IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 8, Issue 5, October - November, 2020
**ISSN: 2320 – 8791 (Impact Factor: 2.317)**
www.ijreat.org

# Diabetic Retinopathy detecting using Kernel PCA

[1]B Yugandhar, [2]RNV Jagan Mohan
[2]Research Scholar, CSE, GIET University, Gunupur,Odisha
[2]Assoc Prof, CSE, SRKR Engineering College, AP, India

***Abstract*-- Diabetic Retinopathy is the most regular motive of avoidable imaginative and prescient impairment, mainly affecting the operating-age populace inside the world. Recent studies has given a higher information of the requirement in clinical eye care exercise tofind better and cheaper ways of identification, management, diagnosis and remedy of retinal disease. The significance of diabetic retinopathy screening applications and issue in reaching reliable early analysis of diabetic retinopathy at an inexpensive value needs attention to develop computer-aided diagnosis tool. Computer-aided disease diagnosis in retinal image analysis could ease mass screening of populations with diabetes mellitus and help clinicians in utilizing their time more efficiently. The latest technological advances in computing power, communication systems, and machine learning techniques provide opportunities to the biomedical engineers and computer scientists to satisfy the requirements of clinical practice. In this paper proposed kernel PCA is used to feature selection, classifiers which include Support Vector Machine (SVM), KNN, Random Forests, Gradient boosting, AdaBoost, Naive Bayes used to detect the DR. Each algorithm is compared to other algorithms.**

***Keywords*-- Diabetic Retinopathy, SVM, KNN, Random forest, Gradient Boost, Adaboost, Naïvebayes.**

## I. INTRODUCTION

Data mining is the method of reading hidden patterns of facts according to distinct perspectives for categorization into useful information. Data mining techniques are used in lots of research regions, which includes mathematics, sciences, genetics and advertising. Data preprocessing is a statistics mining technique that involves transforming raw information into a comprehensible format [1]. Real-world records is regularly incomplete, inconsistent and plenty of mistakes. Data preprocessing is a supported technique of resolving such troubles. Data preprocessing deal with missing values in two methods:-

1. This technique commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a specific feature and a specific column if it has greater than 75% of lacking values. This approach is give advice only when there are sufficient samples inside the data set. One has to ensure that after we've got deleted the data, there is no addition of bias. Removing the data will result in loss of data with the intention to now not give the anticipated results at the same time as predicting the output [1].

2. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation that could add variance to the data set. But the loss of the records can be negated through this technique which yields higher results compared to elimination of rows and columns. Replacing with the above three approximations are a statistical technique of handling the missing values[5]. In our data set categorical records will cause trouble, so will convert into numerical values. To convert Categorical variable into Numerical data we will use Label_Encoder() class from preprocessing library.Label_Encoder is used to shifting categorical data into Numerical data.

The term demographics refer to particular characteristics of a populace. The word is derived from the Greek words for human beings (demos) and picture (graphy). Examples of demographic characteristics include age, race, gender, ethnicity, religion, education, sexual orientation, marital status, disability status and psychiatric prognosis. Patient demographics form the center of the data for any hospital. Demographic information presents data concerning research participants and is necessary for the determination of whether or not the individuals in a specific examine are a representative pattern of the target population for generalization purposes[17]. Usually demographics or research participant characteristics are said inside the techniques section of the research document and function impartial variables in the research design[2,3].

This session deals with the different techniques of machine learning to perform medical data analytics with the help of various techniques. Machine learning is a benevolent that delivers computers with the ability to learn without being explicitly programmed.

It is a well-known approach of data analytics and focuses on the development of computer programs such that they can teach themselves to grow and change when exposed to new data. The process of machine learning is similar to that of developments of actions in data mining. These systems are looking for through the data to look for patterns. However, instead of extracting data for human comprehension -- as is the case in data mining applications -- machine learning uses that data to improve the program's own understanding. Machine learning programs detect patterns in data and regulate program actions accordingly.

Machine Learning is a subfield of Computer Science and the scientific discipline that compacts with the construction and study of Algorithms based upon the data analyzed. Algorithms are designed by giving inputs to the model designed. Machine Learning is closely related to Computational Statistics in which helps us in making predictions or decisions. It has a strong attachment to Artificial Intelligence, Statistics and mathematical optimization. The example applications include Daibetic Retinopathy. This paper entirely describes about the procedure for well-known disease identification problematic with the help of certain mathematical techniques like clustering, regression, classification and it is used to perform machine-learning operations to overwhelm certain problem. These Machine-learning ideas used to overwhelm the memory problems. For Instance, if load any large dataset into Machine memory, an error may occur as the memory cannot be allocated due to memory insufficiency. This can be solved by increasing machine configuration or by parallelizing with commodity hardware.

Naïve Bayes is the simplest classifiers and amazingly well for many applications on the whole individuals linking the text classification. We supposed that a Diabetic Retinopathy Medical Records represented as D to classify the common method is to output that class $C_i$ whose probability of incidence $P(C_i|D)$ is maximum. To estimate the value of $P(C_i|D)$, this classifier naively undertakes that the attributes of D are independent of each other is known as Naïve Bayes. On one occasion individuality has been supposed that the derivation is used to compute $P(C_i | D)$ as follows

$$P(C_i | D) = P(D \wedge C_i) / P(D)$$
$$= P(D| C_i) P(C_i)/P(D)$$
$$\alpha P(D| C_i) P(C_i)$$
$$\alpha P(A_1=d_1| C_i) P(A_2=d_2| C_i) P(A_3=d_3| C_i) \ldots\ldots\ldots P(A_n=d_n| C_i) P(C_i)$$

At this time, the Diabetic Retinopathy Records of D contains attributes $A_n$ with values $d_n$. The denominator P(D) is overlooked, as it is common for all the classes.The last line of the derivation by presumptuous independence between the attributes. For classification is the values of $P(A_n=d_n|C_i)$ are pre-computed and stored for all possible attribute values and classes. At this point, the classification of these probability values are used to approximation $P(D|C_i)$ as per the above derivation and the class with the maximum probability of incidence is output. In this paper, we are using preprocessing, feature selection and classification using different approaches like Support Vector Machine (SVM),KNN(K-Nearest Neighbor) , Random Forests, Gradient boost, AdaBoost, Naive Bayes.

In this paper, we are using preprocessing, feature selection and classification using different approaches like Support Vector Machine (SVM),KNN(K-Nearest Neighbor) , Random Forests, Gradient boost,AdaBoost, Naive Bayes.

The paper is organized as follows: A brief description on Diabetic Retinopathy is given in section II. In section III discusses the dataset. In section IV discusses the Proposed Methods.Then in section V they obtained results and finally conclusions are drawn in section VI.

## II. DIABETIC RETINOPATHY

Diabetic retinopathy damages the retina of the patient. It is maximum frequent in the patients who've had diabetes for longer than 10 years. This problem is occurring in hundreds of thousands of human beings worldwide but medical practitioners and the tools required for detection of diabetic retinopathy is scare for serving the mass populace. The Diabetic Retinopathy (DR) is a clinical condition that damage the retina of eye which cause the blindness. With the right remedy and monitoring of the eyes the new instances of diabetic retinopathy may be decreased up to 90%. The excessive blood sugar damages the tiny blood vessels that elements blood to retina which cause diabetic retinopathy[3,7]. The light detected by retina transformed to sign which passes to mind through the optic nerve. In in advance stages of diabetic retinopathy the blood vessels starts leaking fluid or hemorrhage (bleed), distorting vision. In later

stages it leads to scarring and cell loss because of unusual growth in blood vessels[8].

Diabetic retinopathy has four stages:

1. **Mild Nonproliferative Retinopathy:** At this stage, microaneurysms occur. They are small areas of balloon-like swelling in the retina's tiny blood vessels.

2. **ModerateNonproliferative Retinopathy:** This degree is while blood vessels that nourish the retina are blocked.

3. **Severe Nonproliferative Retinopathy:** In this stage,many more blood vessels are blocked, depriving several areas of the retina with their blood supply. These areas of the retina send alerts to the body to grow new blood vessels for nourishment.

4. **Proliferative Retinopathy:** At this advanced level, the signals sent by the retina for nourishment cause the increase of new blood vessels. These new blood vessels are abnormal and fragile. They grow along the retina and along the surface of the clear, vitreous gel that fills the inside of the eye. By themselves, these blood vessels do not cause symptoms or vision loss. However, they have got skinny, fragile walls. If they leak blood, severe vision loss and even blindness can result[18].

The complete targeted eye tests used for detecting diabetic retinopathy consists of Visual acuity checking out, Tonometry, Pupil dilation and Optical coherence tomography. The visual acuity finished is used to measure the ability of person to see objects at diverse distances. Tonometry check done to measure the stress inside the eye. In Pupil dilation check, drop inside the eyes is placed that widen (dilate) the pupil and then clinician observe the retina and optic nerve. OCT works like extremely sound image. In OCT, the light is penetrated inside the eye which captures the detailed images of the tissues inside the eye[4].

The table-I shows the range of the exudates affected in diabetic retinopathy.

TABLE-I: Ranges of exudates affected in diabetic retinopathy

| Normal | Mild | Moderate | severe |
|---|---|---|---|
| Below 0.15% | 0.15% to 2.5% | 2.5% to 4% | Greater than 4% |

## III. DATASET

The IDRiDdataset is consists of retinal fundus images, database consisting of 516 images categorized in two parts:

- Retinal images with the signs of DR and/or DME.
- Normal retinal images (without signs of DR and/or DME).

The medical experts categorized the full set of 516 images with a variety of pathological conditions of DR and DME. The dataset is divided into training and testing set including of 413 (80%) and 103 (20%) images respectively by maintaining correctcombination of disease stratification. The diabetic retinal images were categorized into separate groups ranging from 0 (No apparent DR) to 4 (Severe DR) according to the International Clinical Diabetic Retinopathy Scale [9],The risk of macular edema can be determined by the existence of exudates [10], severity grading of DME is done based on eventsofhard exudates adjacent to macula center as per the definitions provided by Messidor_database[11].The data provided 516records with two attributes, One is image id and anther one is Retinopathy grade, and the class label is Risk of macular edema consists three outcomes (0,1,2). The table-II show the dataset.

TABLE-II: Dataset

| Image name | Retinopathy grade | Risk of Macular Edema |
|---|---|---|
| IDRiD_001 | 3 | 2 |
| IDRiD_002 | 3 | 2 |
| IDRiD_003 | 2 | 2 |
| IDRiD_004 | 3 | 2 |
| IDRiD_005 | 4 | 0 |
| IDRiD_006 | 4 | 1 |
| IDRiD_007 | 4 | 0 |
| IDRiD_008 | 4 | 2 |
| IDRiD_009 | 3 | 2 |
| IDRiD_010 | 4 | 1 |

The table-II contain the attributes show the following details.

- Image No: Name of identified and renamed patient image.

- DR(Diabetic Retinopathy) Grade: DR severity level in range 0 (No apparent DR) to 4 (Severe DR).
- Risk of DME (Diabetic Macular Edema): Macular edema severity level in range 0 (No DME) to 2 (Severe DME).

## IV. PROPOSED METHOD

The proposed method used in this study can be seen in Figure-I

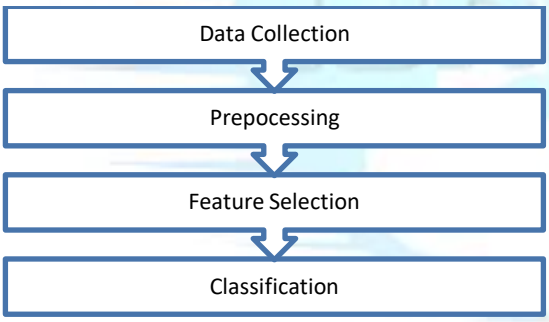


Figure I: Flow chart for Proposed Method

*1. Pre processing*

Preprocessing is required in order to correctly classify, Data preprocessing is a supported method of resolving such issues. Data preprocessing deal with missing values in two methods:- we delete a particular row if it has a null value or We can calculate the mean, median or mode of the feature and replace it with the missing values.

*2. Feature Selection*

Kernel PCA is offered as feature choice to detect disease[13]. Kernel PCA gives exact end result than PCA to select out its feature [13]. So, in this study we used kernel PCA as feature selection process to select the features. Extending the classical principal component analysis (PCA), the kernel PCA [14] extracts nonlinear structures of high dimensional data effectively. The motivation why we used kernel PCA as feature selection, because the features have high dimension.

*3. Classification*

In classification process, SVM,KNN, Random Forest, Gradient Boosting, Adaboost and Naive Bayes, are used to classify the instances[15].The dataset is imbalanced i.e. the class distribution is not identical. Table-III shows the number of examples belonging to each class.

Table-III: Number of Examples belonging to Classes

| Class | No of Examples |
|---|---|
| 0 | 222 |
| 1 | 51 |
| 2 | 243 |

In table-III contains three classes 0, 1 and 2.The total number of training examples becomes 516. And we have divided the dataset into train and test set having 413 and 103 records respectively.

## V. RESULTS AND DISCUSSION

The performance of detection can evaluate the performance of accuracy. This can be examined by using the as described on the Table-IV.

Table-VI: Confusion Matrix

| | | predicted | | Total |
|---|---|---|---|---|
| | | Yes | No | |
| Actual | Yes | TP | FN | T |
| | No | FP | TN | P |
| Total | | T’ | P’ | T+P or T’+P’ |

Here, TP, FP, FN,TN refer to the number of True positive, false positive, false negative, True negative

samples respectively. TP and TN, it means when the classifier is getting right while FP and FN when the classifier is gettingwrong, P is the number of positive set of instances and N is the Number of negative set of instances[16].

Sensitivity: Measurementof true positive ratio

Sensitivity = TP/TP+FN

Specificity: Measurement of true negative ratio

Specificity=TN/TN+FP

Accuracy: Measurement of correct classification

Accuracy = TP+TN/TP+FN+TN+FP

The results of each experiment can be show in Table-V

Table-V: sensitivity, specificity and accuracy

| | | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|
| **SVM** | C1 | 0.89 | 1 | 0.95 |
| | C2 | 0 | 1 | 0.92 |
| | C3 | 1 | 0.76 | 0.87 |
| **Naïve Bayes** | C1 | 0.80 | 1 | 0.59 |
| | C2 | 0 | 1 | 0.58 |
| | C3 | 1 | 0.73 | 0.56 |
| **KNN** | C1 | 0.89 | 1 | 0.94 |
| | C2 | 0 | 1 | 0.92 |
| | C3 | 1 | 0.76 | 0.88 |
| **Random Forest** | C1 | 0.80 | 0.90 | 0.86 |
| | C2 | 0.50 | 0.90 | 0.88 |
| | C3 | 0.83 | 0.91 | 0.86 |
| **Gradient boost** | C1 | 0.30 | 0.96 | 0.44 |
| | C2 | 0 | 0.98 | 0.34 |
| | C3 | 0.64 | 0.94 | 0.73 |
| **Ada boost** | C1 | 0.30 | 0.96 | 0.48 |
| | C2 | 0 | 0.98 | 0.36 |
| | C3 | 0.64 | 0.94 | 0.77 |

## VI. CONCLUSION

The early detection of diabetic retinopathy is crucial as it may result in blindness. Theproposed algorithm uses classifiers which include Support Vector Machine (SVM),KNN, Random Forest,Gradient boosting, AdaBoost, Naive Bayes and every presents one kind of accuracy. The SVMclassifier provides better testing accuracy whereas random forests Technique,KNN and Gradient boosting. The Naïve Bayes andAdaBoost classifier gave poor accuracy.

## REFERENCES

[1] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Gain: Missing data imputation using generative adversarial nets. pages 5689–5698, 2018.

[2] F. Fan, W. Cong, and G. Wang, "A New Type of Neurons for Machine Learning." arXiv preprint arXiv:1704.08362, 2017.

[3] Elia J. Duh, Jennifer K. Sun, and Alan W. Stitt, "Diabetic retinopathy: current understanding, mechanisms, and treatment strategies," JCI Insight, vol. 2, no. 14, pp. 1-13, jul 2017.

[4] KeleXu, Dawei Feng and HaiboMi:"Deep Convolutional Neural Network-Based Early Automated Detection of Diabetic Retinopathy Using Fundus Image", Molecules, 2017 (Open Access Journal), 22,2054; doi:10.3390/molecules22122054.

[5] DušanSovilj, Emil Eirola, YoanMiche, Kaj-Mikael Björk, RuiNian, Anton Akusok, and AmauryLendasse. Extreme learning machine for missing data using multiple imputations. Neurocomputing, 174:220–231, 2016.

[6] Karan Bhatia, Shikhar Arora and Ravi Tomar:"Diagnosis of Diabetic Retinopathy Using Machine Learning Classification Algorithm", 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016.

[7] Anupriyaa Mukherjee, DikshaRathore, Supriya Shree and Asst Prof. Shaikameel:"Diagnosis of Diabetic Retinopathy", Int. Journal of Engineering Research and Applications, ISSN : 2248-9622,Vol. 5, Issue 2, ( Part -4) February 2015.

[8] SohiniRoychowdhury, Student Member, IEEE, Dara D. Koozekanani, Member, IEEE, and Keshab K. Parhi, Fellow, IEEE "DREAM: Diabetic Retinopathy Analysis Using Machine Learning".

[9] Wu, L.; Fernandez-Loaiza, P.; Sauma, J.; Hernandez-Bogantes, E.; Masis, M. Classification of diabetic retinopathy and diabetic macular edema. World J. Diabetes 2013, 4, 290

[10] Giancardo, L.; Meriaudeau, F.; Karnowski, T.P.; Li, Y.; Garg, S.; Tobin, K.W., Jr.; Chaum, E. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. Med. Image Anal. 2012, 16, 216–226

[11] Decencière, E.; Zhang, X.; Cazuguel, G.; Lay, B.; Cochener, B.; Trone, C.; Gain, P.; Ordonez, R.; Massin, P.; Erginay, A.; et al. Feedback on a publicly distributed image database: The Messidor database. Image Anal. Stereol. 2014, 33, 231–234.

[12] D. Sarwinda and A. M. Arymurthy (2013), Feature selection using kernel PCA for Alzheimer's disease detection with 3D MRImages of brain. In Proc. International Conference

on Advanced Computer Science and Information Systems (ICACSIS), 329-333.

[13] B. Scholkopf, A. Smola, and K. R. Muller (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation. 10(1): 1299-1319.

[14] Harini R and Sheela N (2016). Feature extraction and classification of retinal images for automated detection of Diabetic Retinopathy. in Proc. International Conference on Cognitive Computing and Information Processing (CCIP), Mysore, 1-4.

[15] M. P. Paing, S. Choomchuay and M. D. RapeepornYodprom (2016). Detection of lesions and classification of diabeticretinopathy using fundus images. In Proc. Biomedical Engineering International Conference (BMEiCON), 1-5.

[16] M.U. Akram, A. Tariq, M.A. Anjum, M.Y. Javed, "Automated detectionof exudates in colored retinal images for diagnosis of diabeticretinopathy," OSA Journal of Applied Optics, vol. 51, pp. 4858–4866,2012.

[17] YugandharBokkaandR.N.V.Jagan Mohan , "Blockchain can strengthen the trustworthiness of meta-Analysis", AICTE Sponsored International Conference on Recent Trends in IoT and Blockchain. GIET University, Gunupur,19-20 October-2019, ISBN: 978-93-5391-198-0.

[18] YugandharBokkaandR.N.V.Jagan Mohan , "Patient Profile Data Optimization UsingBack Propagation", AICTE SponsoredNational Conference on Productivity, Quality, Reliability, Optimizationand Computational Modeling,SRKR Engineering College,18-20 December-2019.