A Data Mining Inspired Methodology towards the Identification of Diabetic Retinopathy

Pradeep Kumar KG¹, Karunakara K², Thyagaraju GS³, Sunanda Dixit⁴
¹Dept. of Computer Science and Engineering, A. J. Institute of Engineering & Technology, Mangalore, India.
²Dept. of Information Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, India.
³Dept. of Computer Science and Engineering, SDM Institute of Technology, Ujjire, Belthangady, India.
⁴Dept. of Computer Science and Engineering, BMS Institute of Technology and Management, Bangalore, India.

¹putturpradeep@gmail.com, ²karunakarak@gmail.com, ³thyagaraju.gulasinda@gmail.com, ⁴sunandadixit_cse@bmsit.in

Abstract. The Biomedical image analysis technique used in most of the clinical diagnosis activities, which is one of the explorative areas that appeal intense significance among scientists. The retinal fundus images are utilized in clinical diagnosis extensively for the treatment and to observe various eye diseases. Diabetic retinopathy is one of the foremost sources for blindness. The major diagnostic sign of diabetic retinopathy is the damage of blood vessels due to various reasons in the eye and then establishment of lesions in the retina. The screening and detection of Diabetic Retinopathy can be performed using retinal fundus images. The identification and analysis of diabetic retinopathy (DR) by means of color fundus images involves experienced practitioners to recognize the existence of many small topographies with a detailed grading system, makes this a complex and time-consuming mission. In this paper, a novel systematized method for the discovery of exudates in retinal images to diagnose diabetic retinopathy. The color fundus images are characterized and analyzed to find microaneurysms on the retina and provides the severity. The algorithm is tested on datasets provided by ophthalmologists and Messidor dataset, which gave excellent and promising results.

Keywords: Diabetic Retinopathy, Biomedical Image, Microaneurysms, Messidor dataset, lesions.

1. Introduction
In general, the human vision is not just a combination of rods and cones in the retina. A massive amount of managing & processing takes place, few of it immediately in the retina and few in the optical cortex at the rear of the brain; formerly a picture is made obtainable to the cognizant mind[8][9]. The eye is able to apprehend images over a very varied series of illumination stages, concealing about 9 or 10 orders of magnitude from a few dozen photons. Adaptation to varying intensities of illumination consumes a little bit of time, up to several minutes subject to the volume of change. Light transmits across several tiers of managing & treating neurons to stretch the light-sensitive rods and cones. The light recognizing rods and cones exist at the back, and the light needs to pass over several tiers of handling & managing cells to reach them [12] [13]. Diabetic Retinopathy (DR) is a severe eye disorder that arises due to diabetes mellitus and it has become the utmost usual source of blindness in the current situation. The eyesight of the patient might be disturbed by diabetes, which instigates cataracts, glaucoma, and furthermore, harm to blood vessels inside the eye, results in diabetic retinopathy. Effectual treatments for DR are accessible and available although it needs timely diagnosis and the continuous supervision of patients. The Identification of DR is accomplished by the
assessment and evaluation of retinal-fundus images. It emerges when diabetes mellitus damages the minute blood vessels that are the internal part of the retina, the light-sensitive tissue at the rear of the eye[14][15]. Exudates segmentation and detection is proposed by Akshay et al[19]. Heart disease prediction using random forest is proposed in[20].

The initial signs of the presence of diabetic retinopathy are the appearance of microaneurysms; these emerge as tiny-sized red dots among the larger vessels of the retina. The occurrences of microaneurysms are due to high pressure in the nerves or blood vessels of the retina. In various cases, the microaneurysms will burst triggering hemorrhages. As the disease and damage to the vasculature progress larger hemorrhages will emerge. This results in leakage of blood; with this, the vessels will also leak lipids and proteins causing small bright dots called exudates. Also, this will affect the retina and leads to some features such as exudates, cotton wool spots or venous loops. The systematic inspection of diabetic patient’s retina, for the identification of diabetic retinopathy, can decrease the risk of sightlessness in the patients by at least 50%. For this purpose, there is a need for an automated or computerized tool for the analysis of diabetic patient’s retina, which will support ophthalmologists, to screen higher populations of patients. The major reasons behind the diabetic retinopathy are genetic heredity and lifestyle. There is a need for the development of novel approaches in the identification of diabetic retinopathy. There is always room for the development of digital image processing techniques for the identification of the disease. The retinal images are processed in various stages to get clear and accurate results in identification. Depends on the occurrences of microaneurysms and area of the retina which affected by this decides the severity of the disease. [10] [11] [18]

2. Literature Survey
According to Hsin-Yi Tsao [1] there are various DR factors that are risky in nature are required to be studied, but it is a matter of fact as to know which risk factors are related to DR. If we study those DR related risks more accurately then we will be able to prevent the diseases to the different population. Here in the paper the prediction model of DR is built that treats Type 2 Diabetes Mellitus using DM Approach. The paper uses SVM, ANN etc for the study.

This is the paper [2] where the author explains about the annual examinations for the patients suffering from the diabetes to find the DR. Which is nothing but a condition where patient suffers from diabetes also has eye problem. Here the paper uses the method which uses the clinical data from the CDC's NHNES where the ML approach is being used for publically available health data. The data uses the classifiers with the model achieving the AUC of 0.72 and sensitivity of 69%.

Here in this paper [3] the DR there is a main problem of the blindness being occured for the important complication. In the paper the author says that the accurate analysis is required for the treating the disease. In the paper the study explains about the powerful algorithm which treats the DR. The study use the DIARET Database for the selecting the images for diagnosing the DR used. The Matlab software used in these runs in the Windows OS with the processor speed of 2.5 Giga Hertz and the RAM of 4 GB. The results on the proposed model shows that the results are based on the comparison which results in the ROC curve and estimates the model's AUC to show the optimal performance. The paper concludes that the algorithm used are on the basis of Image Processing System and suggests that eye specialists can treat the DR quickly.

The DR [4] uses the database which detects the diabetes in the early symptoms. These days the treatment of the DR requires the Fundus Images using OCT. The OCT instrument are expensive that are beneficial to both the patients and the eye specialists as these instruments provide the exact treatment to be made. In the mentioned paper the algorithm DCNN which is another Convolution NN algorithm. After combining the 2 portions we train the same using SVM which is used to categorize each into different categories. In this paper we perform the tests using the data that are provided by the Kaggle. Where the no of images are 34k trained images and are tested 53k times. The result of the test shows the accuracy is 86%. The Figure 1 shows the proposed framework used in the paper.
According the paper the author Manoj Vairalkar [5] explains that there is one serious problem in the Diabetic patients for DR, where the quantity of the diabetes patients is increasing exponentially. The result of the increase in the outcome is informative too. There are different strategies used for Data Mining that is used to calculate which is diabetic and which is non diabetic. The paper proposes the compelling algorithm to identify the DR. The Figure 2 shows the KDD architecture used in the paper.

The author [6] in this paper explains that there are huge advances that are being made in the field of health technology, biomedical and biotechnology. This has led to the significant rise in the data product such as genetically high throughput of the data that are acquired from the Electronic DB Health Informatics. The paper described about the ML and DM techniques of the bioscience that are
more vital and avails the data that are more vital. The DM which is the disorder that exerts the human health across the globe. The advanced research done in the aspect if treating the Diabetis has become huge advances in the research. The main aim of the study is to conduct a review in the application of ML, DM etc. The study shows 85% were supervised ML approach used and 15% were unsupervised approach. The Figure 3 and 4 explains the process used in the paper and KDD process respectively.

![Classification Process](image1)

**Figure 3.** Classification Process used in the Paper by Ioannis Kavakiotis et.al. [6]

![KDD Process](image2)

**Figure 4.** Basic Steps of KDD Process mentioned in the Paper by Ioannis Kavakiotis et.al. [6]
As per the paper the Diabetes is chronic disease caused by the problem of carbo metabolism and has manipulated world with serious health diseases. There is a technique used for the correct and quick detection of the diabetes that treats the patients without any side effects. The MI method is being in the process. The results show that the technique is highly accurate and successful in k-means algorithm [7].

![Figure 5. The proposed model by Shima Afzali [7]](image)

3. **Methodology**

The proposed framework follows different stages as shown in the Figure 6. The images produced by color fundus camera are acquired and trained for image processing. These images are from the patients for the inspection of diabetic retinopathy and its severity levels. The unaffected images were also considered while building the training data. The Image Enhancement involves grayscale conversion, adjusting brightness, and contrast for more clarity in the picture. Image restoration involves elimination or reduction of deprivations, which may be incorporated through the attainment of images. The use of median filter helps in removal of the noise by maintaining edges [16] [17].

![Figure 6. Various stages of proposed methodology](image)
3.1. Boundary Detection
The boundary of the retinal image is detected using canny edge detection technique. The canny edge detection is the ideal edge detector, it comprises a multi-phase process. The various stages are, noise lessening, gradient computation, non-maximum suppression, double threshold and edge tracing. This will highlight segments of the image with high first spatial derivatives. Edges escalate to creases in the gradient magnitude image appearances. In the first stage, the use of gaussian filtering technique helps in noise reduction. In second stage, we are finding out the intensity gradient. The smoothened image is then filtered with a Sobel kernel in both vertical and horizontal track to attain the first derivative in a horizontal direction ($G_x$) and vertical direction ($G_y$). From these, we can locate edge gradient and direction for every pixel by utilizing the subsequent formula.

\[
\text{Edge Gradient} (G) = \sqrt{G_x^2 + G_y^2}
\]

\[
\text{Angle} (\theta) = \tan^{-1} \left( \frac{G_y}{G_x} \right)
\]

For non-maximum suppression as the third stage, subsequent attainment of gradient magnitude and direction, a complete examination of image is made for the removal of any spurious pixels which may trouble during the formation of edge. The edge is specified via using double threshold values in the last stage. In general, an edge which is having an intensity gradient beyond the UpperLimit are certain to be edges, and those beneath LowerLimit are non-edges. This step also eliminates small pixel noises by edge tracking. Segmentation involves, partitioning the image into constituent parts. It intern helps to recognize the coverage of disease in the retina. The edge-based segmentation process is utilized to trace out till the edges. It is constructed on the swift transformation of intensity cost in an image because a particular intensity value does not deliver decent evidence about the edges. Hence, the edges or creases are distinguished and then are associated together to custom the object margins to fragment the mandatory regions. The Figure 7 and Figure 8 shows the sample images from MESSIDOR dataset which was collected from the FUNDUS camera.

![Figure 7. From Messidor Dataset-1](image-url)
3.2. Operational procedure of algorithm

The registration of the affected region takes place in an efficient manner. When a candidate is generated or extended, which matches the threshold, it is loaded to a separate data structure. So that frequent areas can be matched and mapped. The frequent itemset are generated until there is no match. The mapping will get over once it reaches the boundary. The simple median filter reduces noise by maintaining sharp and continuous boundaries in signal values. It is efficient in lessening hasty noise. The median filter gives the output, as the median value of the input data. In the first iteration, each segment of the image is treated as a separate set and maintained in a separate data structure. The value from each segment is compared to each other. If there’s a similarity then copy it into the second data structure highlighting disease if there is a threshold match depicting the disease. In the second iteration, the neighbouring two segments are grouped together and compared with the next group. If there’s a similarity update in second data structure for the affected region. In the third iteration, three segments are grouped and compared with other regions. This process iteration of comparison is continued until the entire image is treated as one single segment. During this process, unmatched regions are eliminated in each iteration, as unmatched regions are not affected. In this approach, the repeated subcategories are stretched one entry at an interval, a phase recognized as candidate generation, and clusters of candidates are experienced against the threshold data. Non-frequent subsets are eliminated using the process candidate elimination, which depicts a healthy region of the eye. Will when all the subsets are tested for frequent itemset. The content of the data structure is the actual diabetic retinopathy affected region. The regions will be highlighted at the end showing the affected area.

Object recognition is the process of highlighting the affected area. Here the recognized area is compared with the other parts of the image and the final result is obtained. In the last stage, total coverage of diabetic retinopathy and its position in the retina is obtained. The severity can also be diagnosed depends the result. During the experiment, both the local datasets, as well as Messidor datasets were used. All the datasets are annotated and verified by the ophthalmologists. The local dataset containing 110 affected images and 45 non-affected images. The algorithm worked perfectly.
for this dataset and gave 98.1% accuracy. [21] For the Messidor dataset containing 487 images, the experiment showed 92.8% accuracy. [20] proposed 2d to 3D X-ray image reconstruction.

4. Results
The study carried out using the algorithm, able to find the status of the normal image of an eye and the DR image taken from the fundus camera. Using the validation technique, we found out the validation results from the sample of images taken. We used cross-validation technique as shown in the Table 1:

| No | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Accuracy (%) | 88.33 | 90.00 | 90.33 | 90.83 | 91.33 | 91.67 | 92.67 | 92.5  | 93.17 | 93.67 |

With Apriori Knowledge the accuracy of tests carried out randomly indicates the results in a better way. With the more experiments being carried out by increasing dataset inputs as batches. When the batch with 20 datasets, the accuracy also increased as shown in the Table 2:

| No | 11    | 12    | 13    | 14    | 15    | 16    | 17    | 18    | 19    | 20    |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Mean Accuracy (%) | 88.85 | 90.77 | 92.31 | 93.85 | 93.46 | 94.23 | 94.62 | 96.67 | 97.17 | 98.1  |

5. Conclusion
As discussed in the paper, the novel approach using apriori technique showed encouraging results. The datasets are trained and the results were verified by the ophthalmologists. Here we use the extracted input and compare the same with the traditional extraction technique and with the support algorithm, we could find that the nature of fundus images collected bettered. The technique is the most promising method for the detection of the DR at the early stage.

References
[1] Hsin-Yi Tsao, Pei-Ying Chan and Emily Chia-Yu Su, "Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms", BMC Bioinformatics, vol. 19 (S9), pp. 195-205, DOI: 10.1186/s12859-018-2277-0, August 2018.
[2] Omolola Ogunyemi and Dulcie Kermah, "Machine Learning Approaches for Detecting Diabetic Retinopathy from Clinical and Public Health Records", AMIA Annual Symposium Proceedings Archive, pp. 983-990, Nov. 5, 2015.
[3] Masoud Khazaee Fadafen, Nasser Mehrshad and Seyyed Mohammad Razavi, "Detection of diabetic retinopathy using computational model of human visual system", Biomedical Research, An International Journal of Medical Sciences, vol. 29, Iss. 9, 2018.
[4] Yung-Hui Li, Nai-Ning Yeh, Shih-Jen Chen and Yu-Chien Chung, "Computer-Assisted Diagnosis for Diabetic Retinopathy Based on Fundus Images Using Deep Convolutional Neural Network", https://doi.org/10.1155/2019/6142839, Hindawi, 2019.
[5] Manoj Vairalkar, Vijaya Kamble and Kalpana Malpe, "Risk Assessment of Diabetic Retinopathy using Data Mining Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075, Volume-8 Issue-6S2, April 2019.
[6] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, pp. 104-116, vol. 15, 2017.
[7] Shima Afzali and Okatay Yildiz, "An Effective Sample Preparation Method for Diabetes Prediction", IAJIT, Feb 2016.
[8] Marcovecchio. M, Mohn. A, Chiarelli. F, "Type 2 diabetes mellitus in children and adolescents." J.Endocrinol. Invest. Vol 28, Issue 9, 2005, pp. 853-863. DOI= http://doi.acm.org/10.1007/bf03347581

[9] Reinehr.T, "Type 2 diabetes mellitus in children and adolescents." World. J. Diabetes, Vol. 4, Issue 6, 2013, pp. 270-281, DOI= http://dx.doi.org/10.4239/wjd.v4.i6.270

[10] L’Italien. GJ, Paul.S, Hendel. R.C, Leppo. J. A, Fleisher.L.A, Cohen. M. C. Eagle. KA, Giugliano. RP, "Neural network assessment of perioperative cardiac risk in vascular surgery patients." Med. Decis. Making, Vol. 18, Issue 1, 1998, pp. 70-75. DOI= http://doi.acm.org/10.1177/0272989X9801800114

[11] Ashizawa.K, MacMahon. H, Ishida. T, Nakamura. K, Vyborny. CJ, Katsuragawa. S, Doi. K, "Effect of an artificial neural network on radiologists’ performance in the differential diagnosis of interstitial lung disease using chest radiographs." AJR. Am. J. Roentgenol. Vol 172, Issue 5, 1999,1311-1315. DOI=http://dx.doi.org/10.2214/ajr.172.5.10227508

[12] Douglas.TH, Moul. JW, "Applications of neural networks in urologic oncology." Semin. Urol. Oncol. Vol. 16, Issue 1, 1998, pp. 35-39.

[13] Salvi. M, Dazzi. D, Pellistri.I, Neri. F, Wall. J.R, "Classification and prediction of the progression of thyroid-associated ophthalmopathy by an artificial neural network." Ophthalmol. Vol. 109, Issue. 9, 2002, pp. 1703-1708. DOI= http://doi.acm.org/10.1016/s0161-6420(02)01127-2.

[14] Lendaris.GG, Armin. Rest, Misley.TR, "Improving ANN generalization using a priori knowledge to pre-structure ANNs."International Conference on Neural Networks.1, 1997, pp.248-253.

[15] Bellamine. F, Almansoori. A, Elkamel. A, "Modeling of complex dynamic systems using differential neural networks with the incorporation of a priori knowledge." Appl. Math. Comput. Vol. 266(C), 2015, pp. 515-526.

[16] Jiang.K, Zhou.ZX, Geng.XY, Zhang.XF, Tang.LM, Wu.HQ, Dong.JC, "Isotropic undecimated wavelet transform fuzzy algorithm for retinal blood vessel segmentation." J. Med. Imaging. Health. Inform, Vol. 5, Issue 7, 2015, pp. 1524-1527.

[17] Wu.HQ, Geng.XY, Zhang.XF, Qiu.MY, Jiang.K, Tang.LM, Dong.JC, "A self-adaptive distance regularized level set evolution method for optic disk segmentation." Biomed.Mater. Eng., Vol. 24, Issue 6, 2014, pp. 3199-3206. DOI=http://dx.doi.org/10.3233/BME-141141

[18] Pradeep Kumar KG, Dr. Karunakara K, and Dr. Thyagaraju G. S., “An Approach to the Detection of Retinoblastoma based on Apriori Algorithm”, International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), June 17, Volume 5, Issue 6, ISSN: 2321-8169, PP: 733 – 738.

[19] AS Akshaya, Sunanda Dixit, Ngangom Priyobata Singh, Automatic Detection of Diabetic Retinopathy Using Two Phase Tophat Transformations—A Novel Approach, 2017, Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications,pp:561-569.

[20] Sunanda Dixit, Vishwas G Pai, Vivian Claud Rodrigues, Karan Agnani, SR Vaishna Priyan, 3D Reconstruction of 2D X-Ray Images, 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS),pp.1-5.

[21] I Yekkala, S Dixit, Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection, 2018, International Journal of Big Data and Analytics in Healthcare (IJBDAH), Volume 3, Issue 1, pp 1-12.