Comparative Study of Diabetes Retinopathy using Hybrid Model

Prepared By

Anuraj Bal(16900215008)
Jeet Saha(16900215019)
Rounak Gupta(16900215033)
Sanmitra Kumar(16900215039)
Sushmita Shaw(16900215059)

Under the guidance of
Prof. Dr. Soumadip Ghosh

A Project Report
To be submitted in the partial fulfillment of the requirements
For the degree of
Bachelor of Technology in Information Technology

Department of Information Technology
Academy Of Technology

Affiliated to

Maulana Abul Kalam Azad University Of Technology,
West Bengal.
May, 2019

Academy Of Technology

CERTIFICATE

This is to certify that the project entitled (Comparative Study of Diabetes Retinopathy Using Hybrid Model) submitted to MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY in the partial fulfillment of the requirement for the award of the B.TECH degree in INFORMATION TECHNOLOGY is original work carried out by ANURAJ BAL, JEET SAHA, ROUNAK GUPTA, SANMITRA KUMAR, SUSHMITA SHAW Students(Roll.Nos:16900215008,16900215019,16900215033,16900215039,16900215059) under my guidance.

The matter embodied in this project is genuine work done by the student and has not been submitted whether to this University or to any other University/Institute for the fulfillment of the requirement of any course of study.

Dr. Soumadip Ghosh

Dated :

Department of
Information Technology
Academy of Technology
Aedconagar,Hooghly-
712121, West Bengal, India

Countersigned By

Dr. Soumadip Ghosh
Head, Department of Information Technology
Academy of Technology, Aedconagar,
Hooghly-712121, West Bengal, India

STATEMENT BY THE CANDIDATES

Roll- 16900215008
Roll- 16900215019
Roll- 16900215033
Roll- 16900215039
Roll- 16900215059

B. Tech 8th Semester
Dept. of Information Technology
Academy of Technology

We hereby state that the Project Report entitled Comparative Study of Diabetes Retinopathy using Hybrid Model has been prepared by us to fulfill the requirements of **IT 892** during the period July 2018 to May 2019.

_____
_____
_____
_____
_____
Signature

ACKNOWLEDGEMENT

At the very outset, we would like to convey our sincere gratitude to our beloved founder-chairman **Prof. J Banerjee** and respected director **Prof. D Bhattacharya** for all the encouragement and support extended to us during the tenure of this project and also our years of studies in this institute.

We are indebted to our guide **Dr. Soumadip Ghosh** for his/her epitome guidance, assistance and cooperation that facilitated the successful conclusion of our project.

We express our heartfelt thanks to our Head of the Department, **Dr. Soumadip Ghosh**, who has been actively involved and very influential from the start till the completion of our project.

We would also like to thank all teaching and non-teaching staffs of the Information Technology Department for their constant support and encouragement given to us. Last but not the least it is our great pleasure to acknowledge the wishes of friends and well wishers, both in academic and non-academic spheres.

_____

ANURAJ BAL (16900215008)

_____

JEET SAHA (169002150019)

_____

ROUNAK GUPTA (169002150033)

_____

SANMITRA KUMAR (16900215039)

_____

SUSHMITA SHAW (169002150059)

Abstract

Predicting diabetic retinopathy would help the doctors to detect the disease at the early stage. Recently researchers have used various machine learning algorithms to predict diabetic retinopathy. In this study a hybrid method of different types of classification algorithms have been employed for prediction. The proposed method first clusters the dataset using K-Means Clustering, then for each clusters an individual classification model is trained. The proposed method is compared with other classification models like DT (Decision Tree), NB (Naïve Bayes)-Classifier, LR (Logistic Regression) and KNN-Classifier. Experimental results have shown that Hybrid model using Logistic Regression achieved an average Accuracy of 83.71% and outperformed other models.

Keywords—diabetic retinopathy, hybrid classification model, k-means

LIST OF FIGURES

Figures **Page No**

LIST OF TABLES

Tables **Page No**

TABLE OF CONTENTS

Chapter 1

Introduction

Diabetic retinopathy is the most common form of diabetic eye disease. Diabetic retinopathy usually only affects people who have had diabetes (diagnosed or undiagnosed) for a significant number of years. People with diabetes can have an eye disease called diabetic retinopathy. This is when high blood sugar levels cause damage to blood vessels in the retina. These blood vessels can swell and leak fluid into the rear of the eye. Or they can close, stopping blood from passing through. Sometimes abnormal new blood vessels grow on the retina. All of these changes can steal your vision [1]. Retinopathy can affect all diabetics and becomes particularly dangerous, increasing the risk of blindness, if it is left untreated. The risk of developing diabetic retinopathy is known to increase with age as well with less well controlled blood sugar and blood pressure level [2].

1.1 Types of Diabetic Retinopathy

Diabetic retinopathy falls into two main classes: nonproliferative and proliferative. The word "proliferative" refers to whether or not there is neovascularization (abnormal blood vessel growth) in the retina. Early disease without neovascularization is called nonproliferative diabetic retinopathy (NPDR). As the disease progresses, it may evolve into proliferative diabetic retinopathy (PDR), which is defined by the presence of neovascularization and has a greater potential for serious visual consequences [3].

NPDR –Hyperglycemia results in damage to retinal capillaries. This weakens the capillary walls and results in small outpouchings of the vessel lumens, known as microaneurysms. Microaneurysms eventually rupture to form hemorrhages deep within the retina, confined by the internal limiting membrane (ILM). Because of their dot-like appearance, they are called "dot-and-blot" hemorrhages.

PDR – As mentioned earlier, the retina has a high metabolic requirement, so with continued ischemia, retinal cells respond by releasing angiogenic signals such as vascular endothelial growth factor (VEGF). Angiogenic factors, like VEGF, stimulate growth of new retinal blood vessels to bypass the damaged vessels. This is referred to as neovascularization. In PDR, the fibrovascular proliferation extends beyond the ILM. Proliferative retinopathy is an advanced stage of diabetic retinopathy in which the retina becomes blocked causing the growth of abnormal blood vessels. These can then bleed into the eyes, cause the retina to detach, and seriously damage vision. If left untreated, this can cause blindness. If proliferative retinopathy is regularly monitored and treated, the development of retinopathy can help be limited and more severe damage may be prevented.

Fig 1. NPDR

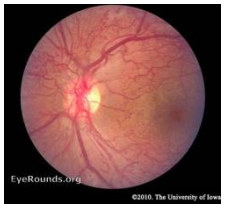

Fig 2. PDR

Diabetic retinopathy includes two different types namely, Background retinopathy and Diabetic maculopathy.

1.1.1 Background Retinopathy

Background retinopathy [4], also known as simple retinopathy, involves tiny swellings in the walls of the blood vessels. Known as blebs, they show up as small dots on the retina and are usually accompanied by yellow patches of exudates (blood proteins). Background diabetic retinopathy requires regular monitoring by an ophthalmologist. It is therefore important to attend regular retinopathy screening appointments.

1.1.2 Diabetic Maculopathy

The macula is the most well used area of the retina and provides us with our central vision. Maculopathy refers to a progression of background retinopathy into the macular and hence the name Diabetic Maculopathy [5]. This can cause vision problems such as difficulty with reading and or seeing faces in the centre of your vision.

1.2 Symptoms of Diabetic Retinopathy

Like many conditions of this nature, the early stages of diabetic retinopathy may occur without symptoms and without pain. An actual influence on the vision will not occur until the disease advances. Macular oedema can result from maculopathy and affect vision occurs if leaking fluid causes the macular to swell. New vessels on the retina can prompt bleeding, which can also block vision in some cases. Symptoms may only become noticeable once the disease advances, but the typical symptoms of retinopathy to look out for include:

- Sudden changes in vision / blurred vision
- Eye floaters and spots
- Double vision
- Eye pain
- Poor night vision
- Colors appear faded

1.3 Medical Treatment

Controlling your blood sugar and blood pressure can stop vision loss. Carefully follow the diet your nutritionist has recommended. Controlling your blood pressure keeps your eye's blood vessels healthy. One type of medication is called anti-VEGF medication. These include Avastin, Eylea, and Lucentis. Anti-VEGF medication helps to reduce swelling of the macula, slowing vision loss and

perhaps improving vision. This drug is given by injections (shots) in the eye. Laser surgery might be used to help seal off leaking blood vessels. This can reduce swelling of the retina.

1.4 Possible Prevention

Long-term good blood glucose level management helps to prevent diabetes retinopathy and lower the risk of developing it. Heart disease risk factors also affect retinopathy risk and include stopping smoking, having regular blood pressure and cholesterol checks and undergoing regular eye check-ups.

The risk of developing diabetic retinopathy can be lessened through taking the following precautions:

- Taking a dilated eye examination once a year.
- Managing diabetes strictly through medicine, insulin, diet and exercise.
- Test blood sugar levels regularly.
- Test urine for ketone levels regularly.

Chapter 2

Literature Overview

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data [6]. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Tom M. Mitchell provided a widely quoted and more formal definition: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [7]. The core of machine learning deals with representation and generalization. Representing the data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the ability of a machine learning system to perform accurately on new, unseen data instances after having experienced a learning data instance. The training examples come from some generally unknown probability distribution and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases. The performance of generalization is 11 usually evaluated with respect to the ability to reproduce known knowledge from newer examples. There are different types of machine learning, but the two main ones are:

- Supervised Learning
- Unsupervised Learning

2.1 Supervised Learning

Supervised learning is the machine learning task of inferring a function from supervised training data [8]. Training data for supervised learning includes a set of examples with paired input subjects and desired output. A supervised learning algorithm analyses the training data and produces an inferred function, which is called classifier or a regression function. The function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a reasonable way.

2.2 Unsupervised Learning

In unsupervised learning the machine simply receives inputs x1, x2. . . , but obtains neither supervised target outputs, nor rewards from its environment [9]. However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised

learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise.

2.3 Machine Learning In Medical Treatment

Computer-aided diagnosis of diabetic retinopathy has been explored in the past to reduce the burden on opthamologists and mitigate diagnostic inconsistencies between manual readers [10]. Automated methods to detect microaneursyms and reliably grade fundoscopic images of diabetic retinopathy patients have been active areas of research in computer vision [11]. The first artificial neural networks explored the ability to classify patches of normal retina without blood vessels, normal retinas with blood vessels, pathologic retinas with exudates, and pathologic retinas with microaneurysms. The accuracy of being able to detect microaneursyms compared to normal patches of retina was reported at 74% [12].

Diabetic retinopathy (DR) is the fastest growing cause of blindness, with nearly 415 million diabetic patients at risk worldwide. If caught early, the disease can be treated; if not, it can lead to irreversible blindness. Unfortunately, medical specialists capable of detecting the disease are not available in many parts of the world where diabetes is prevalent. It is believed that Machine Learning can help doctors identify patients in need, particularly among underserved populations. A few years ago, several of us began wondering if there was a way Google technologies could improve the DR screening process, specifically by taking advantage of recent advances in Machine Learning and Computer Vision. In "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs", published today in JAMA, presents a deep learning algorithm capable of interpreting signs of DR in retinal photographs, potentially helping doctors screen more patients in settings with limited resources.

Currently, detecting DR is a time-consuming and manual process that requires a trained clinician to examine and evaluate digital color fundus photographs of the retina. By the time human readers submit their reviews, often a day or two later, the delayed results lead to lost follow up, miscommunication, and delayed treatment.

Clinicians can identify DR by the presence of lesions associated with the vascular abnormalities caused by the disease. While this approach is effective, its resource demands are high. The expertise and equipment required are often lacking in areas where the rate of diabetes in local populations is high and DR detection is most needed. As the number of individuals with diabetes continues to grow, the infrastructure needed to prevent blindness due to DR will become even more insufficient.

The need for a comprehensive and automated method of DR screening has long been recognized, and previous efforts have made good progress using image classification, pattern recognition, and machine learning. With color fundus photography as input, the goal of this competition is to push an automated detection system to the limit of what is possible – ideally resulting in models with realistic clinical potential. The winning models will be open sourced to maximize the impact such a model can have on improving DR detection.

Chapter 3

Related Work

Diabetic retinopathy is the leading cause of blindness in the working-age population of the developed world. Since 1982, the quantification of diabetic retinopathy and detection of features such as exudates and blood vessels on fundus images were studied. A lot of work has been done in this field. Before starting implementation of main task we go through similar paper to know about the whole system such as what are the things we need to consider in order to detect diabetic retinopathy.

AkaraS. ,Matthew N. Dailey has proposed a "Machine learning approach to automatic exudate detection in retinal images from diabetic patients" [13]. In their paper they presented a series of experiments on feature selection and exudates classification using K- nearest Neighbor (KNN) and support vector machine (SVM) classifiers.

Rajendra Acharya U.,E. Y. K. Ng, Kwan-Hoong Ng and Jasjit S. Suri introduced algorithms for the automated detection of diabetic retinopathy using digital fundus images [14] where they improved an algorithm used for extraction of some features from digital fundus images. Moreover, Varun G. and Lily P. has used deep learning for detection of diabetic retinopathy [17].

In "Diagnosis of Diabetic Retinopathy using Machine Learning" research paper S. Gupta and K. AM tried to detect retinal micro-aneurysms and exudates retinal funds from images [15]. After pre-processing, morphological operations are performed to find the feature and the features are get extracted such as GLCM and splat for classification. They achieved the sensitivity and specificity of 87% and 100% respectively with accuracy of 86%.

Tiago T.G. in his paper "Machine Learning on the Diabetic Retinopathy Debrecen Dataset" has used R language for predicting diabetic retinopathy [16]. He used a dataset in which the features were extracted from images of the eye of a diabetic patient. In his work he used eight different classification algorithms and also shown some comparisons. He achieved 78% accuracy from his work.

Diabetic retinopathy is a leading cause of blindness among working-age adults. Early detection of this condition is critical for good prognosis. In this it demonstrate the use of convolutional neural networks (CNNs) on color fundus images for the recognition task of diabetic retinopathy staging. The network models achieved test metric performance comparable to baseline literature results, with validation sensitivity of 95% [18].

The work focuses on decision about the presence of disease by applying ensemble of machine learning classifying algorithms on features extracted from output of different retinal image processing algorithms, like diameter of optic disk, lesion specific (microaneurysms, exudates), image level (pre-screening, AM/FM, quality assessment). Decision making for predicting the presence of diabetic retinopathy

was performed using alternating decision tree, adaBoost, Naive Bayes, Random Forest and SVM [19].

Those are some related paper of our topic from where we took knowledge and idea to develop new version. In our work we will use different machine learning classification algorithms to classify diabetic retinopathy.

Chapter 4

Dataset Description

The source of the dataset that has been used for our study is collected UCI Repository. This dataset contains features extracted from Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of an anatomical part or an image-level descriptor. The Messidor database has been established to facilitate studies on computer-assisted diagnoses of diabetic retinopathy. We have seen different kind of datasets in kaggle, github and other websites which was used for different kind of projects based on diabetic retinopathy. As we wanted to work with detection of diabetic retinopathy, this dataset will be appropriate for our work as it has different types of features.

Our dataset contains different types of features that is extracted from the Messidor image set. This dataset is used to predict whether an image contains signs of diabetic retinopathy or not. The value here represents different point of retina of diabetic patients. First 19 columns in the dataset are independent variables or input column and last column is dependent variables or output column. Outputs are represented by binary numbers. "1" means the patient has diabetic retinopathy and "0" means absence of the disease.

Data Set Characteristics:	Multivariate	**Number of Instances:**	1151	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	20	**Date Donated**	2014-11-03
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	81963

Fig 3. Dataset Description

The dataset contains 20 columns, where each attributes represents various features of diabetic retinopathy extracted from the Messidor image set. We have total number of 1151 instances. Given below, is a description of the attributes of our dataset.

Feature indexes are

i. x1 – The binary result of quality assessment. 0=bad quality 1= sufficient quality.
ii. x2 –The binary result of pre-screening, where 1 indicates severe retinal abnormality and 0 its lack.
iii. x3 – x8 - The results of microaneurism detection. Each feature value stand for the number of microaneurisms found at the confidence levels alpha = 0.5, . . . , 1, respectively.
iv. x9 – x16 - contains the same information as x3 – x8 for exudates. However, as exudates are represented by a set of points rather than the number of pixels constructing the lesions, these features are normalized by dividing the number of lesions with the diameter of the ROI to compensate different image sizes.

v. x17 - The euclidean distance of the center of the macula and the center of the optic disc to provide important information regarding the patient's condition. This feature is also normalized with the diameter of the ROI.
vi. x18-The diameter of the optic disc.
vii. x19 - The binary result of the AM/FM-based classification.
viii. y - Class label. 1 = contains signs of Diabetic Retinopathy, 0 = no signs of Diabetic Retinopathy.

Here is a glimpse of the first 25 rows of our dataset.

Table 1.

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	Y
1	1	22	22	22	19	18	14	49.89576	17.77599	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	57.70994	23.79999	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	55.83144	27.99393	12.68749	4.852282	1.393889	0.373252	0.041817	0.007744	0.530904	0.128548	0	1
1	1	55	53	53	50	43	31	40.46723	18.44595	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	18.02625	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	15.4484	9.113819	1.633493	0	0	0	0	0	0.541743	0.139575	0	1
1	1	6	6	6	6	2	1	20.67965	9.497786	1.22366	0.150382	0	0	0	0	0.576318	0.071071	1	0
1	1	22	21	18	15	13	10	66.69193	23.54554	6.151117	0.496372	0	0	0	0	0.500073	0.116793	0	1
1	1	79	75	73	71	64	47	22.14178	10.05438	0.874633	0.09978	0.023386	0	0	0	0.560959	0.109134	0	1
1	1	45	45	45	43	40	32	84.3584	50.97746	17.29372	1.974419	0	0	0	0	0.546008	0.112378	0	0
1	0	25	25	25	23	22	18	22.48005	13.95	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	23.0128	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	65.11366	33.1248	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	123.0535	70.57101	37.40989	19.93725	14.78667	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
1	1	64	64	63	58	55	40	19.67346	6.064866	0.907342	0.080105	0	0	0	0	0.551182	0.098591	0	0
1	0	46	41	39	32	23	15	115.5338	21.29331	9.665742	2.276676	0.329396	0.186	0.118458	0.071698	0.540472	0.104949	1	1
1	1	37	37	37	34	31	23	61.35761	35.16591	8.114027	1.204871	0.178499	0.010772	0	0	0.478189	0.110793	0	0
1	1	19	17	15	12	12	7	179.704	34.6782	13.01895	1.045157	0.023003	0.005001	0.002	0	0.470425	0.094014	1	1
1	0	37	34	31	30	28	24	8.818234	3.161544	1.900918	1.524727	1.29287	0.165831	0	0	0.538223	0.09827	0	1
1	1	10	10	9	9	9	6	72.93894	20.28536	9.793215	0.916265	0.040814	0	0	0	0.528929	0.108156	1	0

Chapter 5

Project Objective

This project focuses on the prediction of diabetic retinopathy and analysis performed of different algorithm for the prediction. Machine learning algorithms such as KNN, NB, LR, DT etc. can be trained by providing training datasets to them and then these algorithms can predict the data by comparing the provided data with the training datasets. Our objective is to train our algorithm by providing training datasets to it and our goal is to detect diabetic retinopathy using different types of classification algorithms.

Here we have used a hybrid model to predict diabetic retinopathy along with its base models. Then a comparative study is done between the results obtained from both base model and hybrid model [20].

The hybrid model uses clustering algorithm to break the dataset into different clusters. K-means clustering is applied on training dataset, and each datapoint is assigned a particular cluster based on common features. Then for each cluster an individual model is trained by ML algorithm. Then in the testing phase, the datapoints are tested by the particular model chose based on the centroids of the clusters.

Chapter 6

Approach To Problem Solution

Machine Learning algorithms are the best possible way to solve any kind of prediction problems. A hybrid way of designing the model rather than using the classical models tends to improves the performance.

6.1 Proposed Method

The current study proposes a hybrid classification analysis method in order to improve the performance of classification models. The proposed hypothesis is motivated by the fact that instead of performing the classification analysis on a given dataset, it would be better if the dataset is clustered in to natural groups and then for each of those clusters a separate classification model is applied. Algorithm 1 explains the aforementioned concept. 'D' denotes the initial dataset. clustering_algorithm() depicts any clustering algorithm. After applying the unsupervised clustering algorithm, it returns a set of cluster centers (C) where $|C|$ is at least 2 and set of data points P where each member P_i is the set of data points of i^{th} cluster. In step 1, the clustering algorithm is applied on the initial dataset. In step 2, each set of data (P_i) is used to train a separate classification model (R_i). The train_classifier() can be any classification model training phase. Figure 3 depicts the working principle of the algorithm [20].

Algorithm 1: Hybrid Classification Model Training
Input: Initial Dataset (D)
Output: $C = \mathrm{U_i}\ C_i$, $i \geq 2$, where is C_i cluster center of i^{th} cluster, $P = \mathrm{U_i}\ P_i$, where P_i is the set of data points in i^{th} cluster, $R = \mathrm{U_i}\ R_i$, $i \geq 2$, where R_i is a classifier trained with data points of i^{th} cluster.

---

1. P , C = clustering_algorithm(D) ▷clustering_algorithm() is any unsupervised clustering algorithm
2. for each P_i in P do
   i. R_i = train_ classifier (P_i) ▷ train_ classifier () trains a classification model using data points in P_i
3. End

Algorithm 2 explains the testing phase of the hybrid classification model. In step (i) of 1, we calculate the nearest cluster center of the test data. $\|T_i - C_x\|$ depicts the distance between T_i and x^{th} cluster center C_x. Next, in step (ii), the corresponding classification model is used to calculate the predicted value of the target. Finally, in step 2, a performance metric is calculated. In the current study, Accuracy score is used as performance measure. Figure 4 depicts the working principle of the whole proposed model.

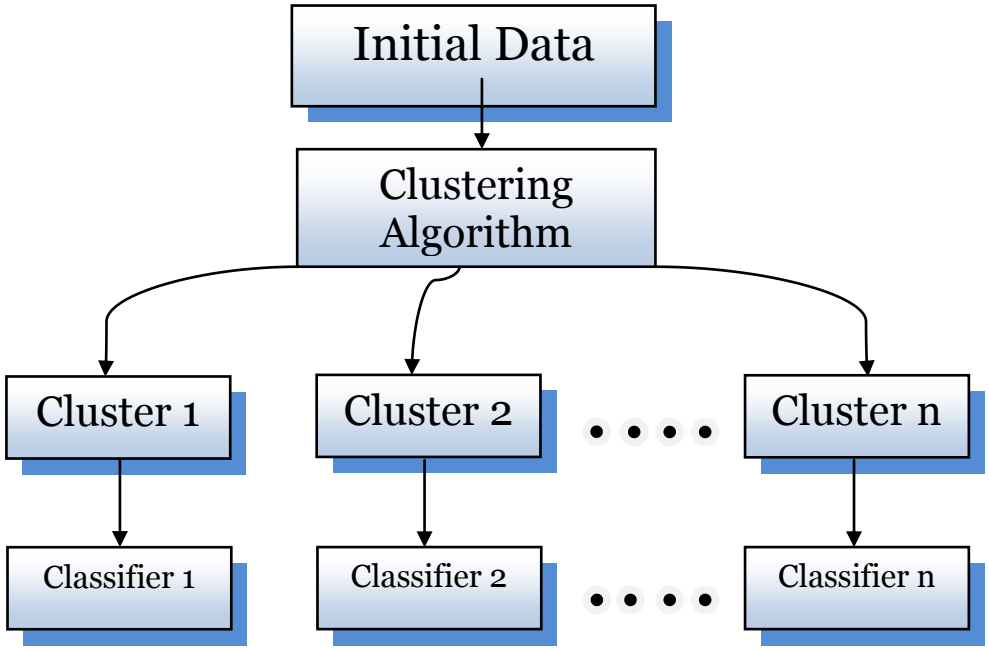

Fig. 4. Hybrid Classification Model training phase

Algorithm 2: Hybrid Classification Model Testing
Input: Testing Dataset (T)
Output: Performance metric

1. for each T_i in do
   i. k = arg min_x ($\|T_i - C_x\|$)
   ii. Predict target value using R_x
2. Calculate performance metric
3. End

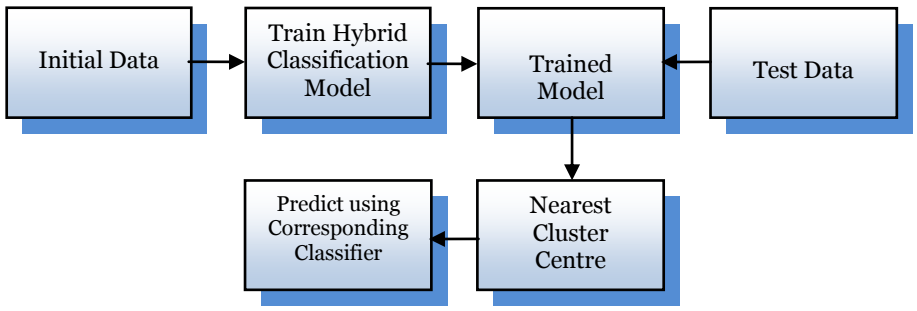

Fig. 5. Proposed Model working principle

6.2 Machine Learning Algorithms

The algorithms that we used to design our prediction model to combine with the hybrid model are discussed below.

6.2.1 Logistic Regression

In logistic regression , we have a hypothesis of the form:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta T x}} \quad ,$$

where g is the logistic function.

We assume that the binary classification labels are drawn from a distribution such that:

P(y = 1 | x; θ) = h_θ(x)

P(y = 0 | x; θ) = 1 − h_θ(x)

Given a set of labelled set of training examples, we choose θ to maximize the log-likelihood:

$$l(\theta) = \sum_i y^{(i)} log h(x^{(i)}) + (1 - y^{(i)}) log(1 - h(x^{(i)}))$$

We can maximize the log-likelihood by stochastic gradient ascent under which the update rule is the following (where α is the learning rate) and we run until the update is smaller than a certain threshold:

$$\theta = \theta + \alpha(y^{(i)} - h_\theta(x^{(i)})x^{(i)}$$

6.2.2 K-NEAREST NEIGHBORS CLASSIFICATION

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

EUCLIDEAN DISTANCE FORMULA: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

$$D^H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y => D = 0$$

$$x \neq y => D = 1$$

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

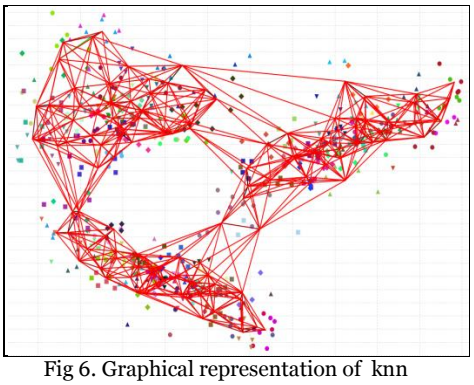

Fig 6. Graphical representation of knn

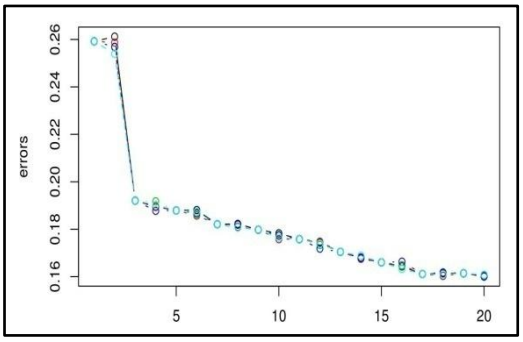

Fig 7. Effect of K value in error

6.2.3 Naïve Bayes

Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. Given a set of variables, X = {x1,x2,x...,xd}, we want to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1,c2,c...,cd}. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule:

$$p(C_j \mid x_1, x_2, \ldots, x_d) \propto p(x_1, x_2, \ldots, x_d \mid C_j) p(C_j)$$

where p(Cj | x1,x2,x...,xd) is the posterior probability of class membership, i.e., the probability that X belongs to Cj. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent we can decompose the likelihood to a product of terms:

$$p(X \mid C_j) \propto \prod_{k=1}^{d} p(x_k \mid C_j)$$

and rewrite the posterior as:

$$p(C_j \mid X) \propto p(C_j) \prod_{k=1}^{d} p(x_k \mid C_j)$$

Using Bayes' rule above, we label a new case X with a class level Cj that achieves the highest posterior probability.

Although the assumption that the predictor (independent) variables are independent is not always accurate, it does simplify the classification task dramatically, since it allows the class conditional densities p(xk | Cj) to be calculated separately for each variable, i.e., it reduces a multidimensional task to a number of one-dimensional ones. In effect, Naive Bayes reduces a high-dimensional density estimation task to a one-dimensional kernel density estimation. Furthermore, the assumption does not seem to greatly affect the posterior probabilities, especially in regions near decision boundaries, thus, leaving the classification task unaffected.

Naive Bayes can be modeled in several different ways including normal, lognormal, gamma and Poisson density functions:

$$p(x_k \mid C_j) = \begin{cases} \frac{1}{\sigma_{kj}\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_{kj})^2}{2\sigma_{kj}}\right), & -\infty < x < \infty, -\infty < \mu_{kj} <, \sigma_{kj} > 0 \quad \text{Normal} \\ \mu_{kj} : \text{mean}, \sigma_{kj} : \text{standard deviation} & \\ \frac{1}{x\sigma_{kj}(2\pi)^{1/2}} \exp\left\{\frac{-[\log(x/m_{kj})]^2}{2\sigma_{kj}^2}\right\}, & 0 < x < \infty, m_{kj} > 0, \sigma_{kj} > 0 \quad \text{Lognormal} \\ m_{kj} : \text{scale parameter}, \sigma_{kj} : \text{shape parameter} & \\ \frac{\left(\frac{x}{b_{kj}}\right)^{c_{kj}-1}}{b_{kj}\Gamma(c_{kj})} \exp\left(\frac{-x}{b_{kj}}\right), & 0 \le x < \infty, b_{kj} > 0, c_{kj} > 0 \quad \text{Gamma} \\ b_{kj} : \text{scale parameter}, c_{kj} : \text{shape parameter} & \\ \frac{\lambda_{kj}\exp(-\lambda_{kj})}{x!}, & 0 \le x < \infty, \lambda_{kj} > 0, x = 0,1,2,\ldots \quad \text{Poisson} \\ \lambda_{kj} : \text{mean} & \end{cases}$$

6.2.4 Decision Tree

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

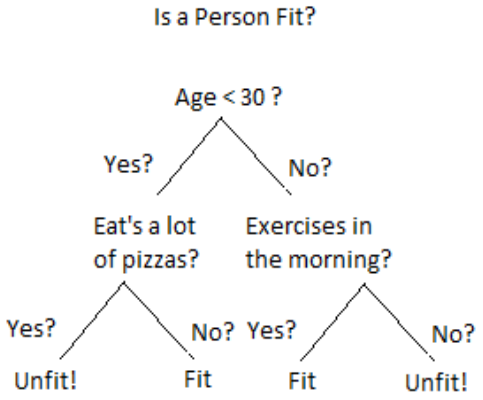

Fig 8. Decision Tree

Entropy

Entropy, also called as Shannon Entropy is denoted by H(S) for a finite set S, is the measure of the amount of uncertainty or randomness in data.

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Intuitively, it tells us about the predictability of a certain event. Example, consider a coin toss whose probability of heads is 0.5 and probability of tails is 0.5. Here the entropy is the highest possible, since there's no way of determining what the outcome might be. Alternatively, consider a coin which has heads on both the sides, the entropy of such an event can be predicted perfectly since we know beforehand that it'll always be heads. In other words, this event has no randomness hence it's entropy is zero.

In particular, lower values imply less uncertainty while higher values imply high uncertainty.

Information Gain

Information gain is also called as Kullback-Leibler divergence denoted by IG(S,A) for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables.

$$IG(S, A) = H(S) - H(S, A)$$

Alternatively,

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

where IG(S, A) is the information gain by applying feature A. H(S) is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature A, where P(x) is the probability of event x.

6.2.5 K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = {x_1,x_2,x_3,.....,x_n} be the set of data points and V = {v_1,v_2,.....,v_c} be the set of centers.

1. Randomly select 'c' cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4. Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3.

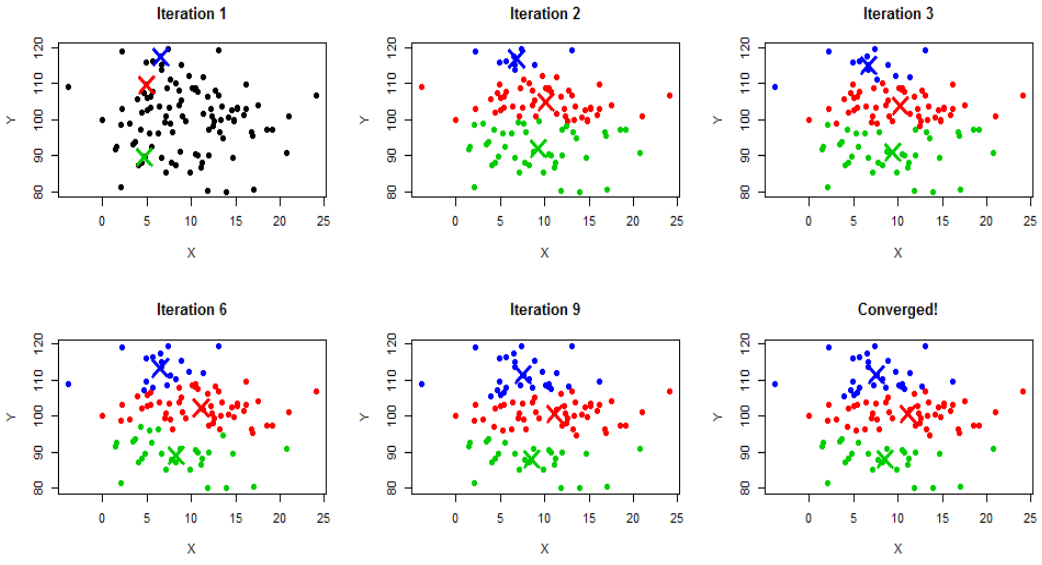

Fig 9. Process of Clustering

Chapter 7

Machine Learning Metrics

Metrics are used to measure the performance of the predicting models.

7.1 Confusion Matrix

The confusion matrix is a tabular representation that provides visualization of the performance of a classification algorithm. Each column of the matrix denotes the examples in a predicted class, while each row indicates the examples in an actual class. This helps to find out any type of misclassification due to the classifier. It provides more detailed analysis than classification accuracy. Classification accuracy is not a reliable metric for assessing the performance of a classifier as it may produce misleading results when the numbers of samples in different classes vary greatly. The confusion matrix entries can be defined as follows;

i. True positive (tp) is the number of 'positive' instances categorized as 'positive'.
ii. False positive (fp) is the number of 'negative' instances categorized as positive'.
iii. False negative (fn) is the number of 'positive' instances categorized as 'negative'.
iv. True negative (tn) is the number of 'negative' instances categorized as 'negative'.

Predicted class / Actual Class	Positive	Negative
Positive	tp	fn
Negative	fp	tn

7.2 Accuracy

Accuracy is defined as a ratio of sum of the instances classified correctly to the total number of instances.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{tp} + \boldsymbol{tn}}{\boldsymbol{tp} + \boldsymbol{fp} + \boldsymbol{fn} + \boldsymbol{tn}}$$

7.3 Precision

Precision is defined as the ratio of correctly classified data in positive class to the total number of data classified as to be in positive class.

$$\boldsymbol{Precision} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fp}}$$

7.4 Recall

Recall or TP rate is defined as the ratio of tp to the total number of instances classified under positive class.

$$\boldsymbol{Recall} = \frac{\boldsymbol{tp}}{\boldsymbol{tp} + \boldsymbol{fn}}$$

7.5 F-measure

F-measure is defined as a combined representation of Precision and Recall and is defined as follows:

$$\boldsymbol{F - measure} = \mathbf{2} * \frac{\boldsymbol{Precision} * \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

Chapter 8

Hardware & Software Requirements

Hardware requirements

- CPU: Intel Core i5 (6th gen), Variant: 8250U
- Clock speed: 1.6 GHz with Turbo Boost Upto 3.4 GHz
- Memory: 8 GB DDR4 RAM, Frequency: 2133 MHz
- Storage: Recommended minimum of 100 GB, or 300
- GPU: NVIDIA Geforce gtx 980 4 GB.
- Internet access to download the files from Anaconda Cloud.

Software requirements

- Software Details

Table 2.

Name	Type	Version	Architecture
Anaconda	Python distributer	Anaconda 2 4.2.0 Python 3.5	64bit (x86)
Spyder	Python IDE	2016.2.3 Build #PC 162.1967.10.	64 bit (x86)
Pandas	Python package	0.16.1 64	64 bit (x86)

- Operating System Details

Table 3.

Name	Microsoft Windows 10 Pro
Version	10.0.10586
Build Number	10586
System type	64 bit

Chapter 9

Pre processing & Feature Selection

Pre processing

The dataset that has been used for the study is an already processed dataset by extracting the features from Messidor image set. The missing values were originally replaced by 0 in the dataset. So the frequency of missing value was counted for each attribute and was replaced by the mean value of that particular attribute. The attributes were renamed for simplicity as x1, x2 ,. . . .x19, and the target variable as y.

Feature Selection

Choosing the best features among all the features is the most important task in order to get the best results. Python provides us with different types of feature selection methods. Among them dimensionality reduction, principle component analysis, correlation matrix, summary are the popular ones. Here we have used both correlation and PCA for choosing the best features.

The corr() method of Pandas dataframe is used to get the correlation matrix. From this matrix the attributes for which the correlation is negative, those attributes are dropped from the dataset.

Table 4.

Attribute Name	Co-Relation Value
X1	0.0628
X2	-0.0769
X3	0.2926
X4	0.2663
X5	0.2346
X6	0.1975
X7	0.1615
X8	0.1278
X9	0.0508
X10	0.0004
X11	0.0382
X12	0.1042
X13	0.1422
X14	0.1514
X15	0.1847
X16	0.1773

X17	0.0084
X18	-0.0306
X19	-0.0421
TARGET	1.000

Principal component analysis (PCA) is a fundamental multivariate data analysis method which is encountered into a variety of areas in neural networks, signal processing, and machine learning. It is an unsupervised method for reducing the dimensionality of the existing data set and extracting important information. PCA does not use any output information; the criterion to be maximized is the variance. PCA is used for dimensionality reduction. With the "n_components" value set to 15, pca is used to combine the attributes and reduce them to 15 attributes.

Principle Component Analysis (PCA)

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. If there are observations with variables, then the number of distinct principal components is min(n-1,p)}. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing n observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

The Figure represents a simple example of PCA. The data points in 2-dimensional space are reduced to 1-dimensional space by mapping the data to new coordinate u1. One crucial benefit of dimensionality reduction is reducing the computational complexity. Meanwhile, PCA can contribute to irrelevant noises removal. Thus, PCA is a good choice to improve the models and reduce the computational cost.

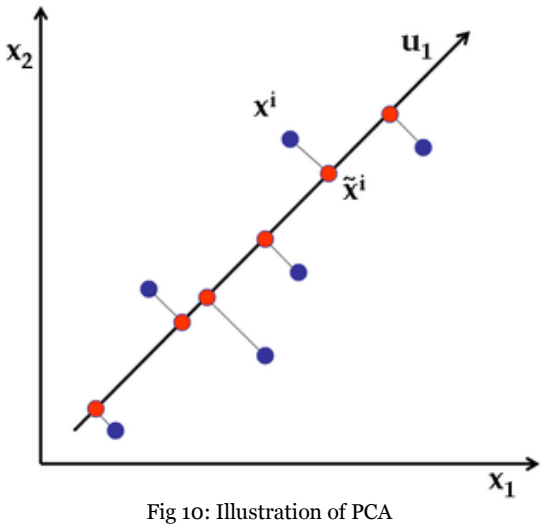

Fig 10: Illustration of PCA

Chapter 10

Prediction Models

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when separating a data set into two parts, most of the data is used for training, and a smaller portion of the data is used for testing. We have also split our dataset into two sets. One is for training and another for testing. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. After the model has been processed by using the training set, we have tested the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that we want to predict, it is easy to determine whether the model's guesses are correct or not. In addition, we have used 80% of our data for training and 20% for testing.

We went through a process of trial and error to settle on a short list of algorithms that provides better result as we are working on classification of diabetic retinopathy, we used some machine learning classification algorithms. The Machine Learning system uses the training data to train models to see patterns, and uses the test data to evaluate the predictive quality of the trained model. Machine learning system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics. So, for our thesis we will evaluate four different machine learning algorithms –

- Logistic Regression (LR)
- Decision Tree(DT)
- K-Nearest Neighbor (KNN)
- Naïve Bayes (NB)

For cross validation, to make sure that the model is neither underfitted nor overfitted, we have used k-folds cross validation.

In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. In our project we used 10-fold cross validation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

So, after preprocessing, feature selection and test-train split, the model is fitted with the training data. At first we have used LR to train the hybrid model. The train dataset is first clustered into 5 clusters. Then for each cluster an individual LR model is trained with the particular datapoints. Now for testing the hybrid model of LR is provided with the test dataset. Now the test datapoints are assigned clusters based on the centroids of the clusters and accordingly the

outputs are predicted from the concerned cluster. Then the same procedure is repeated for NB, KNN and DT.

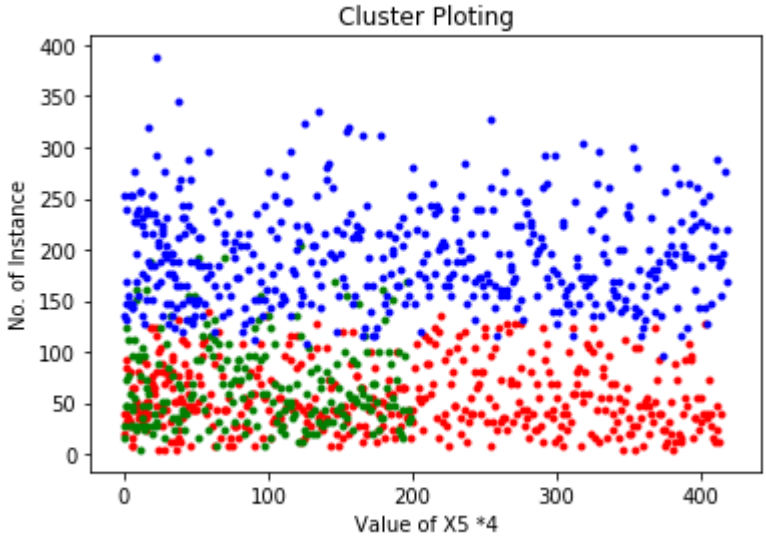

Fig 11. 2D Cluster Plot of Training Data

The fig 11 shows a 2D plot of the clustered training data. As the dataset contains 16 attributes so it's difficult to visualize the actual plot. So for simplicity the plot has been done in 2D, showing only 3 clusters instead of 5.

Now the training and testing is done again, but this time using the normal LR, DB, KNN and DT models.

Chapter 11

Results and Discussion

11.1 Effect of K value in Accuracy

Choosing the right value of K in k-means clustering is the most important aspect in getting the best cluster results. So, the foremost task in our study was to choose the k value. Starting with an initial k value of 3, it has been varied between 3 to 11 (considering the no. of datapoints) at interval of 2. For each k value the resultant Accuracy score of the hybrid model has been calculated. The table 5 shows the variation in Accuracy value with the increasing number of clusters i.e. with increment in the k value. It is clear from the result, that changing the value of k does have a significant effect on accuracy of each model. But for k=5 the hybrid logistic regression model gives the best result and outperformed other models. The dependency of Accuracy with k value can be better understood from the Fig. 12, For this study k value has been chosen as 5 as it gives the best result.

Table 5. Dependency of Accuracy on K.

Value Of K	Hybrid NB	Hybrid LR	Hybrid KNN	Hybrid DT
3	61.08%	73.02%	62.38%	61.75%
5	68.65%	**82.41%**	70.80%	**67.02%**
7	**69.47%**	69.28%	60.61%	59.69%
9	63.06%	70.12%	**73.32%**	63.71%
11	54.10%	63.87%	56.44%	54.82%

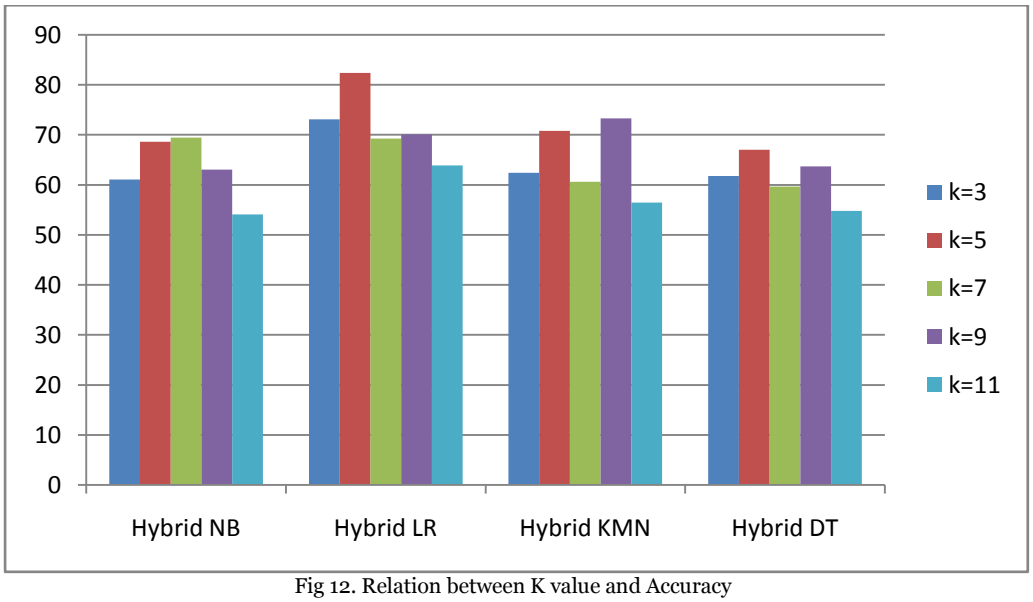

Fig 12. Relation between K value and Accuracy

11.2 Comparison of Hybrid Model to its Base Model

All the classification model has been compared with the hybrid approach of that classification model. Each classification model has been iterated for 10 times and finally the average Accuracy of the 10 iterations has been considered in an attempt to minimize the error. Similarly, the hybrid model of each classifier has been iterated 10 times to obtain the average Accuracy. From Table 6, it is clear that for every classification model, the hybrid model has outperformed its base model. A graphical representation of the comparison is shown in the Fig. 13. And out of the all the hybrid logistic regression has the best performance.

Table 6. Hybrid vs. Normal Classification Model(taking k=5)

NO. Of Iteration	HYBRID NB	NORMAL NB	HYBRID LR	NORMAL LR	HYBRID KNN	NORMAL KNN	HYBRID DT	NORMAL DT
1	82.07%	68.1%	83.54%	79.52%	7368%	67.24%	73.2%	63.79%
2	71.58%	68.1%	81.44%	79.52%	72.32%	67.24%	65.67%	60.34%
3	68.16%	68.1%	81.44%	79.52%	72.73%	67.24%	70.02%	64.65%
4	70.98%	68..1%	83.54%	79.52%	72.73%	67.24%	71.11%	65.51%
5	71.58%	68.1%	83.54%	79.52%	73.68%	67.24%	71.86%	62.93%
6	71.58%	68.1%	80.94%	79.52%	73.32%	67.24%	74.12%	63.79%
7	73.83%	68.1%	83.54%	79.52%	73.68%	67.24%	70.79%	62.93%
8	72.07%	68.1%	81.44%	79.52%	72.73%	67.24%	68.7%	62.06%
9	70.87%	68.1%	81.44%	79.52%	72.73%	67.24%	70.61%	61.2%
10	71.58%	68.1%	81.58%	79.52%	73.68%	67.24%	69.41%	63.79%
Average	**71.43%**	**68.1%**	**82.25%**	**79.52%**	**73.13%**	**67.24%**	**70.55%**	**63.1%**

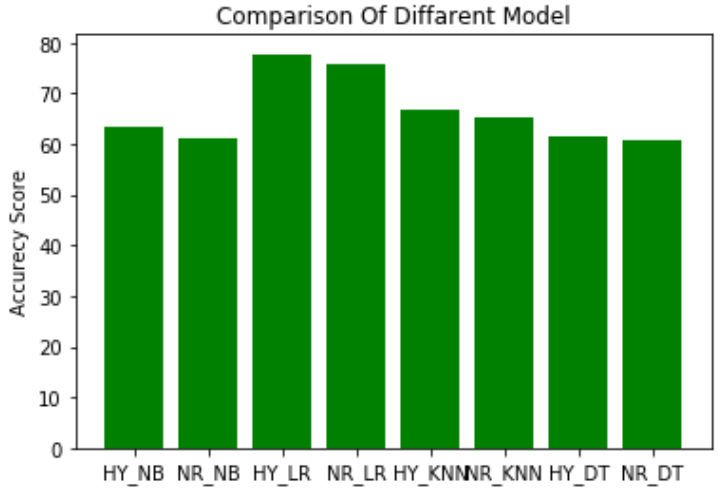

Fig 13. Comparison of Hybrid Model to its Base Model

11.3 Comparison of Hybrid LR Model with other Models

The current study reveals that logistic regression model has outperformed all other classification models. So, logistic regression has been coupled with the hybrid model to predict diabetic retinopathy. The hybrid model of logistic regression outperforms the base model of logistic regression and also the other classification models in every metrics (accuracy, precision, recall, f1 score). The average Accuracy of the hybrid model is calculated as 83.71% which is way better than other models. The Table 7 shows the accuracy, precision, recall, f1 score and confusion matrix of the hybrid logistic regression model and other models. The accuracy, precision, recall, f1 score of all the models has been represented graphically in Fig. 14, Fig. 15, Fig 16 and Fig 17 respectively.

Table 7. Performance of all models

MODEL	ACCURECY SCORE	PRECISION SCORE	RECALL SCORE	F1 SCORE	CONFUSION MATRIX
Naïve Bayes	65.76%	80.66%	49.79%	58.94%	[[46 6] [39 25]]
Hybrid LR	83.71%	87.42%	72.74%	78.32%	[[54 11] [9 42]]
K Nearest Neighbours	72.02%	71.96%	66.03%	68.47%	[[36 18] [20 42]]
Decision Tree	66.80%	68.63%	69.01%	68.24%	[[29 22] [25 40]]

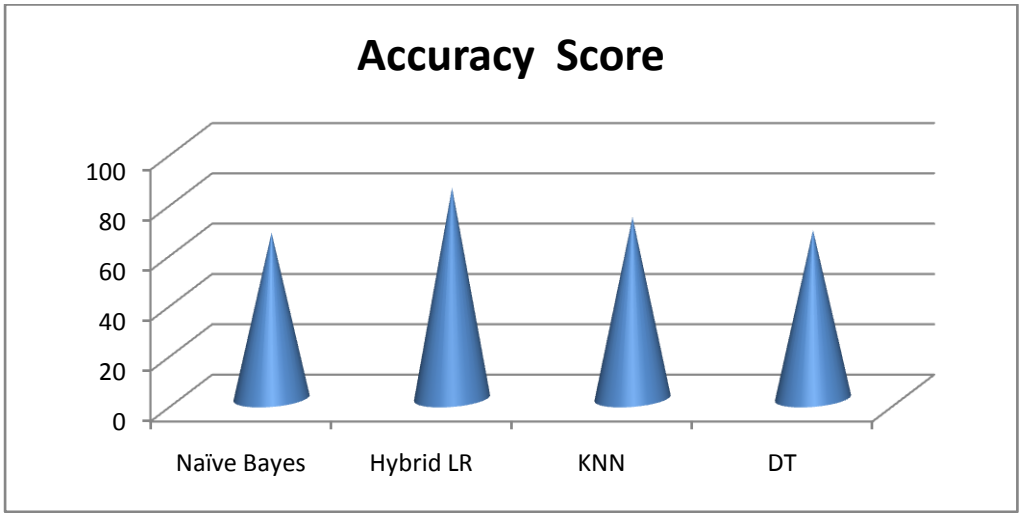

Fig 14. Comparison of accuracy score

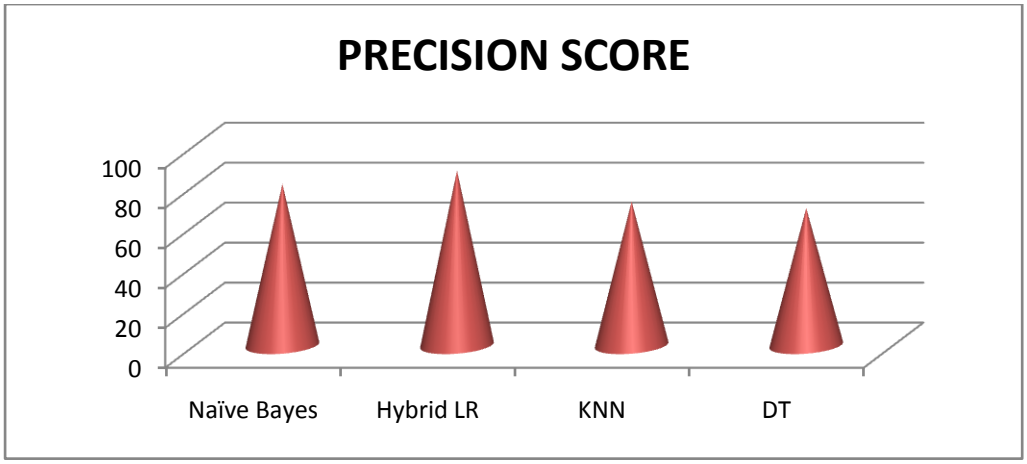

Fig 15. Comparison of precision score

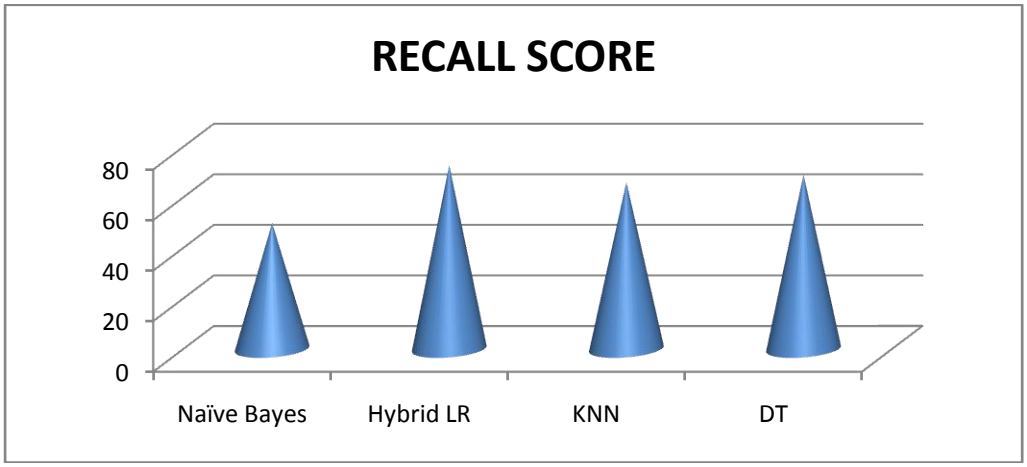

Fig 16. Comparison of recall score

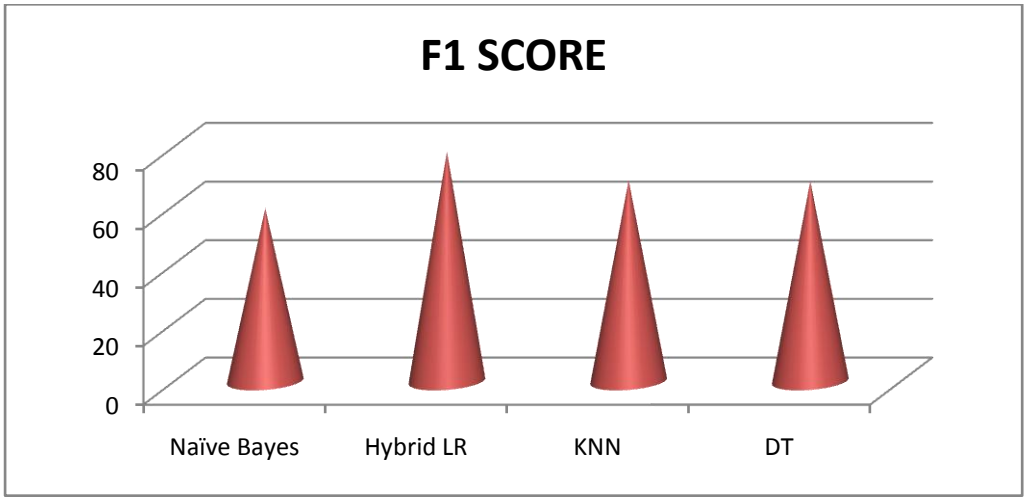

Fig 17. Comparison of f1 score

11.4 Conclusion

The current work proposed a hybrid classification model to predict diabetic retinopathy. The hybrid model involves an initial phase of data clustering. An extensive analysis is done where the proposed hybrid model frame work is compared with the base model on which the model is applied. The results have indicated that the hybrid model is highly capable of improving the performance of classification models to a greater extent in predicting diabetic retinopathy.

11.5 Future Scope

We are interested to do the work in different ways, one such method is image processing, where the different images of retina would be taken as dataset and analyzing would be done on that dataset.

Secondly, this can be improvised with a lot more categorized such as according to ages, genders, background studies, working facilities and so on.

For better result we have to take a large amount of data set, if we take the large amount of dataset then the machine will be trained better to provide the better result based on the algorithms used on it.

We can build a website or an android app for this purpose. In this way patient will be able to upload their data into our server and our machine learning software will let them know about their disease through our website whether it is in a good or bad condition.

We can improve the hybrid model by using different clustering algorithms other than k-means. The hybrid model can also be optimized by optimizing the hyper parameters of the clustering algorithms and the models that we have used.

REFERENCES

[1]. [Kierstan Boyd, G Atma Vemulakonda, MD]

[2]. https://www.diabetes.co.uk/diabetes-complications/diabetic-retinopathy.html

[3]. [Jesse Vislisel and Thomas Oetting, MS, MD]

[4]. https://www.diabetes.co.uk/diabetes-complications/background-retinopathy.html

[5]. https://www.diabetes.co.uk/diabetes-complications/diabetic-maculopathy.html

[6]. BoserB ,Guyon I.G,Vapnik V., "A Training Algorithm for Optimal Margin Classifiers", Proc. Fifth Ann. Workshop Computational Learning Theory,pp. 144-152, 1992.

[7]. Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7., McGrawHill, Inc. New York, NY, USA. Published on March 1, 1997

[8]. Alex C, Boston A. (2016).Artificial Intelligence, Deep Learning, and Neural Networks, Explained (16:n37)

[9]. Zoubin Ghahramani†, "Unsupervised Learning∗", Gatsby Computational Neuroscience Unit University College London, UK

[10]. Mookiah M., Acharya U., Chua C., Lim C., Ng E., Laude A. Computer-aided diagnosis of diabetic retinopathy: A review. In Computers in Biology and Medicine. 2013:2136–2155.[PubMed] [Google Scholar]

[11]. Philip S., Fleming A., Goatman K., Mcnamee P., Scotland G. The efficacy of automated disease/no disease grading for diabetic retinopathy in a systematic screening programme. In British Journal of Ophthalmology, 2007:1512–1517. [PMC free article] [PubMed] [Google Scholar]

[12]. Gardner G., Keating D., Williamson T., Elliott A. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. British journal of Ophthalmology. 1996;80(11):940–944. [PMC free article] [PubMed] [Google Scholar]

[13]. Varun G., Lily P., Mark C., "Development and validation of a deep learning Algorithm for Detection of Diabetic Retinopathy", December 2016.

[14]. Tiago T.G. "Machine Learning on the Diabetic Retinopathy Debrecen Dataset", knowledge-Based System60, 20-27. Published on June 25, 2016.

[15]. Boser B. E, Guyon I. M. and Vapnik V. N. (1992). "A training algorithm for optimal margin classiers".Proceedings of the 5th Annual Workshop on Computational Learning Theory COLT'92, 152 Pittsburgh, PA, USA. ACM Press, July 1992. On Page(s): 144-152

[16]. Leske MC, Wu SY, Nemesure B, Hennis A Barbados Eye Studies Group. Causes of visual loss and their risk factors: An incidence summary from the Barbados Eye Studies. Rev PanamSaludPublica.2010;27:259–67

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

[18]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961805/

[19]. Karan Bhatia ; Shikhar Arora ; Ravi Tomar, "Diagnosis of diabetic retinopathy using machine learning classification algorithm"

[20]. Sankhadeep Chatterjee, Sanmitra Kumar, Jeet Saha, Soumya Sen, " Hybrid Regression Model for Soil Moisture Quantity Prediction"