) 18–26.

[29] Kevin P. Noronha, U. Rajendra Acharya, K. Prabhakar Nayak, Roshan Joy Martis, Sulatha V. Bhandary, Automated classification of glaucoma stages using higher order cumulant features, Biomedical Signal Processing and Control 10 (2014) 174–183.

[30] Sumeet Dua, U. Rajendra Acharya, Pradeep Chowriappa, S. Vinitha Sree, Wavelet-based energy features for glaucomatous image classification, IEEE Transactions on Information Technology in Biomedicine 16 (1) (2012) 80–87.

[31] J. Sousa, A. Paiva, J. Almeida, A. Silva, G. Junior, M. Gattass, Texture based on geostatistic for glaucoma diagnosis from fundus eye image, Multimedia Tools and Applications 76 (2017) 19173–19190.

[32] L. Xiong, H. Li, Y. Zheng, Automatic detection of glaucoma in retinal images, in: 2014 9th IEEE Conference on Industrial Electronics and Applications, 2014, pp. 1016–1019.

[33] X. Bai, I. Niwas S, W. Lin, B.-F. Ju, C.-K. Kwoh, L. Wang, C. Sng, M. Aquino, and P. Chew, "Learning ECOC code matrix for multiclass classification with application to glaucoma diagnosis," Journal of Medical Systems, vol. 40, 40, Article number: 78, 2016.

[34] J. Ker, L. Wang, J. Rao, T. Lim, Deep learning applications in medical image analysis, IEEE Access 6 (2018) 9375–9389.

[35] X. Chen, Y. Xu, D.W. Kee Wong, T.Y. Wong, J. Liu, Glaucoma detection based on deep convolutional neural network, in: 2015 37thAnnual International Conference

of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 715–718.
[36] J. Nayak, U.R. Acharya, P. Bhat, N. Shetty, T.-C. Lim, Automated diagnosis of glaucoma using digital fundus images, J. Med. Syst. 33 (5) (2009) 337–346.
[37] J. Orlando, E. Prokofyeva, M. del Fresno, and M. Blaschko, "Convolutional neural network transfer for automated glaucoma identification," 12th International Symposium on Medical Information Processing and Analysis, 2016, Tandil, Argentina, 2016.
[38] A. Pal, M.R. Moorthy, A. Shahina, G-eyenet: A convolutional autoencoding classifier framework for the detection of glaucoma from retinal fundus images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 2775–2779.
[39] B. Al-Bander, W. Al-Nuaimy, M. A. Al-Taee, and Y. Zheng, "Automated glaucoma diagnosis using deep learning approach," 2017 14th Inter-national Multi-Conference on Systems, Signals Devices (SSD), 2017, pp. 207–210.
[40] A. Li, J. Cheng, D.W.K. Wong, J. Liu, Integrating holistic and local deep features for glaucoma classification, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, pp. 1328–1331.
[41] A.C. de Moura Lima, L. Bezerra Maia, R.M. Pinheiro Pereira, G.B. Junior, J. D. Sousa de Almeida, A. Cardoso de Paiva, Glaucoma diagnosis over eye fundus image through deep features, in: 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP), 2018, pp. 1–4.
[42] M. Norouzifard, A. Nemati, H. GholamHosseini, R. Klette, K. Nouri-Mahdavi, S. Yousefi, Automated glaucoma diagnosis using deep and transfer learning: Proposal of a system for clinical testing, International Conference on Image and Vision Computing New Zealand (IVCNZ) 2018 (2018) 1–6.
[43] A. Li, Y. Wang, J. Cheng, J. Liu, Combining multiple deep features for glaucoma classification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 985–989.
[44] H. Fu, J. Cheng, Y. Xu, C. Zhang, D.W.K. Wong, J. Liu, X. Cao, Disc-aware ensemble network for glaucoma screening from fundus image, IEEE Transactions on Medical Imaging 37 (11) (2018) 2493–2501.
[45] Yidong Chai, Hongyan Liu, Jie Xu, Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models, Knowledge-Based Systems 161 (2018) 147–156.
[46] E. Corona, S. Mitra, M. Wilson, T. Krile, Y.H. Kwon, P. Soliz, Digital stereo image analyzer for generating automated 3-d measures of optic disc deformation in glaucoma, IEEE Transactions on Medical Imaging 21 (10) (2002) 1244–1253.
[47] M. Norouzifard, A. Dawda, A. Abdul-Rahman, H. GholamHosseini, R. Klette, Superpixel segmentation methods on stereo fundus images and disparity map for glaucoma detection, International Conference on Image and Vision Computing New Zealand (IVCNZ) 2018 (2018) 1–6.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 770–778.
[49] Y. Gao, Z. Huang, Y. Dai, Dsan: Double supervised network with attention mechanism for scene text recognition, in: 2019 IEEE Visual Communications and Image Processing (VCIP), 2018, pp. 1–4.
[50] U. Raghavendra, Hamido Fujita, Sulatha V. Bhandary, Anjan Gudigar, Jen Hong Tan, U. Rajendra Acharya, Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images, Information Sciences 441 (2018) 41–49.
[51] Justin Ker, Yeqi Bai, Hwei Yee Lee, Jai Rao, and L. P. Wang, "Automated brain histology classification using machine learning," Journal of Clinical Neuroscience, vol.66, pp. 239–245, 2019.
[52] Justin Ker, Satya P. Singh, Yeqi Bai, Jai Rao, Tchoyoson Lim, Lipo Wang, Image Thresholding Improves 3-Dimensional Convolutional Neural Network Diagnosis of Different Acute Brain Hemorrhages on Computed Tomography Scans, Sensors 19 (9) (2019) 2167, https://doi.org/10.3390/s19092167.
[53] H. Asha Gnana Priya, J. Anitha, Daniela Elena Popescu, Anju Asokan, D. Jude Hemanth, Le Hoang Son, "Detection and Grading of Diabetic Retinopathy in Retinal Images Using Deep Intelligent Systems: A Comprehensive Review", Computers, Materials and Continua, vol. 66, no.3, pp: 2771-27786, 2020.
[54] D. Jude Hemanth, Omer Deperlioglu, Utku Kose, An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network, Neural Computing and Applications 32 (3) (2020) 707–721.
[55] D. Jude Hemanth, J. Anitha, Le Hoang Son, Mamta Mittal, Le Hoang Son and Mamta Mittal, "Diabetic Retinopathy diagnosis from retinal images using modified Hopfield neural network", Journal of Medical Systems 42 (12) (2018) https://doi.org/10.1007/s10916-018-1111-6.
[56] J. Anitha, D. Jude Hemanth, An Efficient Kohonen-Fuzzy Neural Network Based Abnormal Retinal Image Classification System, Neural Network World 23 (6) (2013) 149–167.
[57] J. Anitha, C.K.S. Vijila, A.I. Selvakumar, A. Indumathy, D. Jude Hemanth, Automated multi-level pathology identification techniques for abnormal retinal images using Artificial Neural Networks, British Journal of Ophthalmology 96 (2) (2012) 220–223.
[58] L.P. Wang, Yaoli Wang, C. Qing, "Feature selection methods for big data bioinformatics: a survey from the search perspective," Methods, vol.111, pp.21-31, 2016.
[59] L.P. Wang, Nina Zhou, Feng Chu, A general wrapper approach to selection of class-dependent features, IEEE Trans. Neural Networks 19 (7) (2008) 1267–1278.
[60] Lipo Wang, Feng Chu, Wei Xie, Feng Chu, and Wei Xie, "Accurate cancer classification using expressions of very few genes", IEEE-ACM Trans. Bioinformatics and Computational Biology 4 (1) (2007) 40–53.
[61] Bing Liu, Chunru Wan, and L.P. Wang, "An efficient semi-unsupervised gene selection method via spectral biclustering", IEEE Trans. Nano-Bioscience, vol.5, no.2, pp.110-114, June, 2006.
[62] H. Liu, E.R. Dougherty, J.G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, L. Yu, Z. Zhao, G Forman, Evolving feature selection, IEEE Intelligent Systems 20 (6) (2005) 64–76.
[63] Fu. Xiuju, L.P. Wang, Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance, IEEE Trans. System, Man, Cybern, Part B-Cybernetics 33 (3) (2003) 399–409.
[64] Xi Chen, Zhiqiang Li, Jie Jiang, Zhen Han, Shiyi Deng, Zhihong Li, Tao Fang, Hong Huo, Qingli Li, Min Liu, Adaptive Effective Receptive Field Convolution for Semantic Segmentation of VHR Remote Sensing Images, IEEE Trans. Geosci. Remote. Sens. 59 (4) (2021) 3532–3546.
[65] Qian Wang, Li Sun, Yan Wang, Mei Zhou, Menghan Hu, Jiangang Chen, Ying Wen, Qingli Li, Identification of Melanoma From Hyperspectral Pathology Image Using 3D Convolutional Networks, IEEE Trans. Medical Imaging 40 (1) (2021) 218–227.
[66] Yunlu Wang, Menghan Hu, Yuwen Zhou, Qingli Li, Nan Yao, Guangtao Zhai, Xiao-Ping Zhang, Xiaokang Yang, Unobtrusive and Automatic Classification of Multiple People's Abnormal Respiratory Patterns in Real Time Using Deep Neural Network and Depth Camera, IEEE Internet Things J. 7 (9) (2020) 8559–8571.
[67] Qian Huang, Wei Li, Baochang Zhang, Qingli Li, Ran Tao, Nigel H. Lovell, Blood Cell Classification Based on Hyperspectral Imaging With Modulated Gabor and CNN, IEEE J. Biomed. Health Informatics 24 (1) (2020) 160–170.
[68] Xueling Wei, Wei Li, Mengmeng Zhang, Qingli Li, Medical Hyperspectral Image Classification Based on End-to-End Fusion Deep Neural Network, IEEE Trans. Instrum. Meas. 68 (11) (2019) 4481–4492.