Diabetic Retinopathy Prediction Using Health Records and Machine Learning Classifiers

Pooja Rathi
Research Scholar, Rungta College of Engg. & Technology, Bhilai, India
Assistant Professor, St. Vincent Pallotti College, Raipur, India

Sourabh Rungta
Professor, Department of Computer Science Engineering,
Rungta College of Engg. & Technology, Bhilai, India

S. M. Ghosh
Professor, Department of Computer Science, Sir C.V. Raman University, Kota, Bilaspur, India

Abstract:

Diabetes is a condition that is spreading quickly over the globe. It happens when the body cannot utilise enough insulin or when the pancreas does not create enough of it. Diabetes Retinopathy (DR) and/or blindness is one of the main issues that diabetic individuals face. Diabetes patients are becoming more numerous, which also results in an increase of DR data. Therefore, the use of data mining (DM) techniques becomes important to extract the useful information and undiscovered knowledge from the data. Data Mining has become crucial in DR detection since it can contribute to the society's overall health.

The present study focuses on the quick, straightforward, and simple prediction of diabetic retinopathy using the physical records of the patients. Information is extracted from these records, which are in the form of numerical values, using DM techniques. Numerous classifiers have been used in the dataset for this DR prediction, including logistic regression, KNN, SVM, bagged trees, and boosted trees. Two distinct cross validations are applied to the data to find the best features and guard against overfitting. Dataset is consisted of the records of 900 diabetes patients. This novel work confirms that bagged trees and KNN are effective classifiers for DR prediction utilising physical health records. According to the results of the experiment, boosted trees with a 10% hold-out validation have the best classification accuracy at 90.1%. KNN has also resulted in 88.9% accuracy, which is noteworthy.

Keywords: Diabetic Retinopathy, Data Mining, Logistic regression, SVM, KNN, Boosted tree, Bagged tree, Cross Validation

DOI Number: 10.14704/nq.2022.20.9.NQ44803 **Neuro Quantology 2022; 20(9):6884-6857**

1. Introduction

Diabetes is a very widely spread illness everywhere in the world. This disease falls in the category of continual and persistent disease. It arises when human body stops producing sufficient insulin. Since insulin is one type of hormone and its function is to help the cells inside our body to use glucose in food. Because of the insufficient insulin production, the level of glucose increases in the body. This leads a person to diabetes.

The disease Diabetes is has two categories that are Type 1 and Type 2. Various complications are there which arise due to the high level of blood sugar like, cardiovascular diseases, problems related to nervous system, kidney problems, increase in the blood pressure, vision problem, etc.

2. Diabetic Retinopathy

2.1 Introduction of Diabetic Retinopathy:

Diabetic retinopathy is damage to the retina (retinopathy), specifically blood vessels in the retina, caused by complications of diabetes mellitus [1]. It takes place when high levels of blood sugar damages the blood vessels in the retina [2]. These blood vessels can swell or leak or it may stop blood circulation in that part. At the initial level DR does not reveal any major indications in the patient except, in some cases, mild vision variation. Eventually, it can cause blindness. Approximately 80% of all patients who have had diabetes for at least ten years suffer from some degree of diabetic retinopathy [1]. This state can arise in any person who has type 1 or type 2 diabetes. The longer the duration of diabetes and the less controlled the blood sugar is, the more likely are the chances to develop this eye complication. One might not have symptoms in the early stages of diabetic retinopathy [3]. Diabetic retinopathy symptoms may include:

- Seeing spots or floaters
- Blurred vision
- Fluctuating vision
- Impaired colour vision
- Dark or empty areas in vision
- Blindness

Following figure illustrates the state of diabetic retinopathy:

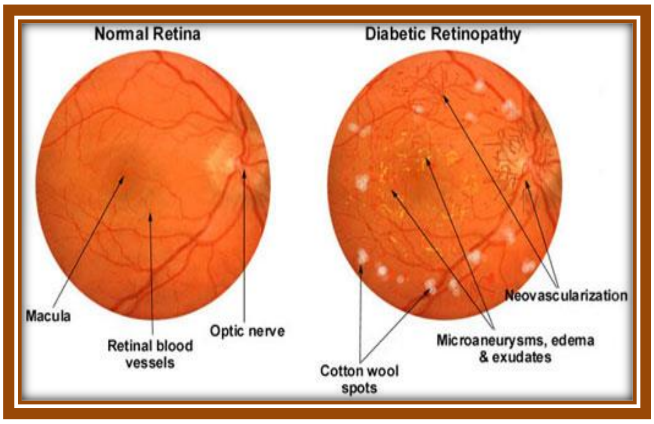

Figure 1: Diabetic Retinopathy

Through a complete eye examination DR can be diagnosed. Diagnosis may include:

- Patient history to determine vision problems, presence of diabetes, and other general health concerns that may be affecting vision
- Retinal detachment
- Visual insight measurements to determine the affected area
- Abnormalities in optic nerve
- Measurement of the pressure within the eye
- Refraction to determine if a new eyeglass prescription is needed

2.2 Types of Diabetic Retinopathy:

Diabetic Retinopathy is broadly classified into two categories: Non-proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR).

A statistical survey found that NPDR is present in 25% of patients 5 years after they were diagnosed with diabetes, 60% at 10 years, and 80% at 15 years [4][5]. These studies also found that the incidence of proliferative diabetic retinopathy (PDR) varied from 2% in those who had diabetes for less than 5 years to 15.5% in those who had diabetes for 15 or more years [6]. NPDR is also known as the early stage where walls of blood vessels of the retina become weaker. At this level, new capillaries are not grown.

The blood leaks from the blood vessels as the microaneurysms(MA) bulge out from walls of the blood vessels in the retina. Sometimes, in this state the central part of the retina known as macula brings to swell; This condition must be treated at the earliest. The weakened capillary walls can leak or rupture, causing haemorrhages [5][7]. Sometimes, small particles form in the retinas which are called exudates. NPDR can switch from mild to severe, as more siege capillary.

PDR is the higher and advanced stage of diabetic retinopathy. In the rear side of the retina, new blood vessels begin to grow. When there is lack of oxygen supply reaching to eye, retinopathy develops in the form of microvascular complication of diabetes, a disease of small vessels.

2.3 DR stages:

Table 1: Diabetic Retinopathy Stages

SNo	Stage Name	Type	Image	Percent age
1	No DR	Patient not suffering from Diabetic Retinopathy		82.6%
2	Miild NPDR	Patients have at least one Microaneurysm (small red swelling in the walls of retina) (8)		51.7%
3	Moderate NPDR	Blocking of capillary starts. Patients have haemorrhages or MAs in one to three retinal quadrants and/or cotton wool spots or hard exudates (8)		35.6%
4	Severe NPDR	Due to more blockages of capillary, retinal areas does not receive adequate blood flow. Owing to which, the retina can't grow new blood vessels to replace the damaged ones. (8)		4.6%
5	PDR	Additional new blood vessels begin to grow in retina, but they are fragile and abnormal. Because of this, they can leak blood leading to vision loss.(8)		8%

3. Role of Data Mining

Data mining is the way toward separating and analysing hidden styles of data in keeping with specific views for arrangement into beneficial and important information [9]. It is an analytical technique intended to discover data looking for steady patterns and discover orderly connections between variables. Data mining is widely applicable in different areas including banking, sports, health care framework, education, training machines, record keeping etc.

In recent years medical data mining has become prominent, as there is an enormous amount of medical data available. This medical data can be used for discovering useful patterns. Huge amounts of medical and scientific information of the patients have been accrued in the form of electronic records. With the help of these databases, new possibilities for medical data mining have been opened. The data mining techniques such as classification, clustering, association, spatio-temporal mining help in finding useful patterns from the huge amount of medical data. With the help of data mining techniques, these clinical data can be utilised to discover improved healthcare at low price. In this research, data mining techniques are applied on health records for the easy and quick prediction of diabetic retinopathy [10].

4. Classification Techniques

Classification is one of the fundamental tasks in Data mining. Classification is a predictive approach and is used to predict the group membership of a data instance. Classification is applied in areas such as weather prediction, medical diagnosing, scientific experiments etc.

Classification techniques are generally used in medical data mining. Various classification techniques are available in data mining like

Decision trees, Bayesian classifier, Random Forest, Random tree, Neural networks, Support Vector Machines, etc.

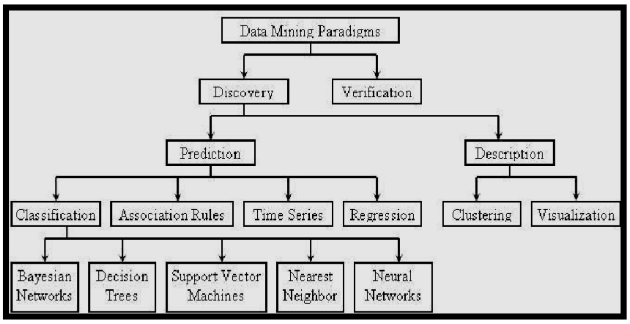

Figure 2: Data Mining Paradigms

Classification is performed in two steps:

Model construction: In this step the prediction model is built using appropriate algorithm.

Model Usage: In this step the prediction model is applied to actual data and prediction is done accordingly.

5. Literature Review

5.1 Neural Network for diagnosis of DR

Usher et al.[2] have proposed an adaptive intensity thresholding method for the extraction of exudates. In their research, ANN has been applied on the features extracted from dataset. Exudates were found by employing their gray-level variation. By applying the method of morphological filtering, the retinal optic disk detection was done. The sensitivity and productivity achieved were 92.8% and 92.4% respectively

R. Priya and P. Aruna[11] have presented a study on SVM and Neural Network based Diagnosis of Diabetic Retinopathy. Automated diagnosis system of diabetic retinopathy has been performed by using a set of 250 images which is a combination of normal, NPDR and PDR affected images. In this study, for the detection of DR, Neural Network and SVM classifiers were used. It was found that SVM performance is better than PNN from obtained results.

PonniBala et al.[3] proposed an automated DR detection system dispelling mathematical morphology. Authors have used the connected component analysis method for the feature extraction processfor. For the classification of the retinal images Fuzzy Inference System was used. Accuracy calculated was 93.84%.

Gabor mark somfai and et.al.[12] have used Artificial neural networks (ANNs) for the detection and classification of diabetic retinopathy related diseases. Here authors have assessed the competency of different features like dimensions, reflectance etc of retinal fundus images to train a Bayesian ANN to discriminate between healthy and diabetic eyes with and with no mild retinopathy.

5.2. Decision Trees and Naive Bayes for diagnosis of DR

Ramon Casanova and et al. [13] have worked on “Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses”. Authors have worked on Random Forests algorithms so that the fundus images data can be classified and analysed for the detection of diabetic retinopathy. It is found that random forest algorithms are more suitable than logistic regression analysis. Furthermore, organized data approach is followed and evaluation of diabetic retinopathy is performed with the help of the metrics. These metrics are receptive to patterns in the data associated with future DR events in subjects not presently affected by DR. Authors concluded that classifiers based on random forests have resulted with more precise results than the traditional logistic regression methods. As compared to logistic regression, Random Forests have performed well with higher accuracy.

K. R. Ananthapadmanaban and G. Parthiban [14] have proposed a system using Naive Bayes and Support Vector Machine algorithms to predict early detection of eye disease and DR. In their study, they have used Rapid Miner tool. The resultant accuracy derived by Naive Bayes is 83.37% and by SVM is 64.91%.

Karan Bhatia, Shikhar Arora, Ravi Tomar [15] in their study “Diagnosis of Diabetic Retinopathy Using Machine Learning Classification Algorithm” have proposed methods to develop an automated system to detect diabetic retinopathy among the diabetic patients. The main aim of the study was to help ophthalmologists in the detection of early symptoms of diabetic retinopathy. In this research, authors have used various technologies for disease detection. Prediction of DR was performed using various classifiers like SVM, RF, adaBoost, Decision Tree and Naive Bayes.

.

5.3. Support Vector Machine

C. Arvind, M. Ponnibala and S. Vijayachitra [16] in their study on "Automatic Detection of Microaneurysms and Classification of Diabetic Retinopathy Images using SVM Technique" have proposed a new method for detecting the microaneurysms from the colour fundus retinal image based on feature classification, to reduce the ophthalmologists' time and energy for verifying the retinal images. Using the colour fundus images microaneurysms are identified and have applied some techniques so that identical blood vessels can be removed. Images were pre-processed for the use of feature extraction. In this study SVM has given the accuracy of 90%

Mahendran Gandhi et al. [17] have performed the diagnosis of DR using morphological process and SVM classifier. In this study, for the automated diagnosis of diabetic retinopathy, a set of 5 images captures by retinal fundus camera were used. For the detection of exudates, the authors have applied erosion and dilation operations on the images. It was followed by segmentation operation. To classify the images into mild, moderate and severe, SVM classifier was applied.

Table 2: Algorithm Performance

Algorithm	Accuracy
Neural Network	93.84%
Naive Bayes	89%
SVM	90%
Decision Tree	72.5%
Random Forest	65%

Table 2 shows the performance accuracy of different classification algorithms. Following chart depicts the pictorial representation of the above given accuracies. From this chart it can be obviously noticed that the Neural Network is providing the best accuracy in the prediction of diabetic retinopathy.

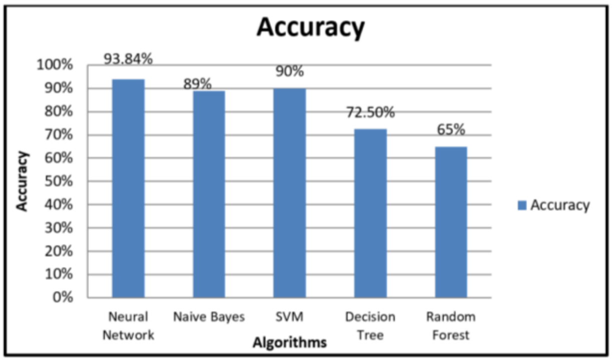

Figure 3: Pictorial representation of algorithm accuracies

6. Basic Structural Design of the System

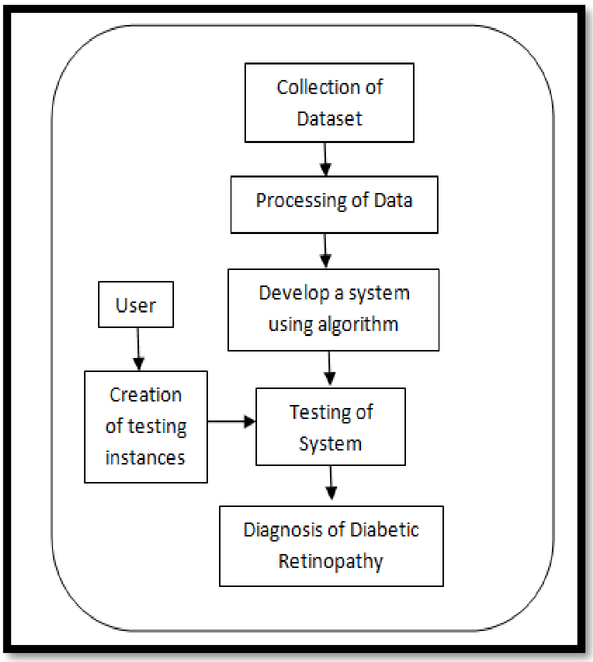

Figure 4: System Architecture

1. Collection of Dataset:

The first step will be collecting data which may contain different attributes like blood group, age, obesity, genetic history, blood pressure, smoking habits, etc. This data will be the helpful for creating a prediction system.

2. Processing of Data:

The dataset obtained may contain some noisy or null values. It may contain some incomplete information. Hence the processing of original data is essential before it can be used as input for the system. Here incomplete information or null values can be corrected or removed.

3. Developing a system:

In this phase different functions are to be performed:

a) Analysis of the dataset.
b) Selection of the activation functions.
c) Applying the algorithm such as Support Vector Machine, Bayesian algorithms, feed forward neural network, naive bayes, decision tree, random forest, etc…
d) Error detection and correction
e) Implement the system.

4. Testing System:

Testing system means to check according to symptoms whether the developed system is diagnosing the disease properly or not. When a user gives query, a testing instance is created. And that test instances gives to testing module. Taking into consideration all the symptoms, the testing routine will diagnose the diabetic retinopathy.

7. Materials and Methods

In this study, for DR prediction, data of a bunch of recurring outpatients has been considered. The information of various patients was collected and hand-picked from some private hospitals. The dataset comprises 900 records. In this research 10 different physical features have been recognized. These features are: Age of the patient, Gender of the patient, Term of the diabetes mellitus, BMI (Body Mass Index), Systolic blood Pressure Measurements, Diastolic blood Pressure Measurements, HbA1C measure, Smoking habit, Level of lipoprotein (LDL), Level of Microalbuminuria..

Patients of all age groups and both genders have been taken into consideration. Duration of diabetes is the key feature in this study. HbA1C level is classified into two terms: less than 8 and more than 8. BMI is a binary measure which takes value 1 if it exceeds 30 otherwise it takes 0. Measure of LDL is divided into three categories; 0, 1 and 2; for values – below 84 mg/dl, 84-111 mg/dl and above 111 mg/dl respectively. Smoking propensity holds 1 for smokers and 0 for non-smokers. Smoking propensity holds 1 for smokers and 0 for non-smokers. The degree of systolic and diastolic pulse estimations assumes an indispensable part. Microalbuminuria is the level of albumin and its value less than 30 is considered to be normal. Value among 30-300 demonstrates kidney issue (nephropathy). Each of the above features was chosen by specialists and assessment was made dependent on these attributes.

8. Experimental Setup

In this work, focus is on prediction of diabetic retinopathy. For this prediction of disease, supervised learning models are best suitable, as the dataset contains labelled variables and values. Five distinct classification algorithms have been applied on the dataset. Before application of the algorithm, as depicted in the structural diagram, the dataset is split into two parts: a training set and a testing set. The training set is used to train the DR prediction model, while the testing set is used to determine the accuracy of the model. Additionally, validation on DR dataset has been implemented, as this progression has assisted with discovering high-quality parameters for this DR prediction version and is likewise beneficial to forestall overfitting. Two distinct procedures have been used to perform validation step on the dataset namely:

i. K-fold Cross Validation
ii. Hold-out Validation

Hold-out is the easy and straightforward category of cross validation. In this type, data is partitioned into two groups: training data and testing data and then the model is developed using these sets. Hold-out validation has been selected as this can facilitate with dividing the data into train and test sets. Training set will be utilized to train the model. Subsequently, a test dataset will be utilized to assess the performance of this model on the inconspicuous data. This experiment have operated on two different types of data splits: one in which 10% of data will be used as test set and remaining as training set and other where 20% data will be used as test set and remaining as training set.

In k-fold cross validation, data has been subjectively isolated into k sets, where k is any positive whole number. In these k groups, there are k-1 training sets and 1 group is a test set. The dataset that have collected for this study contains records of patients which are restricted to 900 records. To generate the more precise and defined outcome, k-fold cross validation has been employed. In this validation, the generated model will be trained for k times. Each time while training the model, it will result in a special fold from training data and afterwards it will be utilized for testing set. Finally, the average value of the outcome of all these sets is calculated to get the execution accuracy. In this manner, overfitting can be avoided that will help to get more accurate results in disease prediction.

9. Algorithm Accuracies and Result Analysis

For DR prediction, different classification algorithms have been applied. These are logistic regression, K nearest neighbour, support vector machine, bagged trees and boosted trees. To evaluate machine learning models on a limited data sample, two different types of validations are applied: k-fold cross validation and hold-out validation.

9.1 The accuracy table for 10 fold and 20 fold cross validation using above listed classification algorithms is given in Table 3.

Table 3 : Accuracy table using k-fold cross validation

Algorithm	Accuracy (in %)	
	10 fold CV	20 fold CV
Logistic Regression	82.4	82.1
SVM	82.7	82.7
KNN	85.6	85.8
Bagged Tree	78.9	79
Boosted Tree	86.1	86.6

From Table 3, it can be observed that the boosted tree has performed superior than all other classifiers. It has resulted in 86.1% accuracy with 10 fold cross validation and has given 86.6% accuracy with 20 fold cross validation. The reason behind getting best accuracy in boosted trees is that boosting is based on the principle of high bias and low variance concept and is build on weak learners. In turn, boosting reduces error by blending a learning algorithm in a chain. Similarly, in case of our DR detection boosting has achieved a strong learner from many weak learners. In this case K nearest neighbours algorithm has also turned to good accuracy with 85.8% in case of 20 fold cross validation. Bagged trees resulted in lowest accuracy which is 79%. A more clear picture of the above accuracies can be seen through below given graphical representation in the form of 3-D pyramids.

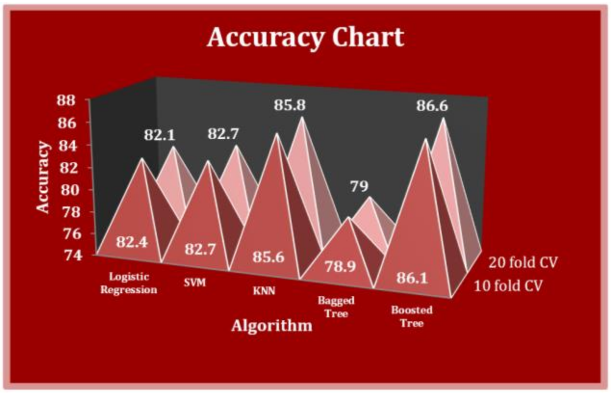

Figure 5: Classification Algorithm Accuracy Chart using k-fold Cross validation

9.2 The accuracy table for 10 fold and 20 fold cross validation using above listed classification algorithms is given below:

After analysing the accuracies after k-fold cross validation, now hold-out validation is employed on the same DR dataset. As per the discussion, in hold-out validation the original dataset is divided into training and testing set, in order to find out the best prediction model. Two most common partitions of the data are chosen where in the first move, 90% data was used as training set and remaining 10% as test set. In the second case, 80% data was utilised to train the model and 20% for testing. The table below depicts the accuracies of different classifiers:

Table 4 : Accuracy table using Hold-out validation

Algorithm	Accuracy (in %)	
	10% hold-out	20% hold-out
Logistic Regression	84.4	85
SVM	84.5	85
KNN	88.9	84.4
Bagged Tree	80	82.2
Boosted Tree	90.1	88.3

From the table above, it can be monitor that boosted trees, once again, have resulted as the best performer. This time algorithms have been applied with hold-out validation where we have used 10% and 20% held out. When applied with 10% hold out, boosted trees resulted in 90.1% accuracy and when combined with 20% hold out, it has given 88.3% accuracy. As mentioned above, boosting is an ensemble which is used to create a strong classifier model from several weak learners. Basically, cross validation allows us to rotate within training and testing in case of a comparatively small dataset. When boosting combined with hold out validation, offers the best result as it lowers the variance.

Along with the boosted trees, k nearest neighbour classifier has also furnished with 88.9% accuracy, which is quite good. Bagged trees have resulted in the lowest accuracy of 80% when used 105 hold out validation.

A comprehensible picture of the above accuracies can be seen through below given graphical representation in the form of 3-D pyramids.

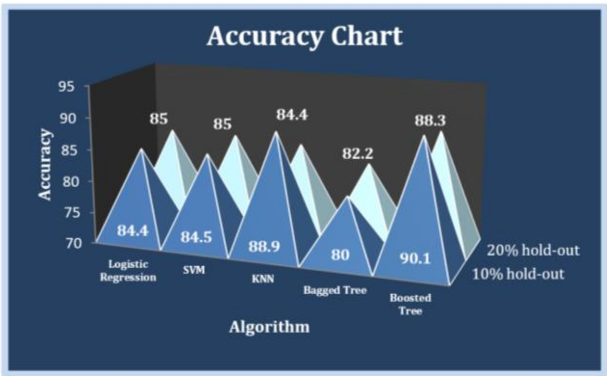

Figure 6: Classification Algorithm Accuracy Chart using Hold-out Validation

10. Conclusion

Diabetic retinopathy is a very dangerous complication causes to eyes of a diabetic patient. Its extreme level may damage both the eyes. Hence diagnosis and prediction of diabetic retinopathy is a real world problem. This study focuses on early prediction of diabetic retinopathy using physical health records of the diabetic patients. Records of 900 patients formed the dataset. Logistic regression, K nearest neighbour, Support Vector Machine, Bagged tree and Boosted trees are the classifiers that have been employed on the dataset for model building. K-fold cross validation and Hold-out validation are the techniques that have been applied along with the classification algorithms to determine the best features and to avoid overfitting. As the experimental outcome of this study, boosted tree have outperformed with the highest accuracy amongst all which comes to be 90.1% when applied with 10% hold-out validation. It produced 88.3% accuracy with 20% hold-out data. K-nearest neighbour classifier also resulted in remarkable outcome with 88.9% accuracy. From this study, it can be infer that Bagged tree and KNN are the optimum classifiers for the prediction of diabetic retinopathy using significant physical health records.

11. References

[1] Dr.V.Ramesh, R.Padmini,"Risk Level Prediction System of Diabetic Retinopathy Using Classification Algorithms", IJSDR, Volume 2, Issue 6, June 2017

[2] Usher D, Dumskyj M, Himaga M, Williamson TH, Nussey S, Boyce J, "Automated detection of diabetic retinopathy in digital retinal images: A tool for diabetic retinopathy screening", Diabetes Med21:84 – 90, 2004

[3] PonniBala M, Vijayachitra S, "Computerized retinal image analysis to detect and quantify exudates associated with diabetic retinopathy", International Journal of Computer Applications 54(2):0975– 8887, 2012.

[4] Klein R, Klein BE, Moss SE, et al. The Wisconsin Epidemiologic Study of Diabetic Retinopathy. II. Prevalence and risk of diabetic retinopathy when age at diagnosis

is less than 30 years. Arch Ophthalmol. 1984; 102(4):520-526.
[5] American Optometric Association. Evidence-based Clinical Practice Guideline: Eye Care of the Patient with Diabetes Mellitus. www.aoa.org/optometrists/tools-and-resources/evidence-based-optometry/evidence-based-clinical-practice-guidlines/cpg-3--eye-care-of-the-patient-with-diabetes-mellitus. Accessed May 2, 2019.
[6] Stratton IM, Kohner EM, Aldington SJ, et al. UKPDS 50: risk factors for incidence and progression of retinopathy in type II diabetes over 6 years from diagnosis. Diabetologia. 2001; 44(2):156-163.
[7] Karma Loday Bhutia, Neiwete Lomi, Sonam Choden Bhutia Prevalence of Diabetic Retinopathy in Type 2 Diabetic Patients Attending Tertiary Care Hospital In Sikkim.DJO 2017;28:19-21
[8] The Four Stages of Diabetic Retinopathy, Cecelia Koetting, OD, FAAO, June 2019
[9] Hayrettin Evirgen, Menduh Çerkezi, "Prediction and Diagnosis of Diabetic Retinopathy using Data Mining Technique", Turkish Online Journal of Science & Technology, Vol. 4 Issue 3, July 2014, pg.32-37
[10] Mrs. Pooja Rathi, Dr. Anurag Sharma, " A Review Paper on Prediction of Diabetic Retinopathy using Data Mining Techniques", IJRIT, Vol 4, Issue 1,292-297, June-2017
[11] R. Priya, P. Aruna, "SVM and Neural Network based Diagnosis of Diabetic Retinopathy", International Journal of Computer Applications (0975–8887)Volume 41–No.1, March 2012
[12] Gábor Márk Somfai, Erika Tátrai, Lenke Laurik, Boglárka Varga, Veronika Ölvedy, Hong Jiang, Jianhua Wang, William E Smiddy, Anikó Somogyi, and Delia Cabrera DeBuc, "Automated classifiers for early detection and diagnosis of retinopathy in diabetic eyes", BMS Bioinformatics, Volumn 5, 2014.
[13] Ramon Casanova, Santiago Saldana, Emily Y. Chew, Ronald P. Danis, Craig M. Greven, Walter T. Ambrosius, "Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses", Plos, June 2014
[14] K. R. Ananthapadmanaban and G. Parthiban. – "Prediction of Chances - Diabetic Retinopathy using Data Mining Classification Techniques.", Indian Journal of Science and Technology, Vol 7(10), 1498–1503, October 2014
[15] Karan Bhatia, Shikhar Arora, Ravi Tomar, "Diagnosis of Diabetic Retinopathy Using Machine Learning Classification Algorithm", 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016
[16] C. Arvind, M Ponnibala, S. Vijaychitra, "Automatic Detection of Microaneurysms and Classification of Diabetic Retinopathy Images using SVM Technique", International Journal of Computer Applications (0975–8887), ICIIIOSP-2013
[17] MahendranGandhi and Dr. R. Dhanasekaran – "Diagnosis of Diabetic Retinopathy Using Morphological Process and SVM Classifier", International conference on Communication and Signal Processing, April 3-5, 2013, India.
[18] Rathi Pooja, Shrivastava Padmavati & Ghosh, S. (2020). Prediction of Diabetic Retinopathy Using Classification Techniques. Solid State Technology. 63. 9479-9494.