BIUST Research and Innovation Symposium 2019 (RDAIS 2019)
Botswana International University of Science and Technology
Palapye, Botswana, 4 - 7 June 2019

ISSN: 2521-2292

Classification of Patients Data to Predict Signs of Diabetic Retinopathy

[1]Utsav Gupta
B.Tech.(Information Technology) Student,
Department of Information Technology,
School of Information Technology and Engineering,
Vellore Institute of Technology, Vellore, India
utsav.gupta2016@vitstudent.ac.in

[2*]B.Valarmathi
Associate Professor,
Department of Software and Systems Engineering,
School of Information Technology and Engineering,
Vellore Institute of Technology,
valargovindan@gmail.com

[3]K.Santhi
Associate Professor, Department of Analytics,
School of Computer Science & Engineering
Vellore Institute of Technology, Vellore, India
santhikrishnan@gmail.com

[4]Abid Yahya
Professor,
Electrical, Computer & Telecommunication
Engineering Department,
College of Engineering & Technology
Botswana International University of Science &
Technology, Palapye, Botswana
yahyabid@gmail.com

Abstract:

In healthcare, data mining is fetching increasingly common, if not gradually essential. Data mining submissions can greatly profit all parties involved in the healthcare commerce. For instance, data mining can help healthcare guarantors notice fraud and exploitation, healthcare officialdoms make purchaser connection administration decisions, doctors recognise effective conducts and best performs, and patients take better and more affordable healthcare amenities. So using the data mining tools only we are here going to predict the efficient algorithms for detecting the certainty of Diabetic Retinopathy on the basis of various vector form of the data from patients images. Diabetic Retinopathy is one of the major causes of blindness in the people at younger age. The support vector machine algorithm (SVM) algorithm is the most efficient classifier for dataset contains features extracted from the Messidor image set, since it produces the highest accuracy value of 71.38% than Decision Tree and Random Forest algorithms.

Keywords: Data Mining, healthcare, Diabetic Retinopathy, blindness, Random Forest, Decision Tree, SVM.

I. INTRODUCTION

Usage of data mining in the field of medical science has been increased significantly on data in the form of required medical data type, to identify the hidden patterns and classify data further. It is difficult for all medical practitioners to analyse the medical data and deduce the diseases since it is a complex task and requires maximum experience and expertise.

Data mining concepts, commonly characterized by volume, variety, velocity, and veracity includes data analysis, such as hypothesis-generation, apart from hypothesis-testing. It focus on data association stability, not on causal relationship, whereas probability distribution measures are essentially not required.

Medicinal facts as factual to be examined has numerous structures not just different from large data of other rectifications, but also dissimilar from outdated proven epidemiology.

Classification, a data mining system, is widely used in healthcare area and it advances the excellence of forecast, and analysis. The area of this project is to converse the part of Diabetic Retinopathy. In United States the carelessness about the health is the main reason of this disease affecting people of the age between 25-74. The careful machine by that polygenic disorder grounds retinopathy-vestiges blurred, however some ideas are hypothesised to elucidate the distinctive course and past of the malady.

The health problem will be analysed by fluoresceine X-ray photography and perception for the next features are given below:

Microaneurysms: The initial medical image of diabetic retinopathy; these happen secondary to vessel wall out pouching because of pericyte loss; they seem as little, red dots within the superficial tissue layerl layers Dot and blot haemorrhages: seem like smallaneurysms if they're minor; they occur as micro aneurysms separation within the deeper sheets of the retina, like the within atomic and outer layers.

Flame-shaped haemorrhages: Splinter haemorrhages that occur within the a lot of superficial fibre layer Retinal hydrops and arduous exudates: Caused by the breakdown of the blood-retina barrier, permitting outpouring of bodily fluid proteins, lipids, and supermolecule from the vessels.

Macular exudates: Leading cause of visual impairment in patients with diabetes.

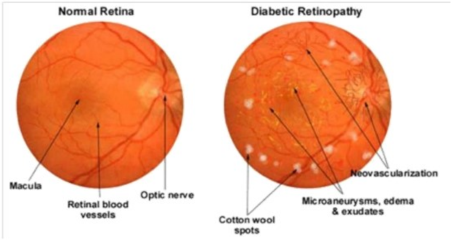


Fig.1 Eyeball Comparison

Fig.1 show the comparison of the eye of an normal individual with the eye of the diabetic retinopathy patient. The eyes of the retinopathy patient contains some cotton wool like structures in their eyes that is the main cause of defective sight.

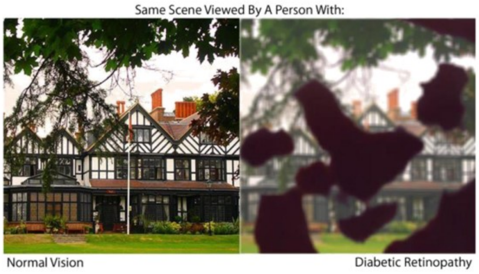


Fig.2 Vision Comparison

The contrast between the visions of the normal person is shown with that of the diabetic patient and it is shown in the Fig.2

So, if we can predict beforehand the possibilities of emergence of such a disease, it will prove to be extremely helpful to counter this disease, and can thus be prevented. Hence, we will try and analyse the patient retina as being diabetic or not diabetic taking into consideration the available features, by using certain techniques lke classification, clustering and etc.

Classification is a pre-specified set of classes**.** Clustering is unsupervised learning used to find groupings in the data using distance metrics. Regression quantifies the relationships between dependent and independent variables.

Thus, with the analysis results obtained by the mentioned techniques, we will try and attain the best possible outcome from the insights of this data.

II. LITERATURE SURVEY

Kotsiliti et al.[1] have given the model that consists of two phase of analysis, first of which is the Retinal fundus preprocessing and the second one is to classify data for detection of this chronic deisease DR.. First one is used to reduce the non-uniform illumination of pictures.

Verma et al.[2] has proposed the most appropriate formula to remove blood vessel which is considered to be the next version of harmonised mesh. He has also given an advanced approach of identifying the depletion.

Kamaladevi et al.[3] for the given journal staed that the new method of MA approach of detection based upon multifeature fused dictionary have been generated. Multifeature dictionary takes two things, first one is the semantic relationship between different features of the images and secondly t takes the content of the pictures.

Ananthpadmanabhan et al.[4] suggested that in the exixting approach, at primary stage, we are required to do the preprocessingfor removing uneven illumination, bad visual features such as brightness and contras filters.. Furthermore, MSCF is every day to recognize all thinkable MA contenders from the fundus similes. Formerly, MA image areas and non-MA image coverings can be mined from these entrants.

Faust et al.[5] stated that the classifications efficiency of different DR systems is discussed. Maximum of the testified structures are extremely augmented with esteem to the examined fundus pictures, so a simplification of distinct outcomes is hard.

Ramani et al.[6] has explained clearly about the changes and damages that can be cuased to eye of the patient suffering from diabetic retinopathy.

And furthermore he has also thrown light on how to detect Dr on the image by applying various image processing techniques such as pre-processing, acquisition, feature extraction.

And after the features are extracted we can apply different data mining techniques to identify patterns using classifications;

Fraser et al.[7]for the section-II of the paper he has given the brief discusiion on blood vessel algorithms, haemorrhage identification algorithm and the last one is SVM classification.

Roychowdhary et al.[8] has given the clear insight of the various distinguished classification algorithms of data mining techniques. Different classifiers of data-mining tools are like GMM, KNN, SVM and Adaboost. Can be used effectively to work it out for non-lesions. He has also identified that GMM and KNN are the best regarding this scope.

Akram et al.[9] suggested that if the eye-disorder patients are taking eye check-up on the regular basis, then there are chances that the DR can be eliminated or detctected frequently by 50%. In such circumstances, mechanical selection plans using fundus pictures prior to physical classifying can be tremendously cost-effective and useful.

Ghamdi et al.[10] has given the advice through his journal that the patients who are suffering from DR should have to go to ophthalmologist for their direct checkup using ophthalmoscope. It can be used to detect random blood glucose which can be further used to analyse the disease.

III. DATASET DESCRIPTION

Our dataset contains the information which is mined from the recognised Messidor dataset, with the last parameter stating the presence or absence of the disease.

TABLE 1: DATASET CHARACTERISTICS

Data Set Characteristics	Multivariate	**No. Of Instances**	1151	**Area**	Life
Attribute Characteristics	Integer, Real	**No. of Attributes**	20	**Date Donated**	2014-11-03
Associated Tasks	Classification	**Missing Values**	N/A	**No. Of Web Hits**	53271

Table 1 shows all the valid information of the dataset release and its further usage in the industry with over 53271 web hits.

There are total 20 attributes in the dataset which are as follows:

0) Quality assessment. 0 = bad quality 1 = sufficient quality.

1) Classification, where 1 specifies severe retinal anomaly and 0 its absence.

2-7) The outcomes of MA finding. Feature value is equals to number of Mas found.

8-15) comprise the similar evidence as 2-7 for exudates. However, as exudates are symbolised by a set of points rather than pixels creating the lesions, these structures are standardised by distributing the number of scratches with the length of the ROI to reimburse dissimilar image dimensions.

16) The Euclidean distance of the centre of the macula and the centre of the optic disc to provide important information regarding the patient's condition.

17) The diameter of the optic disc.

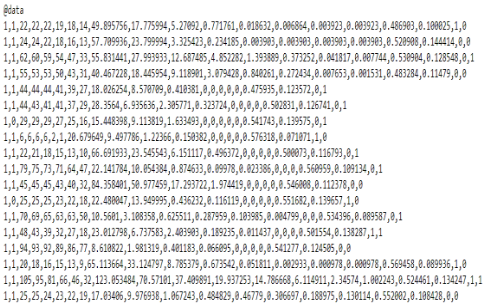
18) The binary result of the AM/FM-based classification.

19) Class label. 1 = marks of DR (Accumulative label for the Messidor classes 1, 2, 3), 0 = no signs of DR.

The sample dataset is shown in Fig. 3.

IV. EXISTING SYSTEM

Diabetes Type-I is a enduring disease that leads to macro and micro vascular complications, such as retinopathy, nephropathy and neuropathy. This disease is the main reason of blindness in the world of economics. Collective occurrence of Type-I Diabetes and elderly people age is the reason behind occurance of diabetic related diseases. Poor glycaemic control and diabetic-tenure are thw risk factors contributing to this retinopathy problem. The Diabetes Control and Complications Trial/Epidemiology of Diabetes Intervention and Complications (DCCT/EDIC) have confirmed that exhaustive conduct of hyperglycaemia efficiently suspensions the onset and slows the progression of complications in T1D, including DR. Nephropathy, hypertension and dyslipidemia are the other danger factors included in this disease.



```
@data
1,1,22,22,22,19,18,14,49.895756,17.775994,5.27092,0.771761,0.018632,0.006864,0.003923,0.003923,0.486903,0.100025,1,0
1,1,24,24,22,18,16,13,57.709936,23.799994,3.325423,0.234185,0.003903,0.003903,0.003903,0.003903,0.520908,0.144414,0,0
1,1,62,60,59,54,47,33,55.831441,27.993933,12.687485,4.852282,1.393889,0.373252,0.041817,0.007744,0.530904,0.128548,0,1
1,1,55,53,53,50,43,31,40.467228,18.445954,9.118901,3.079428,0.840261,0.272434,0.007653,0.001531,0.483284,0.11479,0,0
1,1,44,44,44,41,39,27,18.026254,8.570709,0.410381,0,0,0,0,0,0.475935,0.123572,0,1
1,1,44,43,41,41,37,29,28.3564,6.935636,2.305771,0.323724,0,0,0,0,0.502831,0.126741,0,1
1,0,29,29,29,27,25,16,15.448398,9.113819,1.633493,0,0,0,0,0,0.541743,0.139575,0,1
1,1,6,6,6,6,2,1,20.679649,9.497786,1.22366,0.150382,0,0,0,0,0.576318,0.071071,1,0
1,1,22,21,18,15,13,10,66.691933,23.545543,6.151117,0.496372,0,0,0,0,0.500073,0.116793,0,1
1,1,79,75,73,71,64,47,22.141784,10.054384,0.874633,0.09978,0.023386,0,0,0,0.560959,0.109134,0,1
1,1,45,45,45,43,40,32,84.358401,50.977459,17.293722,1.974419,0,0,0,0,0.546008,0.112378,0,0
1,0,25,25,25,23,22,18,22.480047,13.949995,0.436232,0.116119,0,0,0,0,0.551682,0.139657,1,0
1,1,70,69,65,63,63,50,10.5601,3.108358,0.625511,0.287959,0.103985,0.004799,0,0,0.534396,0.089587,0,1
1,1,48,43,39,32,27,18,23.012798,6.737583,2.403903,0.189235,0.011437,0,0,0,0.501554,0.138287,1,1
1,1,94,93,92,89,86,77,8.610822,1.981319,0.401183,0.066095,0,0,0,0,0.541277,0.124505,0,0
1,1,20,18,16,15,13,9,65.113664,33.124797,8.785379,0.673542,0.051811,0.002933,0.000978,0.000978,0.569458,0.089936,1,0
1,1,105,95,81,66,46,32,123.053484,70.57101,37.409891,19.937253,14.786668,6.114911,2.34574,1.002243,0.524461,0.134247,1,1
1,1,25,25,24,23,22,19,17.03406,9.976938,1.067243,0.484829,0.46779,0.306697,0.188975,0.130114,0.552002,0.108428,0,0
```

Fig.3 Dataset Snapshot

V. GAP IDENTIFIED

From the existing system, we came to know that Vision Threatening DR is the most critical consequence of Type-I Diabetes. And the above system do not come up with exact solution to detect whether the patient will going to have the Diabetic Retinopathy or not. So to resolve that issue came up with the different data mining techniques to predict the same. Then we will observe which technique will give the most optimum result. So that it can be used to predict the DR in patients having Diabetes Type-I.

VI. PROPOSED METHOD

The motivation for this project came from the fact that the highlighted disease is an important disease to address while considering counter measures for any situation. This disease is growing rapidly amongst our society and needs to be immediately countered. This is only possible when sufficient analysis is done on the data that is available from the sources. Thus, we took this as the topic of our project.

Further, even on obtaining the data simply doesn't bring about a change on its own. We must then commence the process of analysis. To bring about the correct results which are apt and reveal information on the concept being tested, we must prepare the model very carefully and precisely. So, we took this challenge upon ourselves so as to learn through this process the depths of knowledge within this field.

In this project we are comparing the accuracy of classification by three different models.

For the analysis to be performed, we are choosing a three classification techniques and they are given below:

a) Decision Tree Algorithm
b) Random Forest Algorithm
c) Support Vector Machine Algorithm

Using these three tree-based classifier methods, we will try and find out which method is best suited for in this scenario, for this dataset. This can be found out by using certain comparison parameters like the accuracy of the prediction of the model.

VII. RESULTS AND DISCUSSIONS

a) Simple Decision Tree

Also, from the below graph, we can clearly see that as the critical performance (cp) value decreases, the size of tree increases, and the relative error in the X-val decreases, thus improving the performance of the Decision Tree classifier. In this classifier Gini Index is used and it gives the accuracy of 59.83%

TABLE 2. DECISION TREE (GINI) CLASSIFICATION REPORT

Average statics	precision	recall	f1-score	support
0.0	0.54	0.77	0.63	156
1.0	0.71	0.46	0.56	190
avg(micro)	060	0.60	0.60	346
avg(macro)	0.62	0.61	0.59	346
avg(weighted)	0.63	0.60	0.59	346

Table 2 shows the performance and accuracy measure of the Decision Tree Algorithm using Gini Index.

TABLE 3. DECISION TREE (ENTROPY) CLASSIFICATION REPORT

Average statics	precision	recall	f1-score	support
0.0	0.55	0.76	0.64	156
1.0	0.71	0.49	0.58	190
avg(micro)	0.61	0.61	0.61	346
avg(macro)	0.63	0.62	0.61	346
avg(weighted)	0.64	0.61	0.60	346

Table 3 shows the performance and accuracy measure of the Decision Tree Algorithm using Entropy and it gives the accuracy of 60.98%

.

b) Random Forest

Table 4 shows the performance and accuracy measure of the Random Forest Algorithm and it gives the accuracy of 67.91%

TABLE 4. RANDOM FOREST CLASSIFICATION REPORT

Average statics	precision	recall	f1-score	support
0.0	0.63	0.71	0.66	156
1.0	0.73	0.66	0.69	190
avg(micro)	0.68	0.68	0.68	346
avg(macro)	0.68	0.68	0.68	346
avg(weighted)	0.68	0.68	0.68	346

c) Support Vector Machine

Table 5 shows the performance and accuracy measure of the Support Vector Machine Algorithm and Table 4 shows the performance and accuracy measure of the Random Forest Algorithm and it gives the accuracy of 71.39%

TABLE 5.SUPPORT VECTOR MACHINE CLASSIFICATION REPORT

Average statics	precision	recall	f1-score	support
0.0	0.63	0.88	0.73	156
1.0	0.85	0.58	0.69	190
avg(micro)	0.71	0.71	0.71	346
avg(macro)	0.74	0.73	0.71	346
avg(weighted)	0.75	0.71	0.71	346

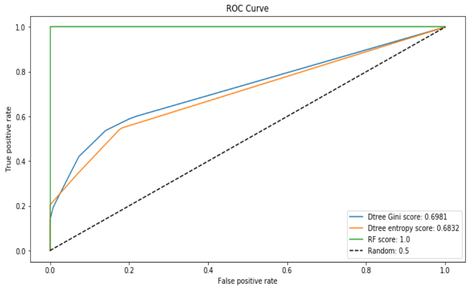


Fig.4 Receiver Operating Characteristic (ROC) curve

As per Fig.4, We can see that the area under the curve decides the efficiency of the algorithm in Receiver Operating Characteristic (ROC) curve. Hence we can see that the area under the SVM curve is maximum with the accuracy of 71.38%. Hence through this we can ensure that the SVM is the most efficient algorithm that can be used to predict the Diabetic Retinopathy in through this dataset.

As we have already discussed about our Messidor dataset which consists of 19 features and two classes where 1 represents the presence of retinal diabetic retinopathy and 0 represents its absence. As SVM uses only classification not regression it is the best algorithm that we can use for the dataset. And it is the one among many of the classification algorithm that can easily take many features and work ont them smoothly to classify data. We have used Decision Tree, Random Forest and SVM algorithms for the dataset, For SVM we are getting 71.38% of accuracy, 59.83% of accuracy for Decision Tree and 67.91% of accuracy for Random Forest respectively. Even SVM can perform better if we increase the features considerably but the features should have to be sparse and classes should be two for the most effective results.

Hence, earlier in all our references we have noticed that SVM is not yet used for comparision and classification of the Messidor Diabetic Retinopathy data. So, by using SVM we have seen a great impact in the accuracy of the system.

VIII. CONCLUSION AND FUTURE WORK

Thus, we have found out that SVM was the most efficient classifier for this particular dataset taken, since it produced the highest accuracy value of 71.38%. This tells us that within the constraints of the given scenario and the techniques on hand, the most useful for analytics was proved to be the SVM classifier.

However, it must always be kept in mind that this conclusion can be only true for this particular scenario, and may vary for different cases. Thus, we must not conclusively establish this result as the benchmark result, for a much larger extent of analytics is required on a much larger scale to actually bring about certain defining characteristics, based on which we may then draw conclusions. In our future work, we are going to use the deep learning concepts for the larger dataset.

REFERENCES

[1] Kotsiliti, E., Al-Diri, B., & Hunter, A, “A classification model for predicting diabetic retinopathy based on patient characteristics and biochemical measures. *Journal for Modeling in Ophthalmology*, Vol. 1, pp. 69-85, 2017.

[2] Verma, K., Deep, P., & Ramakrishnan, A. G., “Detection and classification of diabetic retinopathy using retinal images”, In *India Conference (INDICON), 2016 Annual IEEE,* pp. 1-6, 2016.

[3] Kamaladevi, M., SnehaRupa, S., & Sowmya, T., “Automatic Detection of Diabetic Retinopathy in Large Scale Retinal images”, *International Journal of Pure and Applied Mathematics*, Vol. *119*, pp. 14181-14189, 2018..

[4] Ananthapadmanabhan, K. R., & Parthiban, G., “Prediction of chances-diabetic retinopathy using data mining classification techniques”, *Indian Journal of Science and Technology,* vol.7, pp. 1498-1503, 2014.

[5] Faust, O., Acharya, R., Ng, E. Y. K., Ng, K. H., & Suri, J. S., “Algorithms for the automated detection of diabetic retinopathy using digital fundus images: a review”, *Journal of medical systems*, vol. *36*, pp. 145-157, 2016..

[6] Ramani, R. G., Balasubramanian, L., & Jacob, S. G., “Automatic prediction of Diabetic Retinopathy and Glaucoma through retinal image analysis and data mining techniques”, In *Machine Vision and Image Processing (MVIP), International Conference* IEEE, pp. 149-152, 2012.

[7] Fraser, C., DAmico, D., & Trobe, J., “Diabetic retinopathy: Classification and clinical features”, *Netherlands: Wolters Kluwer*, 2015.

[8] Roychowdhury, S., Koozekanani, D. D., & Parhi, K. K., “DREAM: diabetic retinopathy analysis using machine learning”, *IEEE journal of biomedical and health informatics*, vol. 18, pp.1717-1728, 2014.

[9] Akram, M. U., Khalid, S., & Khan, S. A., “Identification and classification of microaneurysms for early detection of diabetic retinopathy”, *Pattern Recognition*, vol. 46, pp.107-116, 2013.

[10] Al Ghamdi, A. H., Rabiu, M., Hajar, S., Yorston, D., Kuper, H., & Polack, S., “Rapid assessment of avoidable blindness and diabetic retinopathy in Taif”, Saudi Arabia. *British Journal of Ophthalmology*, vol. 96, pp.1168-1172, 2012.