DIABETIC RETINOPATHY USING BIG DATA ANALYTICS

Victor Sarmacharjee, Sanchari Dey, Ashim Kumar Buragohain, Purnendu Bikash Acharjee

Department of Information Technology, The Assam Kaziranga University

Abstract- Diabetic Retinopathy is one of the major cause of blindness in diabetics. We can evolve an impulsive method diabetic retinopathy treatment system . In this diabetic retinopathy project, we will classify whether the patient has diabetic retinopathy or not. In this paper, we examined a large number of diabetic data sets for a variety of patients in order to conduct Big data analytics. We combined and collected various inceptions of diabetic information for several research papers in this project, ranging from primary and secondary treatment information to administrative data, in order to examine novel convictions of patient care processes. In this paper the proposed method we aims to detect the stages of various Diabetic Retinopathy by using Convolutional Neural Network (CNN) to spontaneously diagnose and therefore classify high-resolution fundus images of the disease into 5 stages based on intensity.

Keywords: Diabetic Retinopathy, Machine Learning, Image Recognition & Classification

INTRODUCTION

Micro aneurysms (MA), dot & blot Hemorrhages (HE) are early symptoms of diabetic retinopathy and are caused by excessive permeability and non-perfusion of capillaries. Non-proliferative Diabetic Retinopathy is the name given to the early stages of the disease (NPDR). Fluid leakage from retinal capillaries indicates further disease development, which may lead to death. From retinal capillaries Fluid leaking specify additional progression of the disease which may lead to sight threatening diabetic retinopathy, in the area of the most Severe vision if the leakage is located. In this paper the proposed method we aims to detect the stages of various Diabetic Retinopathy by using Convolutional Neural Network (CNN) to spontaneously diagnose and therefore classify high-resolution fundus images of the disease into 5 stages based on intensity and we develop a network with CNN architecture and data augmentation to identify the complicated features involved in the classification task such as No DR , Mild, Moderate, Severe and Proliferative . In the fundus image classification the main problem is high variability especially in the case of Proliferative diabetic retinopathy which exist retinal proliferation of new blood vessels and retinal detachment.

LITERATURE REVIEW

The combined methods of volume data visualization and data analysis to help better diagnosis and treatment of human retinal diseases and . such diseases can be identified by measuring the thickness of various retinal layers by using optical coherence tomography,[1](Aishwarya Iyer, S. Jeyalatha and Ronak Sumbaly. 2015. Diagnosis Of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1.). To extend and generalize an older CNN approach to support better performance in a clinical setting through performance enhancements and graceful handling of inherent noise in OCT data by considering statistical characteristics at multiple levels of resolution.. A feature, such as a retinal layer, can therefore be modelled . The algorithm begins with a new vessel enhancement method based on a modified corner detector. Later, a weighted version of the vessel enhancement is combined with morphological operators, to detect the four main vessels orientations {0°, 45°, 90°, 135°},[2](Galega officinalis. May 2009. Management Of Type 2 Diabetes Mellitus 4TH Edition.) These four image functions have all the necessary information to determine initial optic disc localization, resulting in two images that are respectively divided along the vertical or horizontal orientations with different division sizes. Each division is averaged creating a 2-D step function, and a cumulative sum of the different sizes step functions is calculated in the vertical and horizontal orientations, resulting in an initial optic disc position. An algorithm that uses a graph-based algorithm to segment both vessel edges simultaneously to calculate the width of retinal vessels in fundus photographs. The simultaneous two boundary segmentation problem is first modelled as a two slice, 3-D surface segmentation problem [3],(which is then transformed into the problem of computing the smallest closed set in a node-weighted graph. A preliminary segmentation is carried out

METHODOLOGY

On the basis of figure the DIABETES RETINOPATHY DISEASE DETECTION process at the center of figure. Two entities that are available which are USER and ADMIN.In the Context Diagram there are 11 data flows available out of which only two outgoing data flow from ADMIN which consist of UPDATE INFORMATION and UPDATE QUESTIONAIRRES. While from USER, there are only five outgoing data flow which consist of SYSTEM EVALUATIONS ,LOGIN, REGISTER, PERSONAL DETAILS and ANSWER OF QUESTIONAIRRES. Thus ingoing data flow, USER have only have two which is DIABETES INFORMATION and RESULTS while ADMIN have only USER INFORMATION as ingoing data flow is as follows:

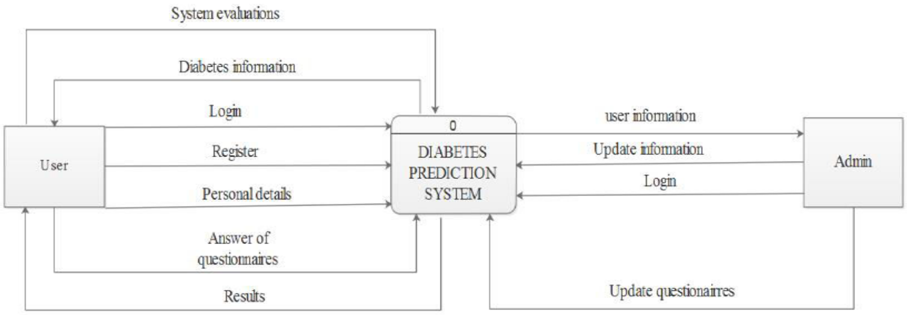


Fig 1.1. Context Diagram [Courtesy : Diabetic Prediction System Project 2017]

Preprocessing : We just need a little preprocessing since we're using the CNN algorithm. Only resizing and rotating the images is needed in this case. It is necessary to rotate the image at 45°C, 90°C, and 180°C. These rotations would increase the number of images in the training dataset, allowing for a more effective training model. The pre-processed dataset will be given to the CNN algorithm after the pre-processing stage.

Training phase : Here the input of the CNN algorithm will be the pre-processed training data in training phase whereas the input goes repetitively through each layer it will recognize the features in the images.

Convolution layer : Let us Consider the N*N squared neuron layer, which is then accompanied by the convolution layer. The filter we're using is the m*m filter, which is. The convolutional layer's contribution would then be of the size (N-m+1)*(N-m+1).

Max pooling layer : This layer will select a k*k region and the o maximum value in that certain region and if N*N is the input layer, then the output will be (N/k)*(N/k) and the output will be determined by reducing each single k*k region by maxfunction

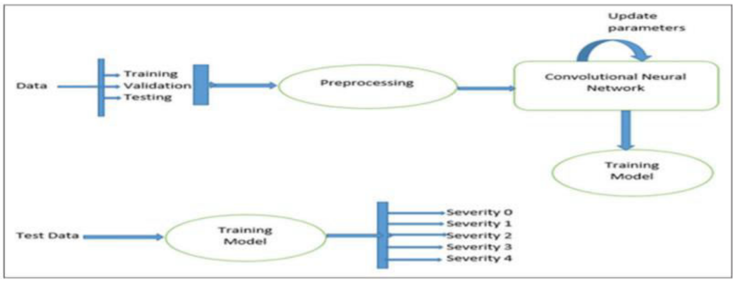


Fig 1.2. Workflow Diagram [Courtesy : Diabetic Prediction System Project 2017]

A Convolutional Neural Networks (CNNs) is put forward to automatize the mechanism of Diabetic Retinopathy screening using the color fundus retinal photography as an input. In this paper the proposed method we aims to detect the stages of various Diabetic Retinopathy by using Convolutional Neural Network (CNN) to spontaneously diagnose and therefore classify high-resolution fundus images of the disease into 5 stages based on intensity and we develop a network with CNN architecture and data augmentation to identify the complicated features involved in the classification task such as No DR , Mild, Moderate, Severe and Proliferative . The main problem in the fundus image classification is high division mostly in case of Proliferative diabetic retinopathy CNN consists of several layers which include convolutions layer pooling, and fully connected layers. The output of a 3×3 convolution layer is defined as follows

$$y(l,m,n) = \sum_{k=1}^{3} \sum_{i=1}^{3} \sum_{j=1}^{3} w(l,i,j,k)x(i+m-1,j+n-1,k) + b(l) \qquad \text{…………..(1)}$$

A fully connected layer contain a set of neurons that are connected with all the activation maps of the previous layers. The outputs of both convolutional and fully connected layers are typically processed using a Rectified Linear Unit (ReLU) which is defined as follows:

$$a_i = \begin{cases} b_i & b_i > 0 \\ 0 & b_i < 0 \end{cases} \qquad \text{…………..(2)}$$

Thus the soft max activation function is established at the end of the network to enumerate the probability distribution of each final fully connected layer output which is given in the below The cross entropy loss *e*, illustrate the deviation of the predicted outputs from the expected outputs, which is : :

$$e = -\sum_{j=0}^{L} \hat{a_j} \log(a_j) \qquad \text{…………(3)}$$

CNN was used in this study to successfully identify DR subjects into non-DR, moderate DR, and serious DR, as well as stage the disease in a mechanised manner. Each image has been resized to a maximum of 1024 pixels in width and height. The smallest category yielded around 8527 images, resulting in a larger collection of 10000 images that are comparable in size to the other categories. Again fundus images were provided to a set of five consecutive stages of convolutional layers with a single 2×2 max pooling layer in between as shown below

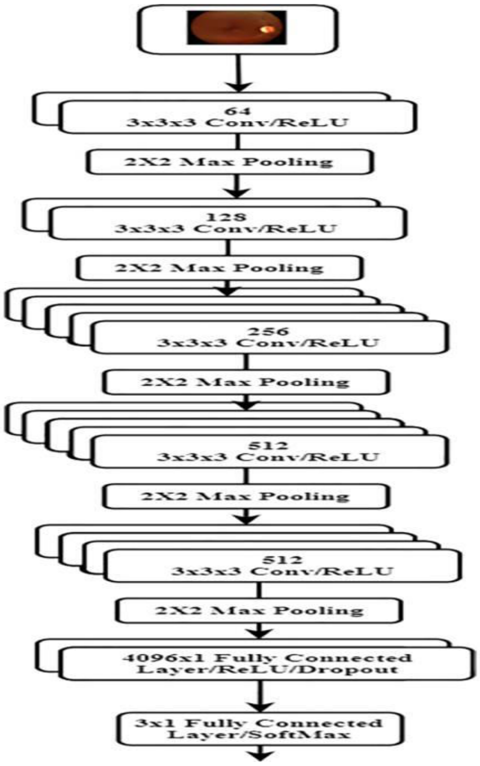


Fig 1.3 : CNN flow diagram

Dataset description

Kaggle has fundus photos that were used in our project's training mode. Almost 8527 high-resolution fundus images were chosen from a Kaggle dataset of 10000 images taken over time with various models and types of cameras in a variety of clinics. Overexposed even more. Images were graded on a scale scale of 0 to 4 .Table1 shows the class labels or score, the similar DR stage, and class size for the dataset.

Table 1.1: Dataset

SL.NO	DR STAGE	CLASS SIZE	CLASS LEVEL
1	NO DR	300	0
2	MILD DR	277	1
3	MODERATE DR	302	2
4	SEVERE DR	194	3
5	PROLIFERATIVE	266	4

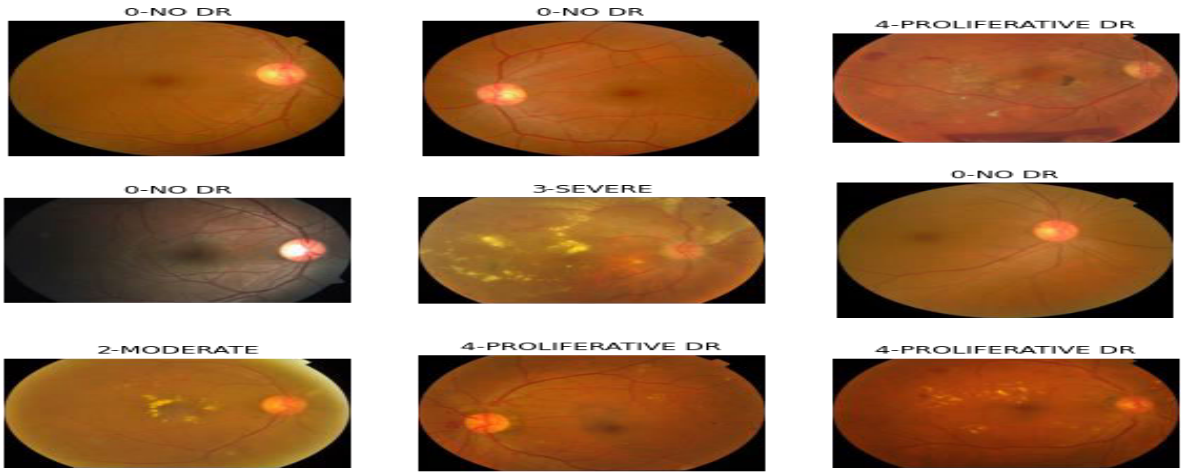
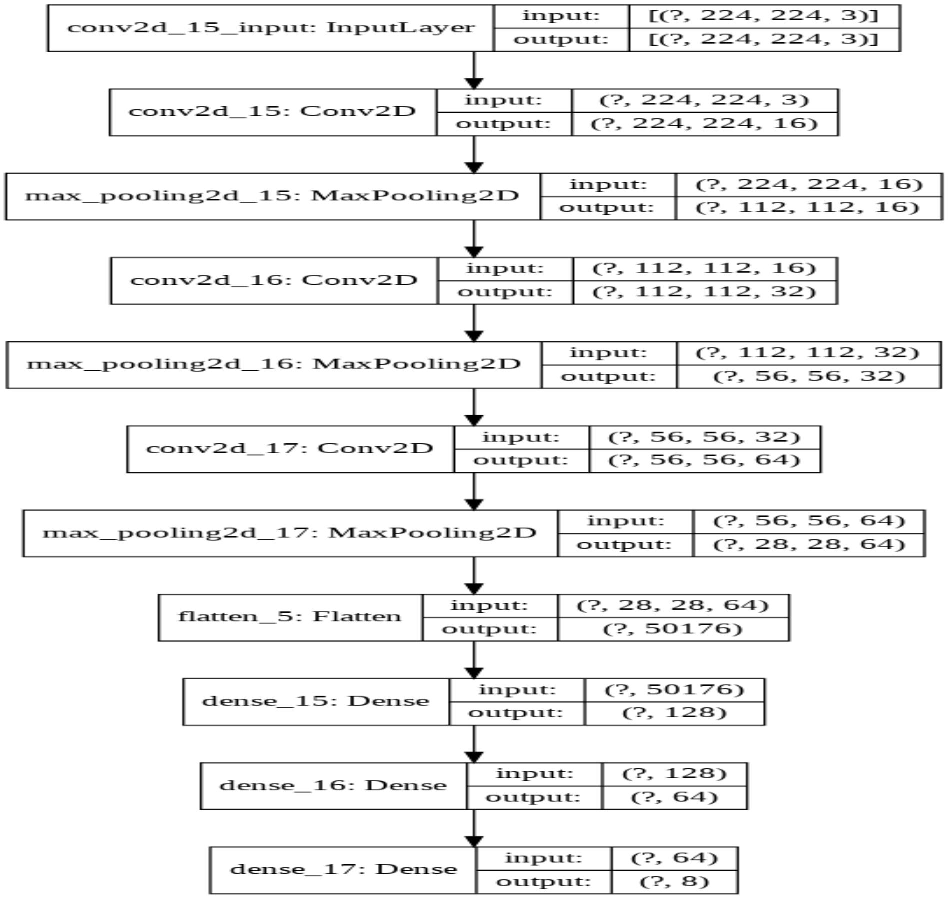


Fig 1.4: sample fundus images after training

Clearly, the proposed CNN architecture of our project is a modified version of the Trained dataset (2019), with two convolutional layers added to the middle two stages and a three-neuron layer replacing the final completely connected layer with 1000 neurons.



conv2d_15_input: InputLayer	input:	[(?, 224, 224, 3)]
	output:	[(?, 224, 224, 3)]
conv2d_15: Conv2D	input:	(?, 224, 224, 3)
	output:	(?, 224, 224, 16)
max_pooling2d_15: MaxPooling2D	input:	(?, 224, 224, 16)
	output:	(?, 112, 112, 16)
conv2d_16: Conv2D	input:	(?, 112, 112, 16)
	output:	(?, 112, 112, 32)
max_pooling2d_16: MaxPooling2D	input:	(?, 112, 112, 32)
	output:	(?, 56, 56, 32)
conv2d_17: Conv2D	input:	(?, 56, 56, 32)
	output:	(?, 56, 56, 64)
max_pooling2d_17: MaxPooling2D	input:	(?, 56, 56, 64)
	output:	(?, 28, 28, 64)
flatten_5: Flatten	input:	(?, 28, 28, 64)
	output:	(?, 50176)
dense_15: Dense	input:	(?, 50176)
	output:	(?, 128)
dense_16: Dense	input:	(?, 128)
	output:	(?, 64)
dense_17: Dense	input:	(?, 64)
	output:	(?, 8)

Fig 1.5: Architecture Layer Diagram

RESULT & DISCUSSION

Now we discussing about the results we obtained from our experiments upon the implementation of this system. We have divided our dataset into two parts- training and testing dataset. In this chapter we will show the outcome of the training and testing dataset. As mentioned before we have used four machine learning algorithms. First, we trained our dataset with the CNN algorithm and its architecture and then we built a model. Then, we tested our testing dataset in this model. If the test set accuracy is near to train set accuracy then we can conclude that we built a good model.

In CNN algorithm we got 0.3732 accuracy and In MobileNet model , for V2_model-1 we got 0.4062 accuracy and for V2_model-2 we got 0.3414.In EfficientNet model , for B7_model-1 we got 0.3781 accuracy ,for B7_model-2 we got 0.3385. for B0_model-1 we got 0.3125 accuracy ,for B0_model-2 we got 0.3099. In ResNet model , for 50_model-1 we got 0.3469 accuracy ,for 50_model-2 we got 0.3125,. for 150_model-1 we got 0.3594 accuracy ,for 150_model-2 we got 0.2917

Models	**Accuracy**		**Loss**
CNN	**0.3732**		**2.9201**
MobileNet	V2_model-1	**0.4062**	**1.4163**
	V2_model-2	**0.3414**	**1.4721**
EfficientNet	B7_model-1	**0.3781**	**1.4072**
	B7_model-2	**0.3385**	**1.4553**
	B0_model-1	**0.3125**	**1.4783**
	B0_model-2	**0.3099**	**1.4790**
ResNet	50_model-1	**0.3469**	**30.1281**
	50_model-2	**0.3125**	**20.7840**
	150_model-1	**0.3594**	**47.8006**
	150_model-2	**0.2917**	**128.1882**

Comparison between all the models

CONCLUSION

This chapter contains the difficulties, future works which will give the summary of our thesis work and also give the indication of our future plan with our thesis project

5.1 Difficulties

While working, we encountered numerous challenges.

First and foremost, there is much more work to be done before an algorithm like this can be widely employed. Second, if we could manage more of our training data, we could improve the accuracy of our algorithm by training it more. Furthermore, we encountered some difficulties while selecting algorithms; for example, we initially chose the Nave Bayes Algorithm, but it did not work out, and it was extremely tough for us to select specific machine learning algorithms that would provide accurate disease classification. Furthermore, we used simple feature selection and scaling strategies; however, we might likely achieve higher results by including mo re advanced feature selection and generation approaches.

For model parameters, we looked at tiny subspaces. There's a chance that alternate parameter spaces would provide better-performing models. We had difficulty implementing the categorization for the small amount of data in our dataset.

5.2 Future Work

For any research, there is always room for improvement. Ours is not an exception of that. We have found some areas where this system can be improvised:

1. Work on more Categories: This can be improvised with a lot more categorized such as according to ages, genders, background studies, working facilities and so on. As an example, A matured man from the IT background has different eye condition that a matured women from Teaching background.
2. Work on more classes: As we working on only two classes whether it is good or bad. In future we are going to add more classes like low, medium, severe condition. In this way patients can know about their condition more accurately
3. Different Algorithms: We have used CNN algorithm and also specialised CNN architecture (Mobile Net , ResNet, Efficient Net)we have used be to train our dataset in order to improve the algorithm.
4. More Analysis: To achieve more accuracy we could use more dataset. If we use huge amount of dataset, machine will train more and it would give us more accurate prediction and accuracy.
5. Hardware Implementation: A hardware product can be the best solution for patient. So, we are looking forward to build a hardware system where we can use our model to implement results on diabetic patients easily. We can then input the data of the patient and wait for the machine to create a new prescription integrated with Doctor’s suggestion.
6. Software Implementation: We can build a website or an android app for this purpose. In this way patient will be able to upload their data into our server and our machine learning software will let them know about their disease through our website whether it is in a good or bad condition.

REFERENCE

(Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly. 2015. Diagnosis Of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1.).

(Galega officinalis. May 2009. Management Of Type 2 Diabetes Mellitus 4TH Edition.)

V. H. Bhat, P. G. Rao, and P. D. Shenoy, “An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques,” Architecture, Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009..

2. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, “Application of data mining: Diabetes health care in young and old patients”, Journal of King Saud University – Computer and Information Sciences, vol. 25, pp. 127–136, 2012.

K. Rajesh, V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis” in International Journal of Engineering and Innovative Technology (IJEIT) Vol 2(3), 2012.

Sadhana, Savitha Shetty, “Analysis of Diabetic Data Set Using Hive and R”, International Journal of Emerging Technology and Advanced Engineering, vol 4(7), 2014.

Sabibullah M, Shanmugasundaram V, Raja Priya K, “Diabetes Patient’s Risk through Soft Computing Model”, International Journal of Emerging Trends & Technology in Computer Science, vol 2(6), 2013.

Kocur, I., Resnikoff, S.. Visual impairment and blindness in europe and their prevention. Brit J Ophthalmol 2002;86(7):716–722

Evans, J., Rooney, C., Ashwood, F., Dattani, N., Wormald, R.. Blindness and partial sight in England and Wales: April 1990-march 1991. Health Trends 1996;28(1):5–12

Philip, S., Fleming, A.D., Goatman, K.A., Fonseca, S., Mcnamee, P., Scotland, G.S., et al. The efficacy of automated disease/no disease grading for diabetic retinopathy in a systematic screening programme. Brit J Ophthalmol 2007;91(11):1512–1517.

Gandhi M. and Dhanasekaran R. (2013). Diagnosis of Diabetic Retinopathy Using Morphological Process and SVM Classifier, IEEE International conference on Communication and Signal Processing, India pp: 873-877

Li T, Meindert N, Reinhardt JM, Garvin MK, Abramoff MD (2013) Classification with Application to Retinal Diabetic Detection in Fundus Images

Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care. 2012;35:556–64

BoserB ,Guyon I.G,Vapnik V., "A Training Algorithm for Diabetic Detction ", Proc. Fifth Ann. Workshop Computational Learning Theory,pp. 144-152, 1992