Received July 6, 2019, accepted July 19, 2019, date of publication July 26, 2019, date of current version August 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2930941

Feasibility of Diagnosing Both Severity and Features of Diabetic Retinopathy in Fundus Photography

JUAN WANG[1], (Member, IEEE), YUJING BAI[2], AND BIN XIA[3]

[1]Delta Micro Technology Inc., Laguna Hills, CA 92653, USA
[2]Shenzhen Sibionics Company Ltd., Shenzhen 518000, China
[3]Shenzhen SiBright Company Ltd., Shenzhen 518000, China

Corresponding author: Juan Wang (wangjuan313@gmail.com)

ABSTRACT Diabetic retinopathy (DR) diagnosis methods in the literature are usually criticized as being limit in diagnosing DR-related features or being lack of interpretability. To deal with these issues, this paper investigates the feasibility of diagnosing both DR severity levels and the presence of DR-related features in a two-step procedure. Specifically, this paper first analyzes the quality of annotations in DR grading by measuring inter-grader variability. Cosine similarity is considered to evaluate the inter-grader variability of the presence of DR-related features, and quadratic weighted Cohen's kappa is employed to assess the inter-grader variability of DR severity levels. Next, different annotation methods as follows are compared to DR severity prediction performance using logistic regression: 1) single annotations by single grader (SASG); 2) single annotations from multiple graders (SAMG); 3) multiple annotations by voting (MAV); and 4) double annotations with adjudication of disagreement (DAAD). Based on the comparison results, the feasibility of diagnosing both DR severity and features is investigated. In the experiments, 1589 fundus images graded by three retinal specialists and four general ophthalmologists are considered. The results demonstrate that retinal specialists are more consistent than general ophthalmologists in grading both the presence of DR-related features and DR severity. The SASG and MAV should be avoided if possible while the DAAD is the good option when prediction performance is the highest priority and the SAMG is especially beneficial when both prediction performance and grading costs are considered. The upper limit performance of DR severity prediction gets accuracy 95.6% and kappa 0.962. When DR-related feature prediction achieves average cosine similarity 0.823, it is potential to get accuracy 91.2% and kappa 0.905 for DR severity prediction in real applications. These results together suggest the potential of diagnosis of both DR severity and the presence of DR-related features in a two-step procedure.

INDEX TERMS Diabetic retinopathy (DR), DR severity, DR related features, inter-grader variability, data annotation.

I. INTRODUCTION

Diabetic retinopathy (DR) is the leading cause of preventable blindness among working-aged adults in the world [1]. It is estimated that 35% of all patients with diabetes mellitus suffer from DR [2], [3]. The risk of DR increases as the longer a person has diabetes mellitus. According to World Health Organization, DR is estimated to affect more than 77% of the patients who have had diabetes 20 years and more [4]. Fundus photography is one of the most commonly used imaging technique for the diagnosis of DR in the retina [5]. It has been widely used for DR screening because of its high resolution, low cost, and easy storage and transmission. To facilitate DR diagnosis, several classification systems have been developed to classify DR severity in fundus images [6]–[8], in which DR severity levels are defined by different DR related features.

Due to the importance of DR for patients with diabetes mellitus, there have been great efforts in development of computerized methods for automatic DR diagnosis in fundus images, which can be broadly divided into two categories: DR related feature detection and DR severity classification. The traditional DR diagnosis methods usually target at DR related feature detection, which have been widely studied in the literature. For example, Sinthanayothin *et al.* [9] developed a

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

method based on recursive region growing and Moat operator to automatically detect important features of nonproliferative DR, including haemorrhages, microaneurysms, and hard exudates. Sopharak *et al.* [10] proposed a set of optimally adjusted morphological operators for exudate detection. Zhang *et al.* [11] designed a method by combining mathematical morphology, feature extraction, and random forest for exudate detection. Seoud *et al.* [12] employed dynamic shape features for detection of microaneurysms and hemorrhages. Although with great success, these methods are usually designed for some pre-selected features and do not work on the other features, thus limiting their application in DR diagnosis.

More recently, with the success of deep learning in medical image analysis, the study on the classification of DR severity becomes popular. For example, Gulshan *et al.* [13] employed GoogleNet to diagnose referable DR. Ting *et al.* [14] developed VGG-like convolutional neural networks (CNNs) for referable DR and vision-threatening DR detection, respectively. Li *et al.* [15] designed a CNN method for detecting vision-threatening referable DR. Achieved high sensitivity and specificity in the relative large test sets, however, these methods are usually criticized as being lack of interpretability. Except for the black box of deep learning models, another important reason, we believe, is the lack of the information about DR related features in the output of these methods. Besides, all of these methods are only optimized for two-class classification tasks, thus can not predict all of DR severity levels.

Due to the problems mentioned above, in this study we investigate the feasibility of developing automatic computerized algorithms for diagnosis of both DR severity and the presence of DR related features. It can be conducted in a two-step procedure: the presence of DR related features in the fundus images is first detected, and then the classification of DR severity is predicted based on the detected features. Note different from the traditional feature detection methods, which provide both the number and the locations of *some pre-selected* DR related features, this method considers the *presence of all* DR related features. Instead of considering two-class classification as the existing methods for DR diagnosis, this method classifies all DR severity levels. In the two-step procedure, DR related feature detection problem can be formulated as a binary multi-label classification problem, and DR severity prediction problem can be formulated as a multi-class classification problem. Both classifiers can be trained with supervised learning, for which data with known annotations have to be available.

In supervised learning, annotation quality has a great impact on the classifier development [16]–[18]. The use of inaccurate annotations may lead to decreased performance in prediction, the high complexity of classification models, and requirement of more training data [16]. For example, Pelletier *et al.* [19] studied effect of noisy annotation on classification performances for land cover mapping with satellite image time series and concluded that classifiers are little influenced for low random noise levels up to 25%–30%, but their performances drop down for higher noise levels. Garcia *et al.* [20] investigated the effect of noisy annotation on the complexity of classification problems by monitoring the sensitivity of several indices of data complexity, and demonstrated that some measures, such as separability of the classes, alterations in the class boundary and densities within the classes were the most affected ones by noise in the annotations.

For DR grading, data annotations are usually provided by human graders, such as retinal specialists and general ophthalmologists. However, due to the subjectivity of graders, the imperfect experience and knowledge of the experts, and the difficulty of the tasks, errors of annotations are unavoidable. Therefore, this study first analyzes the quality of annotations in DR diagnosis. Annotation quality is generally evaluated by metrics that assess inter-grader variability [21]. The low inter-grader variability usually indicate high quality of annotations [22]. In the literature, several studies have demonstrate fair to moderate agreement among graders for the classification of DR severity [23]–[26]. However, to our best knowledge, there is no inter-grader variability studies for the presence of DR related features.

Furthermore, data annotation method also has great effect on model development [27], which fundamentally determines the quality of annotations. In the literature, different annotation methods have been employed, such as single annotation by multiple graders together [27], multiple annotations by voting [13], and double annotations with adjudication of disagreements [28], etc. However, the systematic comparison of different annotation methods is rare. Therefore, this study also quantitatively compares different annotation methods to provide evidence and guidance for the selection of annotation methods, based on which the feasibility of DR diagnosis for both DR severity and the presence of DR related features is investigated.

In summary, as a preliminary study, this work investigates two important issues related to the feasibility of DR diagnosis for both DR severity levels and the presence of DR related features in fundus images. Firstly, we study the quality of annotations in DR grading by analyzing the variability among different graders in identifying the presence of DR related features and classifying DR severity. Secondly, we quantitatively compare different annotation methods for predicting DR severity levels. The recommendations of data annotations are provided and the feasibility of diagnosing both DR severity levels and the presence of DR related features is presented. In this work, analyses are conducted on a set of 1589 fundus images graded by three retinal specialists and four general ophthalmologists. The results demonstrate that it is feasible to diagnose both DR severity levels and DR related features in fundus images with good performance.

The contributions of this study are summarized as follows:

1) First, we study the quality of annotations in DR grading, and present suggestions for the selection order of graders for data annotation.

2) Second, we quantitatively compare different annotation methods, and provide recommendations for annotation method selection.
3) Finally, we investigate the feasibility of diagnosing both DR related features and DR severity using a two-step procedure, and obtain the upper limit performance of DR severity prediction.

II. METHODS

A. IMAGE DATASET AND GRADING

The dataset considered in this study consists of 45° field-of-view digital fundus images. It includes 1589 images collected in a DR screening project. The images are captured in either macula-centered or optic-disk-centered. Seven graders participated in the image grading. Among them, three are retinal specialists (denoted as G_1, G_2 and G_3) and four are general ophthalmologists (denoted as G_4, G_5, G_6, and G_7). All of the three retinal specialists are with certificate of grading in DR screening jointly issued by Gloucestershire Retinal Education Group and Chinese Foundation for Lifeline Express. General ophthalmologists are selected based on their experienced levels in evaluating fundus images measured by the number of years. Specifically, G_4 has more than ten years of experience in fundus image scoring, G_5 have more than five years but less than ten years of experience, G_6 has more than three years but less than five years of experience, and G_7 has more than one year but less than three years of experience.

All images were graded using an online image grading platform developed by Shenzhen SiBright Co. Ltd. (Shenzhen, Guangdong, China), in which the graders can zoom in/out images for better visualization. Before grading, graders were required to participate a training, in which 100 images were given, and the ground truth of each image was displayed immediately after the grading for self-evaluation. Images in the dataset were randomly grouped into 18 sessions for grading, with 90 images in the first 17 sessions and 59 images in the last session. During grading, graders independently provide DR severity levels and detect the presence of DR related features for each image.

The classification of DR severity is assessed based on the International Clinical Diabetic Retinopathy (ICDR) scale [8], which is one of the most popular DR classification system and is recommended by The International Council of Ophthalmology (ICO) Guidelines for Diabetic Eye Care [29]. Based on the progress stage of the disease, ICDR scale defines five levels of DR severity as follows: none, mild, moderate, severe, and proliferative DR.

During grading, all of DR related features as follows are considered: 1) microaneurysms (MA), 2) intraretinal hemorrhages (IRH), 3) superficial retinal hemorrhages (SRH), 4) hard exudate (HE), 5) cotton wool spots (CWS), 6) venous abnormality (VAN), including venous distortion and dilation, venous beading, venous loops, venous reduplication, 7) intraretinal microvascular anomalies (IRMA), 8) new vessels elsewhere (NVE), 9) new vessels at the disc (NVD), 10) preretinal fibrous proliferation (PFP), 11) vitreous or preretinal hemorrhage (VPH), and 12) traction retinal detachment (TRD). However, VPH and TRD are excluded from analysis in this study because they appear rare in the dataset (seven graders detect 1.14 images with VPH on average and none image with TRD). Note in this study, graders are only required to determine the presence or absence of these features, other information such as the number, sizes, locations, and descriptions of the features are not acquired.

For quantitative analysis, in this study different DR severity levels are denoted by integer number from 0 to 4, with value 0 for none and value 4 for proliferative DR. The gradings of ten DR related features are represented by vectors with ten elements, one element for a feature. For each element, value 1 indicates the presence of the corresponding feature, and value 0 otherwise.

B. EVALUATING QUALITY OF ANNOTATIONS FOR DR GRADING

1) INTER-GRADER VARIABILITY FOR IDENTIFYING THE PRESENCE OF DR RELATED FEATURES

To analyze the agreement among different graders for detecting the presence of DR related features, cosine similarity [30] is employed. Cosine similarity is a measure of similarity between two non-zero vectors by evaluating the cosine of the angle between them. Mathematically, for vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$, their cosine similarity is defined as:

$$C = \frac{\mathbf{Y}_1 \cdot \mathbf{Y}_2}{\|\mathbf{Y}_1\|\|\mathbf{Y}_2\|} \tag{1}$$

where $\cdot$ sign denotes dot product and $\|\mathbf{Y}\|$ represents the magnitude of $\mathbf{Y}$. Cosine similarity is ranged from -1 to 1, in which 1 means exactly same similarity, and 0 indicates complete dissimilarity.

In this study, for a pair of two graders, cosine similarity is calculated first for each image from the corresponding feature vectors, and then average cosine similarity among all images are obtained to measure their inter-grader consistency. Note since only the presence of DR related features is interested in this study, the images which are identified as absence of any features by both graders are not considered in the average process above.

2) INTER-GRADER VARIABILITY FOR GRADING DR SEVERITY

To quantitatively measure the agreement between two graders for DR severity, quadratic weighted CohenâĂŹs kappa coefficient κ [31] (short for kappa coefficient for simplicity in this study) is considered. It evaluates the agreement of a pair of graders in classifying N instances into C mutually exclusive categories. Mathematically, it is defined as:

$$\kappa = 1 - \frac{\sum_{i=1}^{C}\sum_{j=1}^{C}\mathbf{W}_{ij}\mathbf{X}_{ij}}{\sum_{i=1}^{C}\sum_{j=1}^{C}\mathbf{W}_{ij}\mathbf{E}_{ij}} \tag{2}$$

where $\mathbf{W}_{ij}$, $\mathbf{X}_{ij}$, and $\mathbf{E}_{ij}$ are the quadratic weight, the number of instances that received grading i by 1st grader and grading

j by 2nd grader, and the expected number of instances that received grading i by 1st grader and grading j by 2nd grader by chance, respectively. The quadratic weight $\mathbf{W}_{ij}$ is defined as follows:

$$\mathbf{W}_{ij} = \frac{(i-j)^2}{(C-1)^2} \tag{3}$$

The use of quadratic weight is to penalize the gradings with high difference between two graders. Kappa coefficient varies from -1 to 1, where $\kappa = 1$ when the two graders are in complete agreement, and $\kappa = 0$ for no agreement other than what would be expected by chance.

3) RELATIONSHIP VISUALIZATION

To visualize the relationships among the graders in DR grading, we apply multidimensional scaling (MDS) [32], a data visualization method by which data points having known relationships to one another can be represented in a two- or three-dimensional scatter plot. In an MDS plot, the data points are arranged in such a way that the distance between any two points reflects the relative degree of ''dissimilarity'' between those points. Consider a set of M data points, MDS seeks to embed these data points in a lower-dimensional space by minimizing the following objective function:

$$\sigma_1^2 = \frac{\sum w_{ij}[d(\mathbf{x}_i, \mathbf{x}_j) - \delta_{ij}]^2}{\sum w_{ij}\delta_{ij}^2}, \tag{4}$$

where $\mathbf{x}_i$ are the embedded data points which are shown in MDS, δ_{ij} are their pairwise proximity measure, w_{ij} are their corresponding weight factors (set as 1 in this study), and $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e., $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|$. The quantity σ_1 is known as Stress-1 [32], which measures the goodness of fit of the MDS model. If two data points i and j have proximity measure δ_{ij}, then their corresponding points in the 2D MDS plot are to be separated approximately by distance $d_{ij} \approx \delta_{ij}$.

In this study, to visualize the relationships among graders for classifying DR severity, each grader is represented as a data point in a two-dimensional (2D) plane, wherein the dissimilarity between a pair of graders is defined as 1 minus their kappa coefficient. Thus, graders close in the MDS plot are those exhibiting higher inter-grader consistency in their readings. Similarly, to visualize the relationships among graders for identifying the presence of DR related features, the dissimilarity between a pair of graders is defined as 1 minus their corresponding average cosine similarity.

C. PREDICTING DR SEVERITY WITH DIFFERENT ANNOTATION METHODS

1) DR SEVERITY PREDICTION AND EVALUATION

In the two-step procedure, DR severity is predicted by the presence of DR related features. For this purpose, logistic regression is considered. Logistic regression is a linear classifier for binary classification. It is one of the most widely used classifiers in the literature due to low computational complexity and good generalization [33]. Mathematically, a linear classifier for feature vector $\mathbf{x}$ is expressed as

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{5}$$

where $\mathbf{w}$ is the discriminant vector and b is the bias. Suppose $(\mathbf{x}_i, y_i), i = 1, 2, \cdots, N$ be a set of N training samples, in which $\mathbf{x}_i$ and y_i are feature vector and label of the ith sample, respectively. For logistic regression with L_1 norm regularization, parameters $\mathbf{w}$ and b are determined by maximizing the loss function as follows:

$$L(\mathbf{w}, b) = \sum \log p(y_i|\mathbf{x}_i; \mathbf{w}, b) + \lambda\|\mathbf{w}\|_1 \tag{6}$$

The parameter λ is used to control the trade-off between log likelihood loss (first term) and model complexity (second term). The probability term is written as a logistic sigmoid acting on the linear classifier model as follows:

$$p(y_i = 1|\mathbf{x}_i; \mathbf{w}, b) = \frac{1}{1 + \exp(-f(\mathbf{x}))} \tag{7}$$

$$p(y_i = 0|\mathbf{x}_i; \mathbf{w}, b) = 1 - p(y_i = 1|\mathbf{x}_i; \mathbf{w}, b) \tag{8}$$

Such choice of the posterior probabilities of the two classes is called logit transformation.

In this study, DR severity prediction problem is formulated as a five-class classification problem. $\mathbf{x}$ denotes the feature vector for the presence of DR related features and y is the DR severity. To achieve five-class classification, logistic regression models are trained using the one-vs-rest scheme, in which each model discriminates between a severity level and all of the other levels. For model selection and performance evaluation, a nested double-loop five-fold cross validation procedure is considered. In the inner loop, a grid search on parameter λ is combined with the inner five-fold cross validation for model selection; the outer loop is for performance evaluation.

To measure the prediction performance, both kappa coefficient and accuracy are considered. Kappa coefficient measures the consistency between predictions and ground truths, while accuracy evaluates the ratio of predictions which exactly match with ground truths.

2) ANNOTATION METHODS FOR COMPARISON

To develop effective data annotation method for DR grading, four different annotation methods as follows are considered, in which suppose the cost of grading an image per each grader is c dollars in expenses and s seconds in time, and let the number of images be N.

(a) *Single annotation by single grader (SASG):* For this method, annotations of all images are obtained by a single grader. The cost of this method is Nc dollars and Ns seconds. In this study, the single grader can be G_i, $i = 1, 2, \cdots, 7$.

(b) *Single annotation by multiple graders together (SAMG):* Suppose n graders (n>1) participate in the data annotation. For this method, images are randomly divided into n non-overlap subsets, and each subset is

assigned to a grader for independent image grading. The cost of this method is *Nc* dollars and *Ns* seconds, which is independent of the number of graders n. In this study, the equivalent grader of this method is denoted as AG_{nr}, where r denotes randomly assigned images to n graders with highest inter-grader agreement.

(c) *Multiple annotations by voting (MAV):* In this method, annotations are obtained from voting of gradings from multiple graders, in which the number of graders (denoted as n) has to be odd. This method costs *nNc* dollars and *nNs* seconds ($n \geq 3$). In this study, the equivalent grader of this method denoted as AG_n, in which n graders with highest inter-grader agreement are considered.

(d) *Double annotations with adjudication of disagreement (DAAD):* This method requires three graders. It is conducted in two steps. Firstly, two baseline graders independently score all images. Secondly, when there is disagreement in the gradings of the two baseline graders, the adjudication is done by a third grader. The third grader is adjudication grader, who has to be more experienced than the two baseline graders. Let the disagreement ratio of the gradings from the two baseline graders be r ($0 \leq r \leq 1$), then the cost of this method is $(2+r)Nc$ dollars and $(2+r)Ns$ seconds. In this study, the grader with highest inter-grader agreement is selected as the adjudication grader (denoted as G_K), and the other graders are randomly paired as the baseline graders. The equivalent grader of this method is denoted as G_i–G_j, where $i, j \in \{1, 2, 3, 4, 5, 6, 7\}$ and $i \neq j \neq K$.

III. RESULTS

A. ANNOTATION QUALITY ANALYSIS FOR DR GRADING

1) LEVEL OF VARIABILITY IN DETECTING THE PRESENCE OF DR RELATED FEATURES

To evaluate variability among the different graders in identifying the presence of DR related features, we computed average cosine similarity in a pairwise fashion for all of the seven graders. The results are shown in Table 1. Owing to the symmetry of the cosine similarity, the entries in the lower triangular portion are omitted in Table 1. As can be seen, among the different graders, the values of average cosine similarity range from 0.217 (between G_1 and G_7) to 0.716 (between G_1 to G_3). For the individual graders, G_3 is most consistent with the others (mean of average cosine similarity $= 0.527$), while G_7 is least consistent (mean of average cosine similarity $= 0.237$). Moreover, from Table 1, the mean of average cosine similarity is 0.349 for general ophthalmologist pairs, which is much lower than 0.681 for retinal specialist pairs. These results indicate that the gradings from retinal specialists tend to be more accurate than those from general ophthalmologists. Therefore, the best annotations can be obtained as majority votings of the gradings from the three retinal specialists (denoted as *RS*), which are considered as ground-truth annotations for the presence of DR related features in this study.

TABLE 1. Average cosine similarity obtained for different pairs of graders for the presence of DR related features.

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	RS
G_1	1	0.654	0.716	0.620	0.421	0.404	0.217	0.829
G_2	-	1	0.672	0.622	0.421	0.426	0.219	0.788
G_3	-	-	1	0.627	0.458	0.451	0.240	0.863
G_4	-	-	-	1	0.461	0.411	0.218	0.658
G_5	-	-	-	-	1	0.477	0.254	0.453
G_6	-	-	-	-	-	1	0.274	0.445
G_7	-	-	-	-	-	-	1	0.237

TABLE 2. Pair-wise inter-grader agreement for DR grading measuring by kappa coefficient.

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	RS
G_1	1	0.786	0.847	0.795	0.692	0.642	0.430	0.914
G_2	-	1	0.810	0.797	0.714	0.666	0.430	0.878
G_3	-	-	1	0.824	0.758	0.689	0.492	0.935
G_4	-	-	-	1	0.770	0.656	0.478	0.834
G_5	-	-	-	-	1	0.692	0.526	0.753
G_6	-	-	-	-	-	1	0.572	0.687
G_7	-	-	-	-	-	-	1	0.473

In Table 1, we also show the average cosine similarity between individual graders and *RS*. It can be seen that the average cosine similarity satisfies $G_3 > G_1 > G_2 > G_4 > G_5 > G_6 > G_7$, which suggests priority of grader selection for annotating the presence of DR related features in fundus images. The average cosine similarity scores are low for G_5 (0.453), G_6 (0.445), G_7 (0.237), which are even lower than their corresponding average cosine similarity scores with some other graders. For example, average cosine similarity is 0.458 between G_5 and G_3, 0.477 between G_6 and G_5, and 0.274 between G_7 and G_6. These results indicate that it is inappropriate to use these graders alone for grading the presence of DR related features.

2) LEVEL OF VARIABILITY FOR GRADING DR SEVERITY

To evaluate variability among the different graders in their gradings of DR severity levels, we computed kappa coefficients in a pairwise fashion for all of the seven graders. The results are shown in Table 2. As can be seen, among the different graders, the kappa values range from 0.430 (between G_1 and G_7, between G_2 and G_7) to 0.847 (between G_1 to G_3), indicating moderate to almost perfect agreement. Among the individual graders, G_3 is most consistent with the others (average kappa $= 0.737$), while G_7 is least agreed with the others (average kappa $= 0.488$). Compared to general ophthalmologists (average kappa $= 0.616$), retinal specialists are more consistent with each other. Their average kappa value is 0.814, indicating almost perfect agreement. Therefore, the best annotations can be obtained as median values of

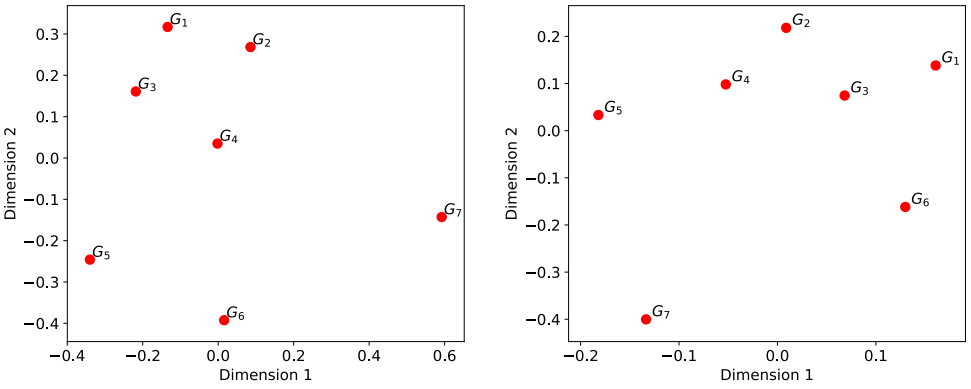

FIGURE 1. MDS plot of seven graders for (a) identifying DR related features and (b) classifying DR severity levels.

the gradings from the three retinal specialists (i.e. *RS*), which are considered as ground-truth annotations for DR severity in this study.

In Table 2, we also show the kappa coefficients between individual graders and *RS*. As can be seen, kappa coefficient satisfies $G_3 > G_1 > G_2 > G_4 > G_5 > G_6 > G_7$, which suggests priority of grader selection for annotating DR severity in fundus images. Moreover, kappa coefficients are low for G_5 (0.753), G_6 (0.687) and G_7 (0.473), which are even lower than their corresponding kappa coefficients with some other graders. For example, kappa coefficient is 0.770 between G_5 and G_4, 0.692 between G_6 and G_5, and 0.572 between G_7 and G_6. The results suggest that it is inappropriate to use these graders alone for grading DR severity levels.

3) VISUALIZATION OF RELATIONSHIPS

Based on the inter-grader variability results of graders in detecting the presence of DR related features in Table 1, an MDS plot of all graders is shown in Figure 1(a) (stress-1 $= 0.274$). Similarly, an MDS plot of all the different graders in grading DR severity levels is shown in Figure 1(b) (stress-1 $= 0.053$) based on the results in Table 2. In these MDS plots, each grader is denoted by a point, and the pairwise distance between the different points indicate how consistent their corresponding graders are.

From Figure 1(a), it can be seen that the retinal specialists are all placed within a small cluster and they are more densely distributed. G_4 is close to the cluster of the retinal specialists, G_5, G_6 and G_7 are further away from the others. These observations are consistent with the results obtained in Table 1. The similar trends are observed from Figure 1(b) as well.

More importantly, it can be seen that the distribution and relative positions of graders in Figure 1 (a) and (b) are very similar to each other. It indicates that the gradings from different graders yield very similar relationships between DR severity levels and the presence of DR related features. It implies the feasibility of accurately predicting DR severity levels by the presence of DR related features.

B. COMPARISON OF DIFFERENT ANNOTATION METHODS

As demonstrated in Section III-A, the best annotations are those provided by *RS*, which are treated as ground truth in this study. To compare different annotation methods in supervised learning, the experiments in this subsection consider DR severity prediction by assuming that the presence of DR related features is known. For this purpose, we fix $\mathbf{x}$ as feature vector provided by *RS*. For performance evaluation, ground truth of y is set as DR severity levels from *RS*.

The different annotation methods only generate different y's in training. According to the results in Section III-A.2, the individual graders satisfy $G_3 > G_1 > G_2 > G_4 > G_5 > G_6 > G_7$ in inter-grader agreement for DR severity, therefore, this order is employed in SAMG and MAV methods for multiple grader selection.

1) SINGLE ANNOTATIONS BY SINGLE GRADER

Table 3 shows the DR severity prediction performance for SASG method. As can be seen, graders G_1, G_2, G_3, and G_4 yield very high accuracy and kappa, indicating the potential of SASG method for data annotation. G_5, G_6 and G_7 get low accuracy and kappa, which are consistent with their low inter-grader agreement results with *RS* in Table 2. Especially, G_7 gets extremely low accuracy (66.4%) and kappa (0.288). It suggests that it might be very dangerous to use annotations from graders with low inter-grader consistency in SASG method. Together, these results indicate that it has to be cautious to use SASG method for data annotation, especially when the inter-grader consistency can not be estimated accurately. Therefore, SASG method should be avoided for data annotation when possible.

2) SINGLE ANNOTATIONS FROM MULTIPLE GRADERS

Table 4 shows the comparison results of different graders in SAMG method for DR severity prediction. Due to the random assignment of images in this method, the prediction performance of each AG_{nr} is evaluated 10 times with different

TABLE 3. Comparison of DR severity prediction performance for different graders in SASG method.

graders	accuracy	kappa
G_1	95.0%	0.954
G_2	94.5%	0.941
G_3	94.9%	0.949
G_4	93.5%	0.946
G_5	87.9%	0.848
G_6	78.2%	0.672
G_7	66.4%	0.288

TABLE 4. Comparison of DR severity prediction performance for different graders in SAMG method. For each grader, the mean and standard deviation (shown in the bracket) of accuracy and kappa obtained from 10 repetitions are given.

graders	accuracy	kappa
AG_{3r}	94.7% (0.2%)	0.948 (0.005)
AG_{4r}	94.6% (0.3%)	0.948 (0.005)
AG_{5r}	93.1% (0.8%)	0.934 (0.007)
AG_{6r}	88.9% (1.3%)	0.871 (0.024)
AG_{7r}	87.9% (0.4%)	0.842 (0.007)

TABLE 5. Comparison of DR severity prediction performance for different graders in MAV method.

graders	accuracy	kappa	training-kappa
RS	95.6%	0.962	1
AG_5	95.0%	0.954	0.941
AG_7	89.9%	0.881	0.909

randomization, and the mean and standard deviation (shown in the bracket) of accuracy and kappa coefficient are given.

From Table 4, it can be seen that AG_{3r} and AG_{4r} has highest mean accuracy and mean kappa, which are close to results of G_1, G_2 and G_3 in Table 3. The prediction performance decreases as the join of graders with low inter-grader consistency. The worst performance is obtained for AG_{7r} (accuracy $=$ 87.9%, kappa $=$ 0.842). Moreover, the standard deviations of both accuracy and kappa are low for all graders, indicating the robustness of SAMG on random image assignment. These results imply that SAMG is a good data annotation method, which is robustness to the presence of graders with low inter-grader agreement.

3) MULTIPLE ANNOTATIONS BY VOTING

Table 5 shows the DR severity prediction performance for different graders in MAV method. Note $RS = AG_3$ in MAV method. As can be seen, *RS* gets accuracy of 95.6% and kappa of 0.962. Due to both $\mathbf{x}$ and y obtained from *RS* are the best annotations, these results are the upper limit of DR severity prediction performance.

Moreover, from Table 5, for both accuracy and kappa, $RS = AG_3 > AG_5 > AG_7$. It implies that inclusion of graders with low inter-grader consistency hurts the predictions. Therefore, it has to be cautious for grader selection in MAV method. Considering that MAV method is also ineffective in cost, it should be avoided, if possible.

TABLE 6. Comparison of DR severity prediction performance for different graders in DAAD method, in which adjudication grader is set as G_3. The adjudication ratio for grading DR severity is also provided.

graders	accuracy	kappa	adjudication ratio
G_1-G_2	95.5%	0.962	15.3%
G_1-G_4	95.2%	0.955	17.5%
G_1-G_5	95.1%	0.956	22.2%
G_1-G_6	95.0%	0.952	26.1%
G_1-G_7	93.7%	0.932	33.0%
G_2-G_4	95.0%	0.953	16.7%
G_2-G_5	93.9%	0.927	19.8%
G_2-G_6	90.9%	0.877	24.2%
G_2-G_7	90.9%	0.871	30.8%
G_4-G_5	93.1%	0.928	16.8%
G_4-G_6	89.7%	0.884	26.7%
G_4-G_7	88.5%	0.864	30.8%
G_5-G_6	88.7%	0.858	20.3%
G_5-G_7	88.0%	0.842	23.3%
G_6-G_7	80.9%	0.702	20.8%

4) DOUBLE ANNOTATIONS WITH ADJUDICATION OF DISAGREEMENT

Table 6 lists the results of graders in DAAD for DR severity prediction. The adjudication grader is set as G_3 due to his/her highest inter-grader agreement with *RS*. For reference, the adjudication ratio is also given in Table 6 (last columns). It can be seen that G_1-G_2 gets highest performance, close to the upper limit of *RS* in Table 5. Comparing to *RS* in MAV method, G_1-G_2 in DAAD method gets 28.2% reduction in cost. Therefore, if the prediction performance is the highest priority, G_1-G_2 in DAAD method should be considered instead.

Moreover, it can be seen that except G_5-G_7 and G_6-G_7, all of the other graders have large kappa values (i.e. $\kappa > 0.85$). These results indicate that it is acceptable to have one baseline grader with low inter-grader agreement, but unacceptable to have both baseline graders with low inter-grader agreement. Therefore, this method is a good candidate for data annotation as well.

Finally, the adjudication ratio is largest for G_1-G_7 (33.0%) and lowest for G_1-G_2 (15.3%), indicating that the former costs 7.60% more than the latter. Therefore, in this method, the difference in cost is small among different graders.

In conclusion, both SAMG and DAAD are good choices for data annotation in supervised learning, while SASG and MAV should be avoided if possible. More importantly, SAMG is the best option if both prediction performance and cost are considered, and DAAD is the choice if the prediction performance has the highest priority.

TABLE 7. Comparison of DR severity prediction performance by different predictions of the presence of DR related features, which are simulated as gradings of different graders in SAMG method. The average cosine similarity for different predictions of the presence of DR related features is provided as well.

graders	accuracy	kappa	average cosine similarity
AG_{3r}	91.2%	0.905	0.823
AG_{4r}	88.0%	0.831	0.778
AG_{5r}	85.8%	0.801	0.732
AG_{6r}	85.0%	0.804	0.671
AG_{7r}	81.6%	0.745	0.621

C. FEASIBILITY OF TWO-STEP PROCEDURE FOR DR DIAGNOSIS

The experiments in the previous subsection assume that $\mathbf{x}$ is the annotations obtained from *RS*, however, such high quality of $\mathbf{x}$ (accuracy $=$ 100%, kappa $=$ 1) can not be obtained in the two-step procedure since $\mathbf{x}$ is the prediction of the classification model in the first step. To investigate how the performance of the classification model in the first step affect the final DR severity prediction, we simulate different predictions of the presence of DR related features (denoted as $\mathbf{x}$) by gradings from different graders in SAMG method. In these experiments, y is obtained by AG_{3r}. The ground truth of $\mathbf{x}$ and y are obtained as the gradings provided by *RS* to evaluate the performance of the DR related feature detection model in the first step and the DR severity prediction model in the second step, respectively.

Table 7 shows the results of different predictions of the presence of DR related features (simulated by different graders in SAMG method) are used as input for DR severity prediction. DR related feature prediction performance is measured by average cosine similarity, and the results are listed in Table 7 as well (in the last column). It can be seen that the best DR severity prediction performance (accuracy $=$ 91.2% and kappa $=$ 0.905) is obtained by AG_{3r} when $\mathbf{x}$ has average cosine similarity of 0.823 (highest among all $\mathbf{x}$'s); while the worst DR severity performance (accuracy $=$ 81.6% and kappa $=$ 0.745) is obtained by AG_{7r} when $\mathbf{x}$ has lowest average cosine similarity of 0.621 (lowest among all $\mathbf{x}$'s). For reference, a 5-class classification with random guessing has accuracy of 20% and kappa of 0. Therefore, even the worst performance in Table 7 is very high and might be acceptable depending on the applications.

Notice in Table 1, the average cosine similarity between retinal specialists and *RS* are 0.863 for G_3, 0.829 for G_1, and 0.788 for G_2. Therefore, among the simulated predictions, highest performance of $\mathbf{x}$ (0.823) is close to the gradings obtained by G_1 and lower than those by G_3. These results indicate that the best $\mathbf{x}$ has similar performance to the gradings provided by single retinal specialists. Numerous studies in medical imaging have demonstrate that deep learning can achieve performance close to or even better than human graders [13], [28]. Therefore, it is highly promising to get deep learning model with DR related feature prediction performance close to 0.823. These results demonstrate that it is feasible to achieve good prediction performance for diagnosis of both DR severity levels and the presence of DR related features in the two-step procedure. Note development of deep learning algorithm for DR related feature prediction requires a large number of fundus images with good annotation quality, which should be conducted based on the data annotation recommendations made in this study and will be studied in the future.

IV. CONCLUSION

This study investigated the feasibility of diagnosing both DR severity levels and the presence of DR features by a two-step procedure. The results demonstrated that the retinal specialists are more consistent than general ophthalmologists in grading both the presence of DR related features and DR severity levels. Among different annotation methods under consideration, SAMG is the good choice when both prediction performance and grading costs are considered, while DAAD is the good option if prediction performance is the highest priority. For DR severity prediction, the upper limit of performance is accuracy of 95.6% and kappa of 0.962. When DR related feature prediction achieves average cosine similarity close to and higher than 0.823, it is potential to get accuracy of 91.2% and kappa of 0.905 or above for DR severity prediction. These results indicate that it is feasible to diagnosing both DR severity and the presence of DR related features in fundus photography. In the future, it would be interesting to apply annotations obtained by SAMG method to DR diagnosis by the two-step procedure, in which a deep learning model is trained first to predict the presence of DR related features, and then DR severity levels are obtained by logistic regression.

ACKNOWLEDGMENT

The authors would like to thank Zhigang Hu from Shenzhen SiBright Co. Ltd. for his assistance in coordinating the graders.

REFERENCES

[1] T. Y. Wong, J. Sun, R. Kawasaki, P. Ruamviboonsuk, N. Gupta, V. C. Lansingh, M. Maia, W. Mathenge, S. Moreker, M. M. K. Muqit, S. Resnikoff, J. Verdaguer, P. Zhao, F. Ferris, L. P. Aiello, and H. R. Taylor, "Guidelines on diabetic eye care: The international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings," *Ophthalmology*, vol. 125, no. 10, pp. 1608–1622, 2018.

[2] J. W. Yau, S. L. Rogers, R. Kawasaki, E. L. Lamoureux, J. W. Kowalski, T. Bek, S.-J. Chen, J. M. Dekker, A. Fletcher, and J. Grauslund, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556–564, 2012.

[3] V. B. Voleti and J.-P. Hubschman, "Age-related eye disease," *Maturitas*, vol. 75, no. 1, pp. 29–33, 2013.

[4] *Prevention of Blindness from Diabetes Mellitus: Report of a WHO Consultation in Geneva*, World Health Org., Geneva, Switzerland, 2006.

[5] E. D. Cole, E. A. Novais, R. N. Louzada, and N. K. Waheed, "Contemporary retinal imaging techniques in diabetic retinopathy: A review," *Clin. Exp. Ophthalmol.*, vol. 44, no. 4, pp. 289–299, 2016.

[6] Early Treatment Diabetic Retinopathy Study Research Group, "Grading diabetic retinopathy from stereoscopic color fundus photographs—An extension of the modified airlie house classification: ETDRS report number 10," *Ophthalmology*, vol. 98, no. 5, pp. 786–806, 1991.

[7] (2003). *Diabetic Retinopathy Screening Services in Scotland: A Training Handbook—July 2003: Page 17*. Accessed: May 3, 2019. [Online]. Available: https://www.ndrs.scot.nhs.uk/?page_id=1609
[8] AO Association. (2014). *Optometric Clinical Practice Guideline: Care of the Patient with Diabetes Mellitus*. Accessed: May 3, 2019. [Online]. Available: https://www.aoa.org/Documents/EBO/EyeCareOfThePatientWithDiabetesMellitus%20CPG3.pdf
[9] C. Sinthanayothin, J. F. Boyce, T. H. Williamson, H. L. Cook, E. Mensah, S. Lal, and D. Usher, "Automated detection of diabetic retinopathy on digital fundus images," *Diabetic Med.*, vol. 19, no. 2, pp. 105–112, 2002.
[10] A. Sopharak, B. Uyyanonvara, S. Barman, and T. H. Williamson, "Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods," *Comput. Med. Imag. Graph.*, vol. 32, no. 8, pp. 720–727, 2008.
[11] X. Zhang, G. Thibault, E. Decencière, B. Marcotegui, B. Laÿ, R. Danno, G. Cazuguel, G. Quellec, M. Lamard, P. Massin, A. Chabouis, Z. Victor, and A. Erginay, "Exudate detection in color retinal images for mass screening of diabetic retinopathy," *Med. Image Anal.*, vol. 18, no. 7, pp. 1026–1043, 2014.
[12] L. Seoud, T. Hurtut, J. Chelbi, F. Cheriet, and J. M. P. Langlois, "Red lesion detection using dynamic shape features for diabetic retinopathy screening," *IEEE Trans. Med. Imag.*, vol. 35, no. 4, pp. 1116–1126, Apr. 2016.
[13] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.
[14] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, and S. Y. Lee, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *J. Amer. Med. Assoc.*, vol. 318, no. 22, pp. 2211–2223, 2017.
[15] Z. Li, S. Keel, C. Liu, Y. He, W. Meng, J. Scheetz, P. Y. Lee, J. Shaw, D. Ting, T. Y. Wong, H. Taylor, R. Chang, and M. He, "An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs," *Diabetes Care*, vol. 41, no. 12, pp. 2509–2516, 2018.
[16] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
[17] A. Ghosh, N. Manwani, and P. S. Sastry, "Making risk minimization tolerant to label noise," *Neurocomputing*, vol. 160, pp. 93–107, Jul. 2015.
[18] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1196–1204.
[19] C. Pelletier, S. Valero, J. Inglada, N. Champion, C. M. Sicre, and G. Dedieu, "Effect of training class label noise on classification performances for land cover mapping with satellite image time series," *Remote Sens.*, vol. 9, no. 2, p. 173, 2017.
[20] L. P. F. Garcia, A. C. P. L. F. de Carvalho, and A. C. Lorena, "Effect of label noise in the complexity of classification problems," *Neurocomputing*, vol. 160, pp. 108–119, Jul. 2015.
[21] R. J. Passonneau and B. Carpenter, "The benefits of a model of annotation," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 311–326, Dec. 2014.
[22] J. Wang and B. Xia, "Relationships of Cohen's kappa, sensitivity, and specificity for unbiased annotations," in *Proc. 4th Int. Conf. Biomed. Signal Image Process.*, 2019.
[23] P. Ruamviboonsuk, K. Teerasuwanajak, M. Tiensuwan, K. Yuttitham, and Thai Screening for Diabetic Retinopathy Study Group, "Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening," *Ophthalmology*, vol. 113, no. 5, pp. 826–832, 2006.
[24] S. Patra, E. M. W. Gomm, M. Macipe, and C. Bailey, "Interobserver agreement between primary graders and an expert grader in the bristol and weston diabetic retinopathy screening programme: A quality assurance audit," *Diabetic Med.*, vol. 26, no. 8, pp. 820–823, 2009.
[25] H. K. Li, L. D. Hubbard, R. P. Danis, A. Esquivel, J. F. Florez-Arango, N. J. Ferrier, and E. A. Krupinski, "Digital versus film fundus photography for research grading of diabetic retinopathy severity," *Investigative Ophthalmol. Vis. Sci.*, vol. 51, no. 11, pp. 5846–5852, 2010.
[26] S. Gangaputra, J. F. Lovato, L. Hubbard, M. D. Davis, B. A. Esser, W. T. Ambrosius, E. Y. Chew, C. Greven, L. H. Perdue, W. T. Wong, A. Condren, C. P. Wilkinso, E. Agrón, S. Adler, and R. P. Danis, "Comparison of standardized clinical classification with fundus photograph grading for the assessment of diabetic retinopathy and diabetic macular edema severity," *Retina*, vol. 33, no. 7, pp. 1393–1399, 2013.
[27] D. Dligach, R. D. Nielsen, and M. Palmer, "To annotate more accurately or to annotate more," in *Proc. 4th Linguistic Annotation Workshop*, 2010, pp. 64–72.
[28] J. Wang, H. Ding, F. A. Bidgoli, B. Zhou, C. Iribarren, S. Molloi, and P. Baldi, "Detecting cardiovascular disease from mammograms with deep learning," *IEEE Trans. Med. Imag.*, vol. 36, no. 5, pp. 1172–1181, May 2017.
[29] H. C. Taylor, S. Binder, T. Das, M. Farah, R. Ferris, P. Massin, W. Mathenge, S. Resnikoff, B. E. Spivey, J. Verdaguer, T. Y. Wong, and P. Zhao. (2017). *ICO Guidelines for Diabetic Eye Care*. Accessed: Mar. 29, 2019. [Online]. Available: http://www.icoph.org/downloads/ICOGuidelinesforDiabeticEyeCare.pdf
[30] S. Pramanik and K. Mondal, "Cosine similarity measure of rough neutrosophic sets and its application in medical diagnosis," *Infinite Study*, vol. 2, no. 1, pp. 212–220, 2015.
[31] P. Li, "A note on the linearly and quadratically weighted kappa coefficients," *Psychometrika*, vol. 81, no. 3, pp. 795–801, 2016.
[32] I. Borg, P. Groenen, and P. Mair, *Applied Multidimensional Scaling and Unfolding*. New York, NY, USA: Springer, 2018.
[33] Z. Zhang, "Model building strategy for logistic regression: Purposeful selection," *Ann. Transl. Med.*, vol. 4, no. 6, pp. 1–7, 2016.

JUAN WANG received the B.S. and M.S. degrees in electrical engineering from the University of Electronic Science and Technology of China, in 2007 and 2010, respectively, and the Ph.D. degree in electrical engineering from the Illinois Institute of Technology, in 2015. She was an Assistant Specialist with the University of California Irvine, in 2016. She is currently an Information Scientist with Delta Micro Technology Inc., Laguna Hills, CA, USA. Her research interests include computer-aided diagnosis, medical imaging, machine learning, and deep learning.

YUJING BAI was born in Nenjiang, Heilongjiang, China, in 1983. She received the M.D. and Ph.D. degrees in ophthalmology from the Zhongshan Ophthalmic Center, Sun Yat-sen University, in 2011.

From 2011 to 2016, she was an Ophthalmologist with Peking University People's Hospital. She has been an Assistant Professor, since 2014. Since 2016, she has been with Shenzhen Sibionics Company Ltd., as the Chief Medical Officer. She is the author of 1 book and more than 80 articles. She holds more than 10 patents. Her research interests include the applications of artificial intelligence in the automated detection of retinopathy, neurodegeneration and regeneration, artificial vision, the molecular mechanisms of age-related macular degeneration, and diabetic retinopathy.

Dr. Bai was a recipient of the Beijing Nova Programme Star, in 2013, the Overseas High-Caliber Personnel of Shenzhen (Level C), in 2016, the Association for Research in Vision and Ophthalmology Travel Grant Award, in 2016, and the Overseas High-Caliber Personnel of Guangdong Province (Young Talents), in 2018.

BIN XIA received the bachelor's degree from Beijing University, in 1988, and the master's and Ph.D. degrees in physics from the University of Washington, in 1990 and 1996, respectively. From 1997 to 2009, he was a Development Engineer with Teradyne Inc. From 2010 to 2016, he was the Director of integrated circuit design with Nurotron Biotechnology Inc. He is the Co-Founder of Shenzhen SiBionics Company Ltd., and Shenzhen SiBright Company Ltd. His research interests include AI applications in medical instruments.