Diabetic retinopathy grading based on Lesion correlation graph

Daming LUO
Graduate School of Information,
Production and Systems, Waseda University
damingsisu@toki.waseda.jp

Sei-ichiro KAMATA
Graduate School of Information,
Production and Systems, Waseda University
kam@waseda.jp

***Abstract*—Diabetic Retinopathy (DR) is a leading cause of blindness. It often happens to people who suffer from diabetes and seldom has early warning signs. Automatically DR detection and severity grading are helpful for clinicians by providing a second opinion. An automatic classification system classifies fundus images into 5 degrees of severity. In this paper, we propose a DR grading model based on lesion correlation graph using Graph Convolution Network (GCN) and Convolution Neural Network (CNN). We extract the irregular lesion region by calculating SURF descriptors in the fundus image. We then clusters descriptors into a number of cluster centroids which is regarded as node representation. With the assistance of GCN, we learn lesion correlation. After fusing correlation information and fundus image feature, which is derived from CNN model, we obtain the final classification result. Furthermore, we provide two evaluation measures: accuracy and Cohen's Kappa value for comparison on different experiments. So far, our model achieves good result in several DR datasets.**

***Contribution*—We introduce the idea of utilizing correlations among lesions learned by GCN to improve the grading result.**

***Keywords*—Diabetic retinopathy, GCN, SURF, Cluster, LD-matrix**

I. Introduction

Diabetic Retinopathy (DR) is a complication of diabetes and the leading cause of blinding among the world. It often happens to people of working age (20 to 74 years old) with diabetes [1]. Among persons with diabetes aged 40 years old or older, 28.5% of individuals have DR [2]. Among those patients, 4.4% developed severe vision problems.

DR causes gradual difference in vasculature structure. The changing of vasculature shows the abnormal of fundus images. [3]. Excessive sugar can block the blood vessels and stop nutrition from reaching cells. Blood vessels block can cause vessel breaking and leaking of different exudates. Lack of nutrition leading eyes to regenerate new blood vessels. Yet newborn blood vessels are too fragile that can cause even more serious hemorrhage and exudates, results in blindness due to photoreceptor cell no longer receive light.

Early DR can be prevented with prompt detection and adequate treatment. However, about 30% of DR patients have no symptoms of visual problems [4]. The lack of symptoms leads to the progression of the disease without treatment. Consequently, regular eye condition monitoring is necessary for early stage DR diagnosis.

DR screening programs have been developed across the world. Mainly in United Kingdom, Ireland, Netherlands [5] and India [6]. Trained clinicians evaluate DR severity based on their professional knowledge. Organizations take eye fundus images from patients and volunteers. Base on the fundus images, clinicians define the lesion parts of DR as Figure 1 depicted.

DR can be classified into 5 degrees according to the existence of lesion parts. The presence of new generated blood vessel decides whether it belongs to proliferative DR (PDR) or non-proliferative DR (NPDR) [7]. Figure 2 demonstrates an example classification result base on the existence of lesion parts. The classification base on the risk of progression. The criterion of DR grading judgement shown in Figure 3.

In this paper, we proposed a lesion correlation based Graph Convolutional Network (GCN) and Convolutional Neural Network (CNN) model to automatically detect multiple lesions and judge the DR severity. Training network takes lesion representation and relation matrix as input, and then output multi-lesion existence relation for final detection. The testing network takes fundus image as input and outputs its severity grading result. From the criterion of how we decide the grading scale shows in Figure 3 and the fundus images in datasets, we learn that the existence and amount of specific lesion directly influence the existence of another lesion. For example, the coexistence of hemorrhage and massive microaneurysms mostly like to incur hard and soft exudates. We utilize GCN to make use of this kind of information and fuse it with image feature representation to predict DR grade.

Our work achieves 3 goals:

1) We firstly develop a lesion representation method to describe the abnormal parts in fundus images. We utilize these representations to GCN and get the correlation of different lesion parts that define the DR grade.
2) We rethink the DR classification as a multi-label detection and classification. Based on the detected multi-lesion we can judge fundus image severity grade.
3) Our model trains and tests on several dataset and performs well. Our work has more extensive ability to fit DR grading task than many existing models.

The reminder of this paper is organized as follows. Section II states related work of DR classification. Section III introduces our network in detail and explains our creative

978-1-7281-9331-1/20/$31.00

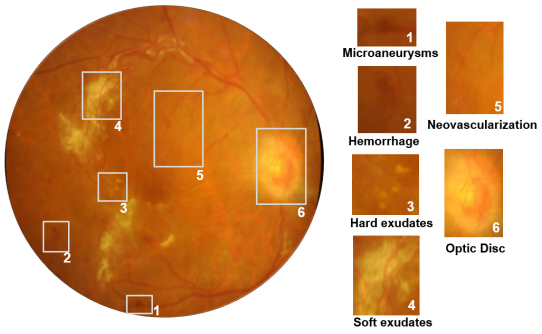

Fig. 1. Example of typical lesions (1-5) and optic disc

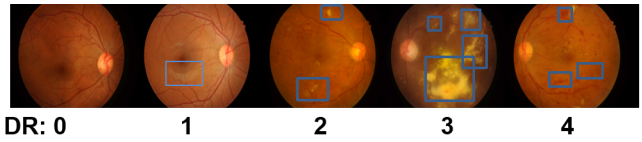

Fig. 2. Example of classification result containing 5 grades

idea. Section IV explains our experiment and furthermore brings comparison experiments. Finally, section V presents the conclusion of the paper.

II. Related Work

Computer-aided diagnosis (CAD) system [8] was developed to provide a second opinion about DR diagnosis. Efforts have been made for recognizing blood vessels, neovascularization, NPDR lesion parts. The basic idea of DR classification is to find out and recognize symptoms region and predict label base on the amount.

SVM classifier-based method does the classification using higher-order spectral information [9]. Another SVM based method suggested using a three-dimensional polynomial kernel [10]. With the development of deep learning methods, medical image processing has a massive improvement. Deep learning has also been widely used in DR classification because lesions show a similar pattern. Recently, there are many proposals about automatically identification of DR using CNN models. A customized CNN model [11] was proposed for DR detections. Transfer learning also was introduced in DR severity grading, it learned the network from ImageNet dataset. Attention mechanism also was added into the neural network and achieved trained and tested on IDRiD dataset [12]. To deal with the problem of trustworthy, uncertainty-aware was proposed to evaluate the confidence of the result [13]. Inspired by face recognition, graph neural network-based method [14] was proposed. This work utilizes correlation to measure inter- and outer- relationship. However, this work has difficulty in capturing lesion regions due to randomly distributed lesions.

In this paper, we mainly focus on the multi-lesion detection and classification, based on the correlation learned by GCN and DR grading criterion, we judge how to classify images into five degrees.

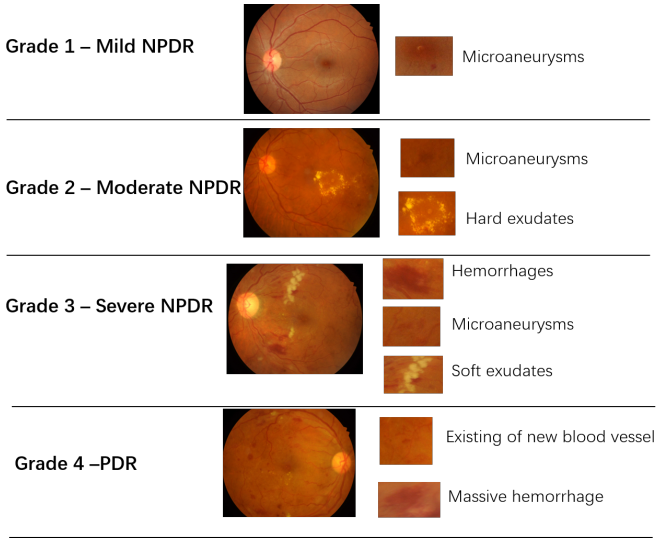

Fig. 3. International criterion on DR image grading

III. Our Method

In this part, we introduce our model for DR grading task. Firstly, the motivation for our method will be introduced. Then nodes feature construction and relation matrix construction will be demonstrated. Furthermore, we will present the network propagation process and prediction procedure.

A. Motivation

For the classification task, effectively capture region of interest to the corresponding label are both important. Diabetic retinopathy classification needs to predict a given fundus image its DR severity, which requires lesion parts extraction and utilize their correlation to predict the label.

In this paper, we use graph structure to state the correlation of symptoms. To better describe lesion as node representation, we use SURF algorithm to extract abnormal parts in a fundus image and cluster all descriptor features into K classes. We take cluster centroids as node representation and lesion cooccurrence as the correlation information to pass through a graph convolutional network, learning the relations of K lesions. The construction process shows in Figure 4. Our work is motivated by the unique of fundus image and DR classification criterion shown in Figure 3.

The appearing locations and amount of lesion vary in every image because blood vessel organize differently for every patient. In Figure 2, we can see that abnormal parts have no identical size, shape or color in the blue box.

The DR classification criterion as shown in Figure 3 reveals that irregular region often incurs another lesion, suggesting their correlation can directly impact the classification result. Main network structure is shown in Figure 7.

B. Node representation

DR lesion detection often focuses on the region of interest. Different lesion shows different size and shape. As shown in Figure 1, microaneurysms shows little red spot while

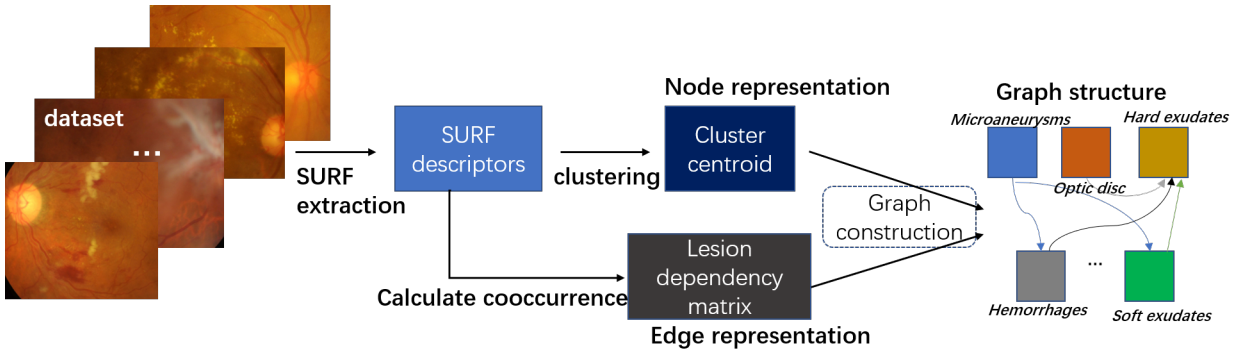

Fig. 4. Illustration of how we construct lesion correlation graph. From lesion images we calculate SURF descriptors, then we do cluster to get node representation. We calculate cooccurrence to learn Lesion dependency matrix as edge in graph. Node and edge information construct the graph.

hemorrhage appears with red irregular area. Exudates also has spot or large area appearance. Our target is to find them and classifies lesions into right categories. There is no such dataset provide lesion segmentation mask, that leads us to think extracting lesion parts using traditional methods.

Trying to find out all irregular regions which has no exact location nor appearance in an image, scale-invariant feature transform (SIFT) [15] can be used. SIFT can extract scale-invariant features which suit for most lesion region. Furthermore, Speeded Up Robust Features (SURF) [16] improves detection accuracy and reduce the detection time. In this work, we adopt SURF algorithm to find out all irregular region show in fundus image. Figure 5 shows comparison result between both algorithms, showing SURF descriptor has a better performance.

To ignore the black background and sudden sharp edge which caused by fundus image standard influence SURF description. We preprocess images by finding the inscribed circle and cut the rest.

For one fundus image, we take K SURF descriptors to capture most distinct abnormal location as $X = [x_1, x_2, ...x_K]$, where $x_i \in \mathbb{R}^{64}$ as one SURF descriptor. Totally, we can get over 26000 descriptors.

Purpose of lesion extraction is to find a way to mapping lesions into vertexes. We need to utilize these SURF descriptors and aggregate them into K classes, which we can consider K classes as lesion labels and node representation. To make

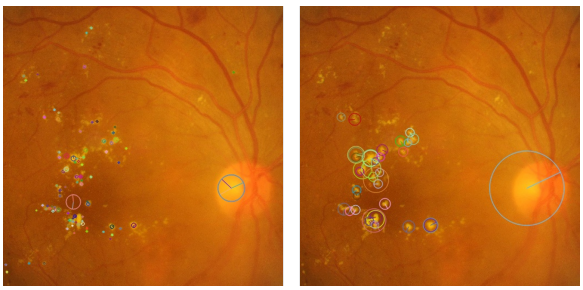

Fig. 5. comparison on SIFT(left) and SURF(right) description, SURF describes lesions more clearly

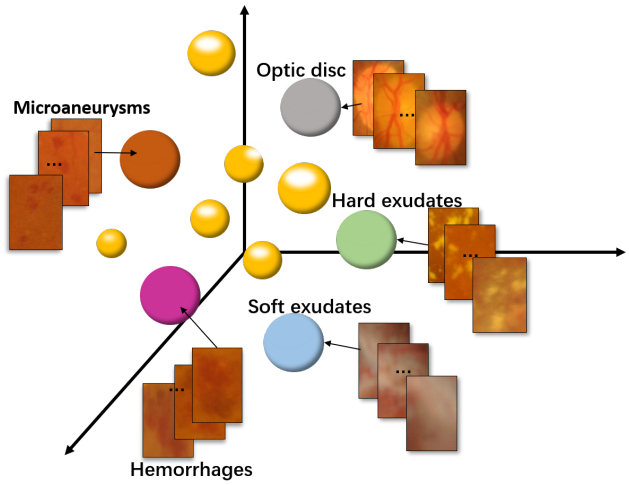

Fig. 6. Illustration of correspondence between lesion sub-images and cluster centroids.

this aggregation, we use Euclidean distance in 64-dimension to calculate similarity within descriptors. We collect all descriptors and cluster them into K classes. In our experiment, we use K-means to make K classes as $\hat{X} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_K]$ where $\hat{x}_i \in \mathbb{R}^{64}$ is the cluster centroid of class.

Furthermore, we consider correspondence of lesion symptoms and clusters as in Figure 6. This figure shows correspondence between lesion parts and cluster centroids. This is a simple demonstration of possible situation. Due to the data complexity, exact correspondence remains undiscovered. In our work, we choose cluster number K to be 20 to be more robust and to get a better experiment result.

C. Lesion dependency matrix construction

The propagation of GCN requires an adjacency matrix to state edge information. However, in any standard DR dataset, this correlation is not provided. In our model, we adopt Lesion dependency matrix (LD matrix) $A \in \mathbb{R}^{K \times K}$ to be the reference in GCN.

To calculate cooccurrence, we first take the clustering result as inference. For each fundus image, we have K SURF

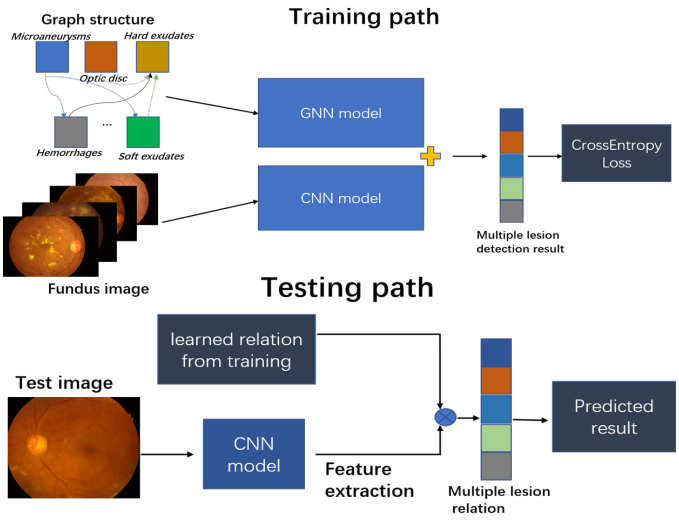

Fig. 7. Structure of our network. We divide our model into training and testing parts. Model updates base on CrossEntropy Loss. Based on the learned relation from training, we can predict DR grade of the test image.

descriptors. After the clustering process described before, we can divide SURF descriptors into specific clusters and regard it as lesion label $\hat{x}_i$. We then calculate the cooccurrence of different lesion labels.

For element in LD matrix,$A_{i,j}$ indicates that times lesion $\hat{x}_j$ appear given the situation of existing $\hat{x}_i$.

D. Learning correlation through GCN

After getting node representation and adjacency matrix, we can learn the correlation of fundus images. Unlike normal CNN network, GCN updates node representation which does not have spatial consistency. The propagation of graphs base on convolution theorem [17] as:

$$f * h = \mathcal{F}^{-1}[\hat{f}(\omega)\hat{h(\omega)}] \tag{1}$$

where $*$ stands for convolution operation and $\mathcal{F}^{-1}()$ is inverse Fourier transform. The message passing from layer to layer can be described as:

$$L^{l+1} = h(L^l, A) \tag{2}$$

$h(\cdot,\cdot)$represent the updating process, which takes feature descriptions $L^l \in \mathbb{R}^{K\times d}$ and LD matrix $A \in \mathbb{R}^K \times K$ as inputs and updates the node features as $L^{l+1} \in \mathbb{R}^{K\times d'}$.

E. Feature fusion

GCN learns the relation of cluster centroids, which regarded as lesion labels. After learning from GCN, the model can only get relation hidden in the image. However, images can tell more information. CNN can extract and combine local information by sliding windows. In DR classification, each fundus image also contains its own feature. From this aspect, we adopt this multi-label model [18] to do feature fusion.

This multi-label model uses GCN and ResNet101 to separately learn label correlation and image-wise feature.

Finally, we concatenate two features and after linear transformation we can get $x \in \mathbb{R}^5$ which is the output of our model. We can get the prediction result base on probability.

IV. Experiment

In this part, we will describe the implementation details and evaluation method. Then illustrate comparison experiment result given different dataset and variables. Finally, we will make the graphic statement for presenting and followed by analysis.

A. Implementation details

Our model contains two branches: Graph neural network and Resnet101 [19]. GCN takes cluster centroids $\hat{X}$ as input and expands dimensionality from 64 to 2048. For ResNet101, we take fundus images as input, but firstly resize images into 448×448 with random flips. Then extract local relation and output feature vector Y. We apply SGD as optimizer where we set learning rate 0.001. Learning rate decays by a factor of 10 for every 10 epochs. The loss function we use here is CrossEntropy Loss as:

$$\mathcal{L}(x, y) = -log(\frac{\exp x[y]}{\Sigma_j \exp x[j]}) \tag{3}$$

where x is predicted label and y is actual label. We train our model in different dataset with 50 epochs. The network is implemented on Pytorch.

B. Evaluation method

In evaluation period, we adopt two criterions to judge whether the result is good or not. One is accuracy, another is Cohen's Kappa value with quadratic weight (QWK). Accuracy test how many fundus images network correctly evaluate, which is intuitive for most tasks. However, for medical image processing, Cohen's Kappa [20] value shows even more importance as:

$$kappa = (p_o - p_e)/(1 - p_e) \tag{4}$$

This equation based on confusion matrix C, where p_o is the empirical probability calculated by

$$p_o = \frac{\Sigma C_{i,i}}{\Sigma C}, p_e = \frac{\Sigma(C_{i,} \times C_{,i})}{(\Sigma C)^2} \tag{5}$$

where C is the confusion matrix. p_o is the empirical probability denotes accuracy and p_e is the penalty term that gives low score when model is highly biased. QWK value further highlight the evaluation criterion. Its values range from -1 (complete disagreement) to 1 (perfect agreement). In medical image processing, even specialist or doctors can not totally sure about their diagnosis about grading scale. So, we want the result to achieve a reasonable range.

Figure 8 illustrates on APTOPS 2019 dataset and on EyePACS dataset. We train and test our model with 50 epochs. In loss and accuracy curves, we record both training and testing process. Blue curve stands for training process and red curve stands for testing process.

on APTOS 2019 dataset (row1) and EyePac dataset (row2)

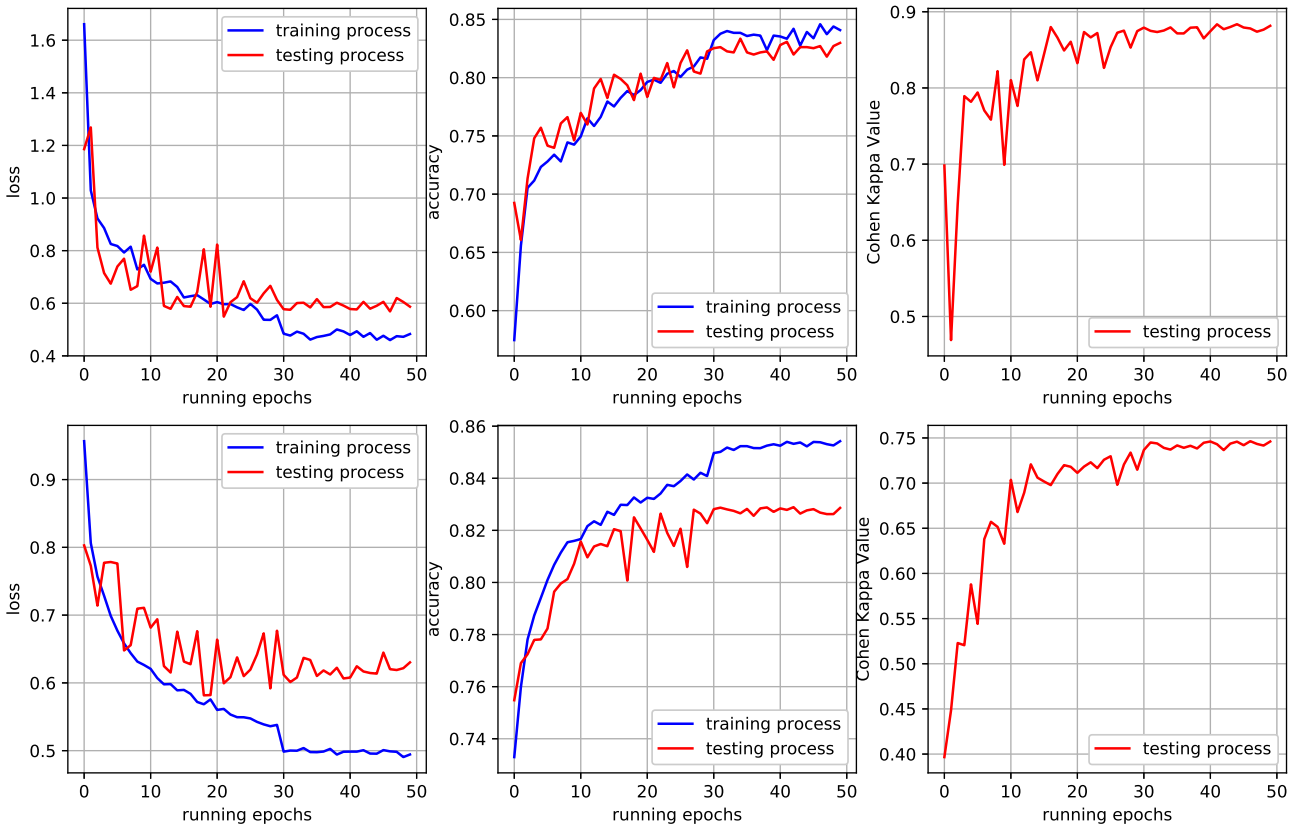

Fig. 8. An illustration on APTOPS 2019 dataset and on EyePACS dataset. We train and test our model with 50 epochs. In loss and accuracy curves, we record both training and testing process. Blue curve stands for training process and red curve stands for testing process. Cohen's Kappa value shows in the third column, which appears in red curve

Besides the training and testing process of our model, we also illustrate other visualization of results. Figure 9 is the confusion matrices of four datasets. Confusion matrices give us a clear view about how the system do the classification. Data gathered around the diagonal means good quality of classification. Especially in medical images, adjacent grades are classified into nearby labels is reasonable. Figure 10 plots Receiver Operating Characteristic (ROC) curve for different datasets. ROC curve shows the ability of the classifier. Area Under Curve (AUC) quantify this ability. The more AUC area, the better classification result. In the figure, we plot the ROC curve in different colors for each grade. Basically, ROC and AUC only evaluate binary classification. Here we extend these criteria for multi-class grading.

EyePACS [21] is the dataset that describes retinal images, especially diabetic retinopathy, which provided by California Healthcare Foundation. It contains 35127 training images and 53577 testing images, in experiment, we only use training dataset. Messidor-2 dataset is a collection of DR examinations, each consisting of two macula-centered eye fundus images (one per eye). It is provided by the Messidor program partners. It contains 1749 training images. Indian Diabetic Retinopathy Image Dataset (IDRiD) dataset [22] is a 2018 Kaggle competition dataset. In DR grading subtask, it contains 413 training images and 104 testing images. APTOS 2019 dataset is provided by 2019 Kaggle competition APTOS. It contains 3649 training images and 1929 testing images.

To evaluate the performance of our model, we deploy different datasets and compare with other methods. The result is shown in Table I. We evaluate experiment by accuracy and quadratic weight Kappa value.

We evaluate our model and it has a good performance in many existing DR datasets. However, model cannot perform well in a small and unbalanced dataset. In EyePACS dataset, we achieve good result while QWK is not quite high because of the unbalanced data. In IDRiD dataset, model does not achieve the best due to the small dataset size and unbalanced distributed dataset. APTOS dataset contains thousands of images and rather evenly distributed. In APTOS 2019 dataset, our model can achieve accuracy to 0.8480, and overpass 0.90 QWK value, which in medical image processing, 0.75 QWK can be regarded as a believable prediction. [23] compares different deep learning models and find out the best accuracy. In this competition, rank 1 adopts ensemble models and achieves 0.9361 QWK value.

Besides the references shown in Table I, there are several

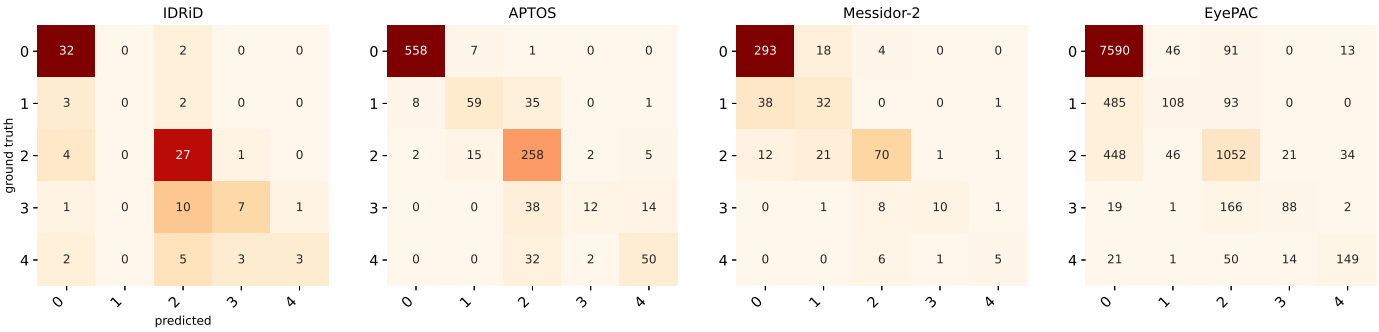

Fig. 9. Confusion matrices of four datasets, column shows predict number and row shows ground truth label

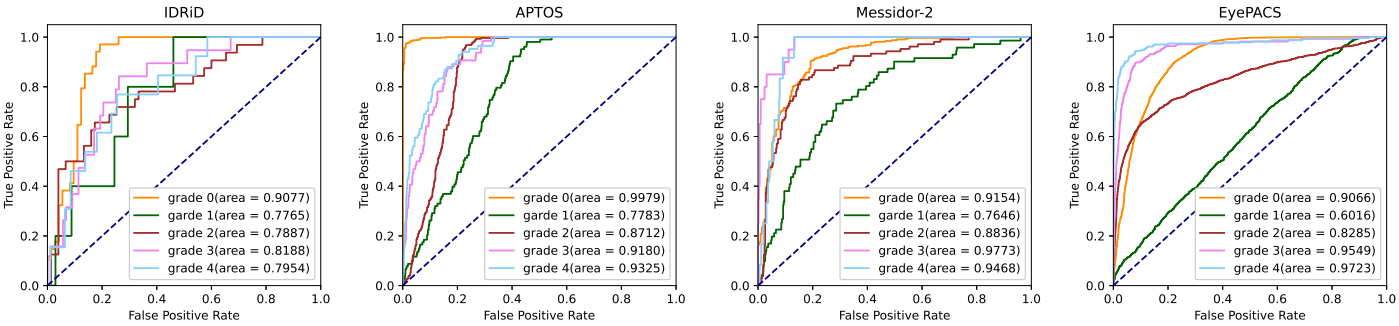

Fig. 10. ROC curves of four datasets, horizontal axis shows False Positive Rate and vertical axis shows TruePositive Rate. ROC curve of different gardes are shown in different color

other experiments focus on DR grading yet using other evaluation methods like sensitivity and specificity, or AUC. However, many of them are focusing on binary classification between NPDR and PDR. [24] does experiments on Messidor-2 dataset get 0.966 AUC. Compare to our multi-class experiment, we do not think they are comparable. [25] uses deep learning models to get sensitivity on three categories, No DR, Mild and Severe. No DR and Severe classes are around 0.80 but Mild grade is around 0.30. Compare to our ROC curve, we also get the result that severe and healthy grade are distinguishable but mild grades are not so separable. [25] also plots confusion matrix of EyePACS dataset experiment. However, instead of showing five grades, confusion matrix only shows three classes, which neglect grade 4 and combine R2 and R3 together.

C. Analysis

In the experiment section, we show the performance of our model by experiment result comparison and other evaluation standards. In Table I, we compare different results relate to different methods. We get reasonable results. However, some research like [23] on APTOS and [13] on IDRiD get a much higher result. Despite their higher resolution figure and ensemble model, it still tells us that our network has a lot of improvement to do.

Also, our method face the problem that it cannot distinguish grade 0 and grade 1 well. In Figure 9, number in confusion matrices $CM_{1,1}$ may less than $CM_{1,0}$. It means model often classify a grade 1 image into grade 0. Network cannot gather specific lesion region in fundus images. Naturally, it decides

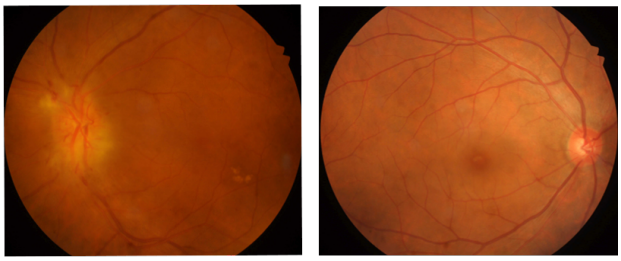

Fig. 11. comparison of a grade 4 (left) and grade 0(right) fundus image

image to be grade 0. In IDRiD confusion matrix, we can see that grade 4 are classified in grade 3 grade 2, even grade 0. Bad quality of fundus images can mislead our system. Shown in Figure 11, a grade 4 image seems missing distinct lesions and similar to a grade 0 image. This happens a lot in several datasets. In Figure 10, ROC curve reveals the system performance. Yet dealing with grade 1, which is the green curve, the performance is weak. Combining confusion matrices, we know that model struggle in classifying grade 1 from grade 0.

V. Conclusion

Diabetic retinopathy is a threatening disease that endanger eyesight to people who have diabetes. For it is hard to detect and diagnosis, we usually need CAD to provide a second opinion. However, the lesion symptoms in fundus image do not show any pattern like fixed size, shape or color.

In this paper, we propose a lesion correlation based GCN

TABLE I
COMPARISON EXPERIMENT ON DIFFERENT DATASET. WE TRY ON ACCURACY, QWK FOR EVALUATION

Methods and Dataset	Accuracy	QWK
EyePACS dataset		
Lesion correlation graph model(ours)	0.8528	0.7850
Cristina González-Gonzalo [26]	-	0.72
DR—GRADUATE [13]	-	0.74
IDRiD dataset		
Lesion correlation graph model(ours)	0.6699	0.7175
Fundus image classification [14]	0.793	-
DR—GRADUATE [13]	-	0.84
Messidor-2 dataset		
Lesion correlation graph model(ours)	0.7352	0.7533
DR—GRADUATE [13]	-	0.71
APTOS 2019 dataset		
Lesion correlation graph model(ours)	0.8480	0.9066
Deep Transfer Learning [23]	0.979	-

and CNN model to automatically predict DR severity. We utilize GCN to learn relations between lesion part which calculated by SURF descriptor and further clustered. Combined with image-wise feature vector that extracted by ResNet101, we get the fusion feature that from correlation and local information. We test network on different dataset and do achieve a good result. In assess part, we adopt two evaluation criterion that are accuracy and Cohen's Kappa value. Compare to other methods, our model has more universality to run well on several datasets.

However, our model has a drawback. Network performs not well on a small and unbalanced dataset. Also, bad quality of fundus images like Figure 11 can jeopardize the grading performance. Our model can find healthy and severe fundus images well, but has difficulty deciding mild images shown in Figure 9 and Figure 10. Despite this disadvantage, Lesion Correlation Graph still works well and surpass many methods.

Future problem to be solved is to consider more lesion extraction methods. Also, we shall utilize more effective clustering methods for node representation. Furthermore, we will focus on the exact correspondence between lesion parts and cluster centroids. Last but not least, we will continue to focus on classify mild fundus images precisely.

REFERENCES

[1] R. C. of Ophthalmologists, *Diabetic retinopathy guidelines*. Royal College of Ophthalmologists, 2012.

[2] R. P. Singh, *Managing Diabetic Eye Disease in Clinical Practice*. Springer, 2015.

[3] F. Bandello, M. B. Parodi, P. Lanzetta, A. Loewenstein, P. Massin, F. Menchini, and D. Veritti, "Diabetic macular edema," in *Macular Edema*. Karger Publishers, 2010, vol. 47, pp. 73–110.

[4] U. R. Acharya, E. Y.-K. Ng, J.-H. Tan, S. V. Sree, and K.-H. Ng, "An integrated index for the identification of diabetic retinopathy stages using texture parameters," *Journal of medical systems*, vol. 36, no. 3, pp. 2011–2020, 2012.

[5] M. D. Abramoff and M. S. Suttorp-Schulten, "Web-based screening for diabetic retinopathy in a primary care population: the eyecheck project," *Telemedicine Journal & e-Health*, vol. 11, no. 6, pp. 668–674, 2005.

[6] I. D. Atlas, "Brussels, belgium: international diabetes federation; 2013," *International Diabetes Federation (IDF)*, p. 147, 2017.

[7] J. F. Arévalo, *Retinal and choroidal manifestations of selected systemic diseases*. Springer Science & Business Media, 2012.

[8] R. Pires, H. F. Jelinek, J. Wainer, and A. Rocha, "Retinal image quality analysis for automatic diabetic retinopathy detection," in *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2012, pp. 229–236.

[9] R. Acharya, C. K. Chua, E. Ng, W. Yu, and C. Chee, "Application of higher order spectra for the identification of diabetes retinopathy stages," *Journal of medical systems*, vol. 32, no. 6, pp. 481–488, 2008.

[10] U. R. Acharya, E. Y.-K. Ng, J.-H. Tan, S. V. Sree, and K.-H. Ng, "An integrated index for the identification of diabetic retinopathy stages using texture parameters," *Journal of medical systems*, vol. 36, no. 3, pp. 2011–2020, 2012.

[11] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016.

[12] Z. Zhao, K. Zhang, X. Hao, J. Tian, M. C. H. Chua, L. Chen, and X. Xu, "Bira-net: Bilinear attention net for diabetic retinopathy grading," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1385–1389.

[13] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Â. Carneiro, A. M. Mendonça, and A. Campilho, "Dr— graduate: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images," *Medical Image Analysis*, p. 101715, 2020.

[14] A. Sakaguchi, R. Wu, and S.-i. Kamata, "Fundus image classification for diabetic retinopathy using disease severity grading," in *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology*, 2019, pp. 190–196.

[15] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[16] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[18] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[21] J. Cuadros and G. Bresnick, "Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening," *Journal of diabetes science and technology*, vol. 3, no. 3, pp. 509–516, 2009.

[22] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, 2018.

[23] N. E. M. Khalifa, M. Loey, M. H. N. Taha, and H. N. E. T. Mohamed, "Deep transfer learning models for medical diabetic retinopathy detection," *Acta Informatica Medica*, vol. 27, no. 5, p. 327, 2019.

[24] J. Sahlsten, J. Jaskari, J. Kivinen, L. Turunen, E. Jaanio, K. Hietala, and K. Kaski, "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[25] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," *AMIA summits on translational science proceedings*, vol. 2018, p. 147, 2018.

[26] C. González-Gonzalo, B. Liefers, B. van Ginneken, and C. I. Sánchez, "Improving weakly-supervised lesion localization with iterative saliency map refinement," 2018.