Diabetic Retinopathy Detection Using Machine Learning

By

Maisha Maliha – 14101013

Ahmed Tareque – 17341009

Sourav Saha Roy– 13301087

Inspiring Excellence

Supervisor

Hossain Arif

Assistant Professor

Department of Computer Science and

Engineering BRAC University

Thesis report submitted to BRAC University in accordance with the requirements for the degree of Bachelor of Science in Computer Science & Engineering

Submission Date: 05.04.2018

Author’s Declaration

Thesis Submission to the Department of Computer Science and Engineering, BRAC University, Dhaka, submitted by the authors for the purpose of obtaining the degree of Bachelor of Science in Computer Science. We hereby announce that the results of this thesis are entirely based on our research. Resources taken from any research conducted by other researchers are mentioned through reference. This thesis either in whole or in part, has not been previously submitted for any degree.

Signature of Supervisor:

……………………………

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
BRAC University

Signature of Authors:

…..………………………..

Maisha Maliha

……………………………

Sourav Saha Roy

……………………………

Ahmed Tareque

Acknowledgement

All the thanks and gratitude goes to Almighty ALLAHU SUBHANU WA TAWALA who is the creator of this vast universe and the source of all kind of knowledge and intelligence, the most merciful, gracious, who gave us the strength, guidance, patience and ability to complete the thesis.

Our heartfelt gratitude to our thesis supervisor Hossain Arif for his generous guidance and continued support and inspiration throughout our work. Without his help and assistance, we wouldn’t be able to finish our research successfully.

Besides this, we are thankful to Kalyan Banik, who provided us valuable guidance and suggestions in many stages of this work. His support has enabled us to go against all odds and complete this research properly.

We are extremely thankful to our parents, family members and friends for their support and encouragement. The journey would be much harder if they weren’t present for us in every moment whenever we needed them.

Finally we thank BRAC University for giving us the opportunity to use their resources and complete our thesis in the university.

Abstract

Diabetic Retinopathy (DR) is human eye disease among people with diabetics which causes damage to retina of eye and may eventually lead to complete blindness. Detection of diabetic retinopathy in early stage is essential to avoid complete blindness. Effective treatments for DR are available though it requires early diagnosis and the continuous monitoring of diabetic patients. Also many physical tests like visual acuity test, pupil dilation, and optical coherence tomography can be used to detect diabetic retinopathy but are time consuming. The objective of our thesis is to give decision about the presence of diabetic retinopathy by applying ensemble of machine learning classifying algorithms on features extracted from output of different retinal image. It will give us accuracy of which algorithm will be suitable and more accurate for prediction of the disease. Decision making for predicting the presence of diabetic retinopathy is performed using K-Nearest Neighbor, Random Forest, Support Vector Machine and Neural Networks.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DR – Diabetic Retinopathy

MA –Microaneurysms

FSH –Flame Shaped Hemorrhages.

KNN -K nearest Neighbors

SVM - Support Vector Machine

CART – Classification and Regression Tree

NNET- Neural Networks

CRF – Conditional Random Field

NPDR – Non- Proliferative Diabetic Retinopathy

PDR - Proliferative Diabetic Retinopathy

CHAPTER 1

INTRODUCTION

Diabetes is a chronic and organ disease that occurs when the pancreas does not secrete enough insulin or the body is unable to process it properly. Over time, diabetes affects the circular system, including that of the retina. Diabetes retinopathy (DR) is a medical condition where the retina is damaged because of fluid leaks from blood vessels into the retina. It is one of the most common diabetic eye diseases and a leading cause of blindness. Nearly 415 million diabetic patients are at risk of having blindness because of diabetics. It occurs when diabetes damages the tiny blood vessels inside the retina, the light sensitive tissue at the back of the eye. This tiny blood vessel will leak blood and fluid on the retina forms features such as micro-aneurysms, haemorrhages, hard exudates, cotton wool spots or venous loops. Diabetic retinopathy can be classified as non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR). Depending on the presence of features on the retina, the stages of DR can be identified. In the NPDR stage, the disease can advance from mild, moderate to severe stage with various levels of features except less growth of new blood vessels. PDR is the advanced stage where the fluids sent by the retina for nourishment trigger the growth of new blood vessels. They grow along the retina and over the surface of the clear, vitreous gel that fills the inside of the eye. If they leak blood, severe vision loss and even blindness can result.

Currently, detecting DR is a time-consuming and manual process that requires a trained clinician to examine and evaluate digital colour 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.

1.1 Motivation

As a research group, we wanted to do our undergraduate thesis on a research that will assist a huge amount of people in their healthy lives. The number of people with diabetic retinopathy is growing higher day by day. It is estimated that the number will grow from 126.6 million to 191.0 million by 2030 and the number with vision-threatening diabetic retinopathy (VTDR) will increase from 37.3 million to 56.3 million, if any proper action is not taken [4]. Despite growing evidence documenting the effectiveness of routine DR screening and early treatment, it is frequently leads to poor visual functioning and represents the leading cause of blindness. Most of the time it has been neglected in health care and in many low income countries because of inadequate medical service. While researching about these factors we get motivated to work with this topic. As there is insufficient ways to detect about diabetic retinopathy, we will build a system which will give prediction about diabetic retinopathy. Thus, we decided to use Machine Learning Algorithms for the prediction of this disease.

1.2 Objectives & Goals:

This thesis mainly focuses on the prediction of diabetic retinopathy and analysis performed of different algorithm for the prediction. Machine learning algorithms such as KNN, RF, SVM, NNET 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.

1.3 Thesis Orientation

Chapter 1 is the INTRODUCTION of the thesis. The motivation and objective & Goals of the thesis are described here.

Chapter 2 is LITERATURE REVIEW. This chapter consists of “Literature Review” which indicates our information collection repository. This chapter also consists of “Related Works and research” which indicates to the real life works and researches done by others, which are related to our thesis work in many ways.

Chapter 3 is PROPOSED MODEL FOR PREDICTION. This chapter consists of “Proposed Model” and “Implementation”.

Chapter 4 is EXPERIMENTAL RESULT ANALYSIS where we have shown the machine learning algorithms, which model gives maximum accuracy, which has better prediction. For analysing, we have used histograms, plots and different kind of comparison graphs and so on.

Chapter 5 is CONCLUSION which consists of “Conclusion Remarks” and “Future Works”.

CHAPTER 2

LITERATURE REVIEW

This chapter contains literature review related with supervised learning model, classification algorithms like KNN, NNET, random forest, and SVM. This chapter also refers the related works and research. Besides it will give information about our research activity.

2.1 Machine Learning

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data [4]. 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 [5].

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

2.2 Supervised Learning Model

Supervised learning is the machine learning task of inferring a function from supervised training data [6]. 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.

A simple analogy to supervised learning is the relationship between a student and a teacher. Initially the teacher teaches the student about a particular topic. Teaching the student the concepts of the topic and then giving answers to many questions regarding the topic. Then the teacher sets an exam paper for the student to take, where the student answers newer questions.

Figure 2.1 describes that the system learns from the data provided which contains the features and the output as well. After it has done learning, newer data is provided without outputs, and the system generates the output using the knowledge it gained from the data on which it trained. Here is how supervised learning model works.

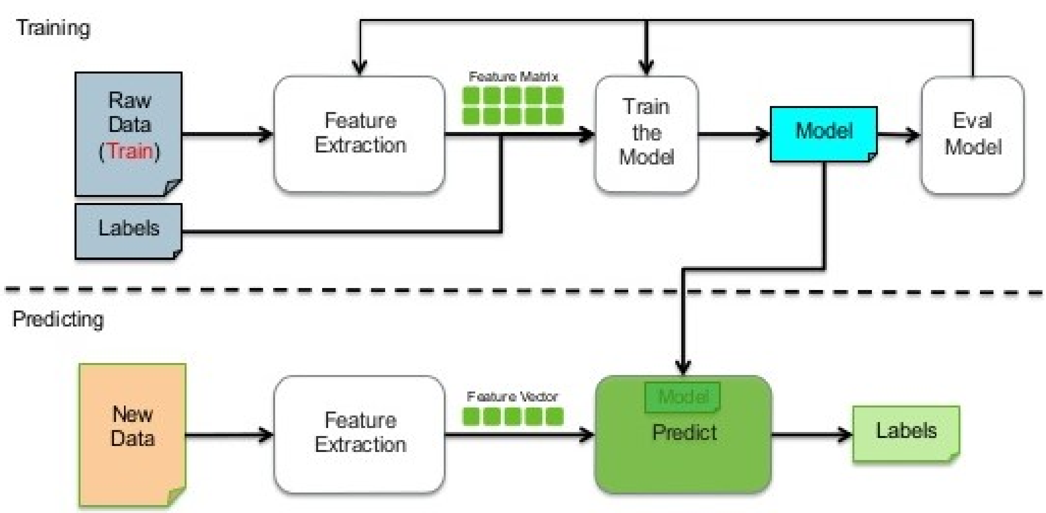

Figure 2.1: Workflow of supervised learning model

2.3 Algorithms

Since there are so many algorithms for machine learning, it is not possible to use all of them for analysis. For this research paper, we will be using four of them neural networks (NNET), random forest (RF), K-Nearest Neighbor (KNN) and support vector machine (SVM).

2.3.1 Neural Networks

Within the field of machine learning n neural networks are a subset of algorithms built around a model of artificial neurons spread across three or more layers [7]. There are plenty of other machine learning model which is notable for being adaptive in nature. Every node of neural network has their own sphere of knowledge about rules and functionalities to develop it-self through experiences learned from previous techniques that don't rely on neural networks. Neural networks are well-suited to identifying non-linear patterns, as in patterns where there isn't a direct, one-to-one relationship between the input and output [8]. This is a learning training. Neural networks are characterize by containing adaptive weights along paths between neurons that can be tuned by a learning algorithm that learns from observed data in order to improve model. One must choose an appropriate cost function. The cost function is what is used to learn the optimal solution to the problem being solved [7]. In a nutshell, it can adjust itself to the changing environment as it learns from initial training and subsequent runs provide more information about the world.

2.3.2 Random Forest

Random forest algorithm can use both for classification and the regression kind of problems. It is supervised classification algorithm which creates the forest with a number of tress [9]. In general, the more trees in the forest the more robust the forest looks like. It could be also said that the higher the number of trees in the forest gives the high accuracy results. There are many advantages of random forest algorithms. The classifier can handle the missing values. It can also model the random forest classifier for categorical values [10]. The over fitting problem will never come when we use the random forest algorithm in any classification problem. Most importantly it can be used for feature engineering which means identifying the most important feature out of the available feature from the training dataset.

2.3.3 K-Nearest Neighbors

K-nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure [11]. KNN has been used in statistical estimation and pattern recognition. KNN makes prediction for a new instance (x) by searching through the entire training set for the k most similar instances and summarizing the output variable for those k instances. For regression this might be the mean output variable, in classification this might be the mode class determine which of the k instances in the training dataset are most similar to new input many distance measure is used like Euclidean distance, Manhattan distance, Minkowski distance.

Distance functions

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

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

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Figure 2.2 Distance functions of KNN

2.3.4 Support Vector Machine

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik [12].

A more formal definition is that a support vector machine constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [13].

SVMs belong to the general category of kernel methods. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. This has two advantages: First, the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Second, the use of kernel functions allows the user to apply a classier to data that have no obvious fixed-dimensional vector space representation [14].

2.4 Cross Validation

Cross validation is the step where the best parameters for the algorithm are selected. The problem of overfitting and underfitting is discovered using cross validation. Normally a machine learning problem has many input feature, so it is not possible to visualize the data or the problems that might be occurring. Using cross validation, such problems can be identified via the learning curves. The two main problems encountered are underfitting and overfitting.

2.4.1 Underfitting

Underfitting occurs when the algorithm cannot properly fit the training set. The curve produced is probably not complex enough for the classification purpose. A synonym to underfitting is high bias.

To identify the presence of underfitting, learning curves need to be plotted. A learning curve with the training error and cross validation error needs to be plotted. If both the training error and 17 cross validation are high and there is a small gap between the curves, it can be positively inferred that the algorithm has underfit the training set.

Figure 2 shows a learning curve indication underfitting.

v

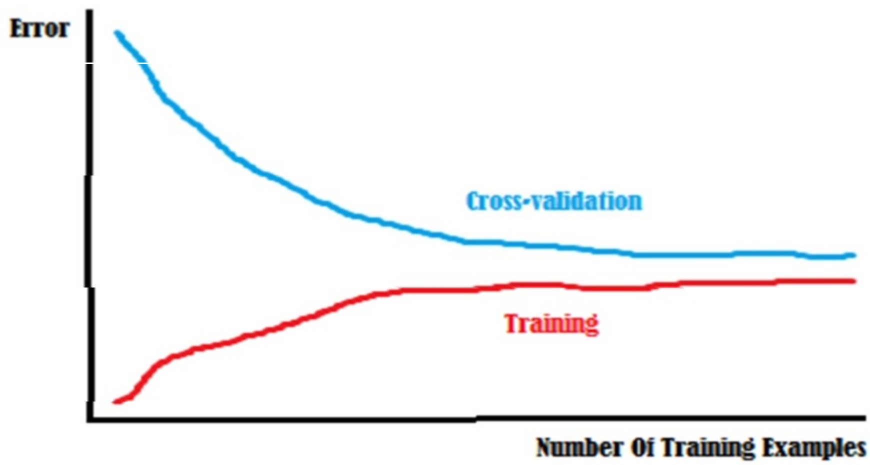

Figure 2.3 Curve showing underfitting

2.4.2 Overfitting

Overfitting occurs when the algorithm fits the training set a bit too well and does poorly in the test set. The algorithm fit the training set a bit too much, thus it was not able to generalize for unseen examples in the test set. A synonym to overfitting is high variance. Figure 3 shows a learning curve sample for the case of overfitting.

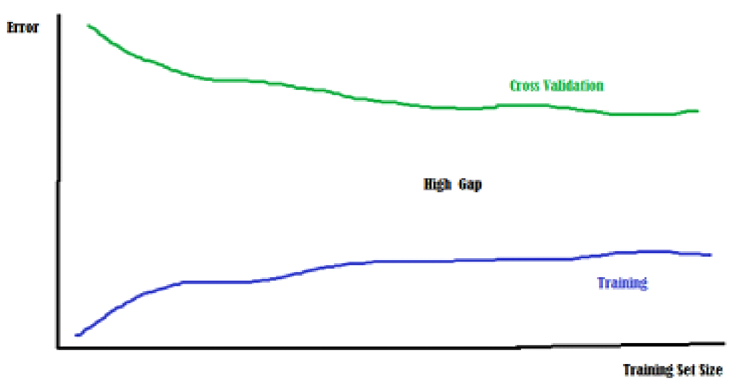

Figure 2.4 Curve showing Overfitting

2.5 Related Works & Research

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”[15]. 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 [16] 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 [3].

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 [17]. 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 [18]. 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.

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 3

PROPOSED MODEL FOR PREDICTION

This chapter contains proposed model, dataset collection, description, data visualization and also classifying algorithms that are used for analysis performance.

3.1 Proposed Model

Our First phase is data collection. We have collected our dataset from UCI Machine Learning repository website. The dataset contains features extracted from Messidor image set to predict whether an image have signs of diabetic retinopathy or not. Then features and labels of the dataset are identified. After that the dataset is divided into two sets, one for training where most of the data is used and the other one is testing. In training set four different classification algorithms has been fitted for the analysis performance of the model. The algorithms we used are k-Nearest Neighbor, random forest, support vector machine and neural networks. After the system has done learning from training datasets, newer data is provided without outputs. The final model generates the output using the knowledge it gained from the data on which it was trained. In final phase we get the accuracy of each algorithm and get to know which particular algorithm will give us more accurate results for the prediction of diabetic retinopathy.

Figure 3.1 shows the proposed model of our system. All the steps in sequential order are given.

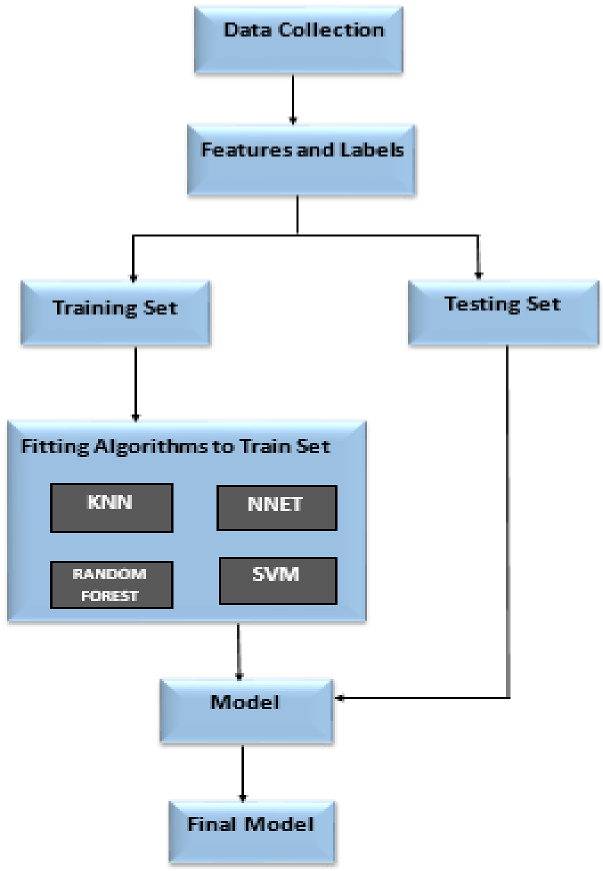

Figure 3.1 Proposed Model

3.2 Implementation

3.2.1 Data Collection

In our project we have used a dataset that is obtained from the UCI Machine Learning 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 detectedv 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.

3.2.2 Data Description

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.

The dataset has following features:

Table 3.1 Features of dataset

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:**	52330

Our dataset contains 20 columns, where each attributes represents various features of diabetic retinopathy. We have total number of 1151 instances. Here is a glimpse of the first 20 rows of our dataset.

Table 3.2 Sample of column headers in raw data

1	q	ps	nma.a	nma.b	nma.c	nma.d	nma.e	nma.f	nex.a	nex.b	nex.c	nex.d	nex.e	nex.f	nex.g	nex.h	dd	dm	class
2	1	1	22	22	22	19	18	14	49.8958	17.776	5.27092	0.77176	0.01863	0.00686	0.00392	0.00392	0.4869	0.10003	1
3	1	1	24	24	22	18	16	13	57.7099	23.8	3.32542	0.23419	0.0039	0.0039	0.0039	0.0039	0.52091	0.14441	0
4	1	1	62	60	59	54	47	33	55.8314	27.9939	12.6875	4.85228	1.39389	0.37325	0.04182	0.00774	0.5309	0.12855	0
5	1	1	55	53	53	50	43	31	40.4672	18.446	9.1189	3.07943	0.84026	0.27243	0.00765	0.00153	0.48328	0.11479	0
6	1	1	44	44	44	41	39	27	18.0263	8.57071	0.41038	0	0	0	0	0	0.47594	0.12357	0
7	1	1	44	43	41	41	37	29	28.3564	6.93564	2.30577	0.32372	0	0	0	0	0.50283	0.12674	0
8	1	0	29	29	29	27	25	16	15.4484	9.11382	1.63349	0	0	0	0	0	0.54174	0.13958	0
9	1	1	6	6	6	6	2	1	20.6796	9.49779	1.22366	0.15038	0	0	0	0	0.57632	0.07107	1
10	1	1	22	21	18	15	13	10	66.6919	23.5455	6.15112	0.49637	0	0	0	0	0.50007	0.11679	0
11	1	1	79	75	73	71	64	47	22.1418	10.0544	0.87463	0.09978	0.02339	0	0	0	0.56096	0.10913	0
12	1	1	45	45	45	43	40	32	84.3584	50.9775	17.2937	1.97442	0	0	0	0	0.54601	0.11238	0
13	1	0	25	25	25	23	22	18	22.48	13.95	0.43623	0.11612	0	0	0	0	0.55168	0.13966	1
14	1	1	70	69	65	63	63	50	10.5601	3.10836	0.62551	0.28796	0.10399	0.0048	0	0	0.5344	0.08959	0
15	1	1	48	43	39	32	27	18	23.0128	6.73758	2.4039	0.18924	0.01144	0	0	0	0.50155	0.13829	1
16	1	1	94	93	92	89	86	77	8.61082	1.98132	0.40118	0.0661	0	0	0	0	0.54128	0.12451	0
17	1	1	20	18	16	15	13	9	65.1137	33.1248	8.78538	0.67354	0.05181	0.00293	0.00098	0.00098	0.56946	0.08994	1
18	1	1	105	95	81	66	46	32	123.053	70.571	37.4099	19.9373	14.7867	6.11491	2.34574	1.00224	0.52446	0.13425	1
19	1	1	25	25	24	23	22	19	17.0341	9.97694	1.06724	0.48483	0.46779	0.3067	0.18898	0.13011	0.552	0.10843	0
20	1	1	64	64	63	58	55	40	19.6735	6.06487	0.90734	0.08011	0	0	0	0	0.55118	0.09859	0

Feature indexes are-

i. q – The binary result of quality assessment. 0=bad quality 1= sufficient quality.

ii. ps –The binary result of pre-screening, where 1 indicates severe retinal abnormality and 0 its lack.

iii. nma.a - nma.f - 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. nex.a – nex.h - contains the same information as nma.a - nma.f for exudates. However, as exudates are represented by a set of points rather than the number of pixels constructingthe lesions, these features are normalized by dividing the number of lesions with the diameter of the ROI to compensate different image sizes.

v. dd - 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. dm-The diameter of the optic disc.

vii. amfm - The binary result of the AM/FM-based classification.

viii. class - Class label. 1 = contains signs of Diabetic Retinopathy, 0 = no signs of Diabetic Retinopathy.

We have also calculated count, mean, max, standard deviation of the values in our dataset.

	q	ps	nma.a	nma.b	nma.c
count	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000
mean	0.996525	0.918332	38.428323	36.909644	35.140747
std	0.058874	0.273977	25.620913	24.105612	22.805400
min	0.000000	0.000000	1.000000	1.000000	1.000000
25%	1.000000	1.000000	16.000000	16.000000	15.000000
50%	1.000000	1.000000	35.000000	35.000000	32.000000
75%	1.000000	1.000000	55.000000	53.000000	51.000000
max	1.000000	1.000000	151.000000	132.000000	120.000000

	nma.d	nma.e	nma.f	nex.a	nex.b
count	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000
mean	32.297133	28.747176	21.151173	64.096674	23.088012
std	21.114767	19.509227	15.101560	58.485289	21.602696
min	1.000000	1.000000	1.000000	0.349274	0.000000
25%	14.000000	11.000000	8.000000	22.271597	7.939315
50%	29.000000	25.000000	18.000000	44.249119	17.038020
75%	48.000000	43.000000	32.000000	87.804112	31.305692
max	105.000000	97.000000	89.000000	403.939108	167.131427

	nex.c	nex.d	nex.e	nex.f	nex.g
count	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000
mean	8.704610	1.836489	0.560738	0.212290	0.085674
std	11.567589	3.923224	2.484111	1.057126	0.398717
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.249050	0.081554	0.000000	0.000000	0.000000
50%	4.423472	0.484829	0.022248	0.001554	0.000000
75%	11.766880	1.921649	0.191953	0.038450	0.004832
max	106.070092	59.766121	51.423208	20.098605	5.937799

	nex.h	dd	dm	amfm	class
count	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000
mean	0.037225	0.523212	0.108431	0.336229	0.530843
std	0.178959	0.028055	0.017945	0.472624	0.499265
min	0.000000	0.367762	0.057906	0.000000	0.000000
25%	0.000000	0.502855	0.095799	0.000000	0.000000
50%	0.000000	0.523308	0.106623	0.000000	1.000000
75%	0.003851	0.543670	0.119591	1.000000	1.000000
max	3.086753	0.592217	0.219199	1.000000	1.000000

Figure 3.2 Descriptive Statistics of Dataset

3.2.3 Data Visualization

Another important feature in the data distribution is the skewness of each class. Data visualization helps to see how the data looks like and also what kind of data correlation we have. The dataset distribution of each feature is shown below in figure 3.5. This is a histogram. A histogram is an accurate graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. Histograms are a great way to get to know your data. They allow you to easily see where a large and a little amount of the data can be found. In short, the histogram consists of an x-axis and a y-axis, where the y-axis shows how frequently the values on the x-axis occur in the data.

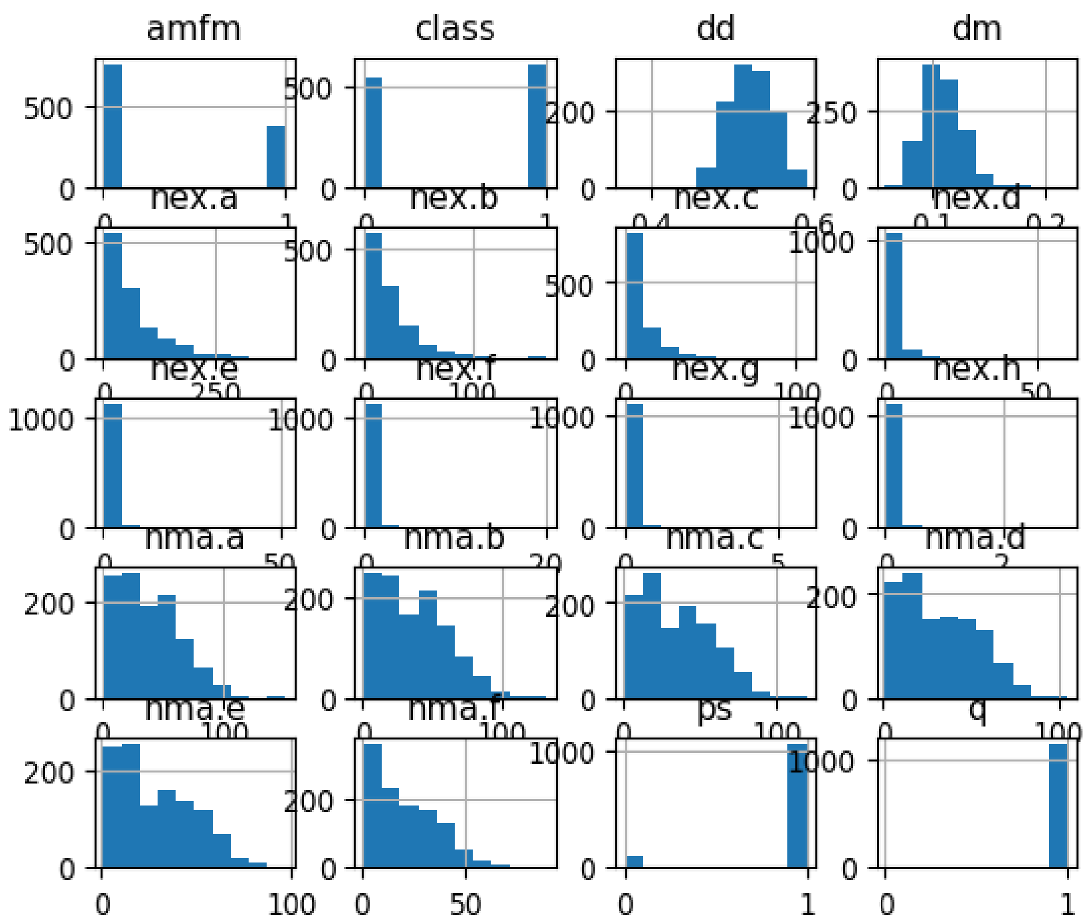

Figure 3.3 Data Distribution plot for each of the feature

As the given input variables are numeric, we can also create box plot.

A Boxplot typically provides the median, 25th and 75th percentile, min/max that is not an outlier and explicitly separates the points that are considered outliers.

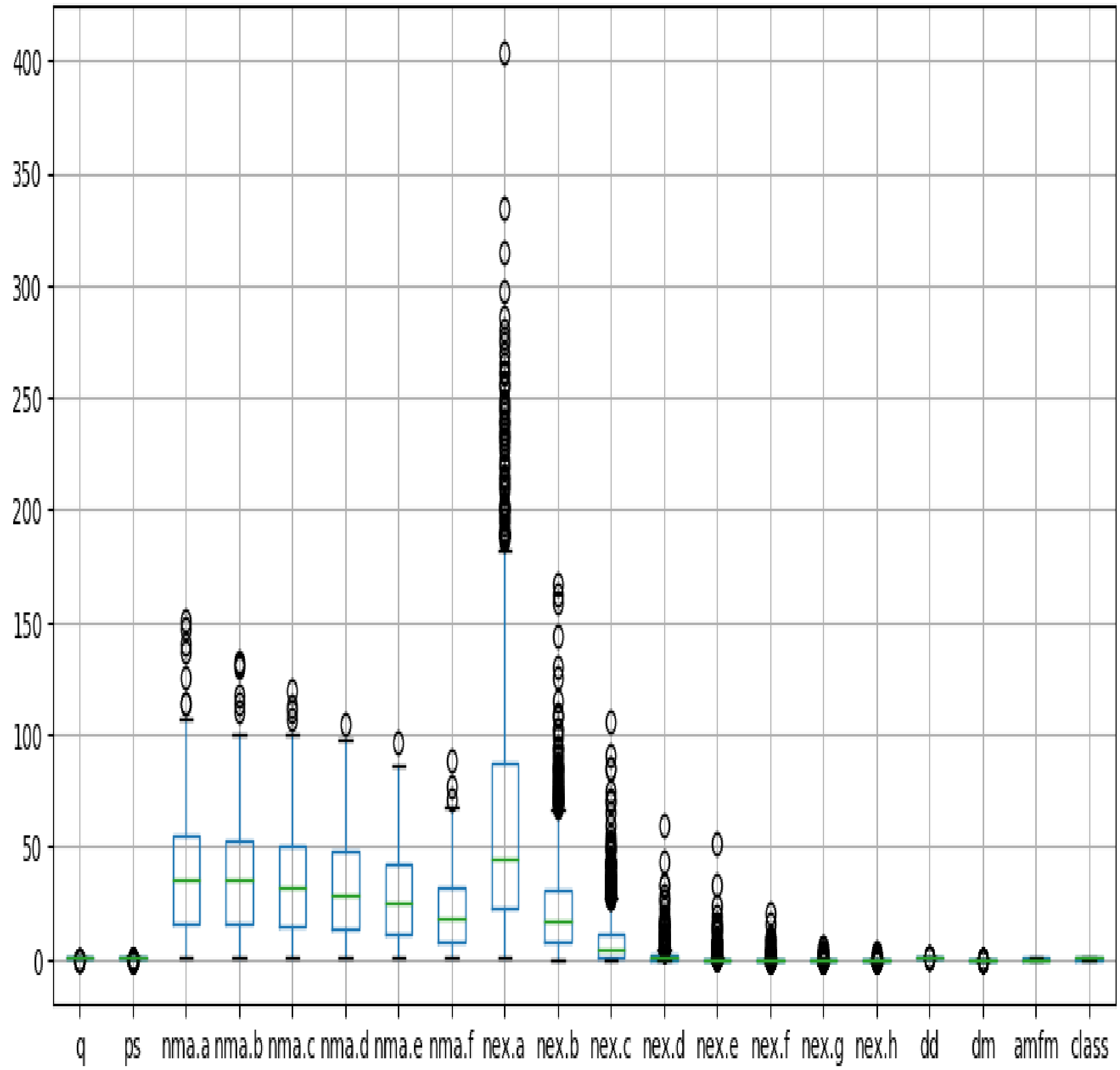

Figure 3.4 Box plot for each feature in dataset

3.2.4 Split Dataset

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.

3.3 Applying Algorithm

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. We get an idea from the data visualizations plots which algorithms will be suitable for the classification problem. 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 –

- Neural Networks (NNET)
- Random Forest
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)

3.4 K-Fold Cross Validation

K-Fold Cross Validation is common types of cross validation that is widely used in machine learning. 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.

3.5 System Setup

Hardware and software used in this research played a big role in terms of results. Both hardware and software specifications have been mentioned here.

3.5.1 Hardware Specification

CPU:

Table 3.3: CPU Specification

Name	AMD FX(tm)-8300
Cores	8
Clock speed (mhz)	3300
Typical TDP	95W
Socket	Socket AM3+
Micro architecture	Pile driver
Platform	Volan

Processor core	Vishera
Core stepping	OR-C0
CPUID	600F20
Manufacturing process	0.032 micron
Data width	64 bit
Level 1 cache size	4 x 64 KB 2-way set associative shared instruction caches 8 x 16 KB 4-way set associative data caches
Level 2 cache size	4 x 2 MB 16-way set associative shared exclusive caches
Level 3 cache size	8 MB 64-way set associative shared cache

Memory:

Table 3.4: GPU Specification

Physical memory	16GB
GPU	NVIDIA GeForce GT 620

3.5.2 Software Specification

Table 3.5: Software details

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 bit (x86)

OS:

Table 3.6: Operating System Details

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

CHAPTER 4

EXPERIMENTAL RESULTS & ANALYSIS

In the previous chapter we have discussed about proposed system and implementation of our thesis. We have demonstrated how we collected our dataset, dataset description, visualization and algorithms we used. Now we discussing about the results we obtained from our experiments upon the implementation of this system. We have divided our dataset into two parts- training and testing dataset. In this chapter we will show the outcome of the training and testing dataset. As mentioned before we have used four machine learning algorithms. First, we trained our dataset with these four algorithms and then we built a model. Then, we tested our testing dataset in this model. If the test set accuracy is near to train set accuracy then we can conclude that we built a good model.

We have total 1151 data of different individual in our dataset. There are 1151 rows and 20 columns in the dataset. After splitting the data into two parts now we have 920 rows for train data and for test data we have 231 rows. When we trained our train data for analysis performance of different algorithms. This is the result we got-

4.1 Training Accuracy of SVM Algorithm

For SVM algorithm we got training accuracy of 57.93%.We know Support Vector Machines a classifier that is defined using a separating Hyper plane between the classes. As SVM is capable of doing both classification and regression and also can capture much more complex relationships between data points we choose this algorithm. But for our training dataset the accuracy of SVM is 57.93% which is not quite satisfactory.

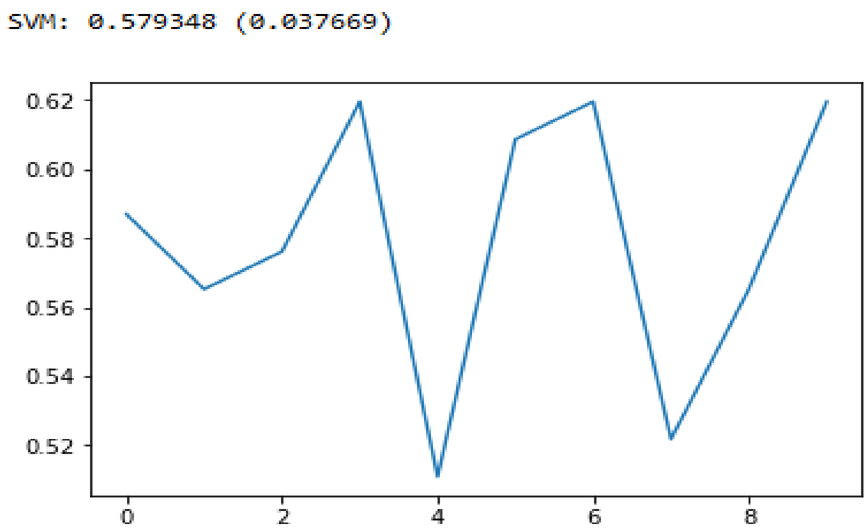

Figure 4.1: Training accuracy of SVM

4.2 Training Accuracy of KNN Algorithm

For KNN algorithm our training accuracy is 66.74%. K-Nearest Neighbor makes predictions using the training dataset directly. When KNN is used for classification, the out can be calculated as the class with the highest frequency from k-most similar instances. As we want to classify our result into two part we decided to use this algorithm. In Figure 4.2 shows that KNN gives around 66.74% accuracy for the training set which is quite acceptable.

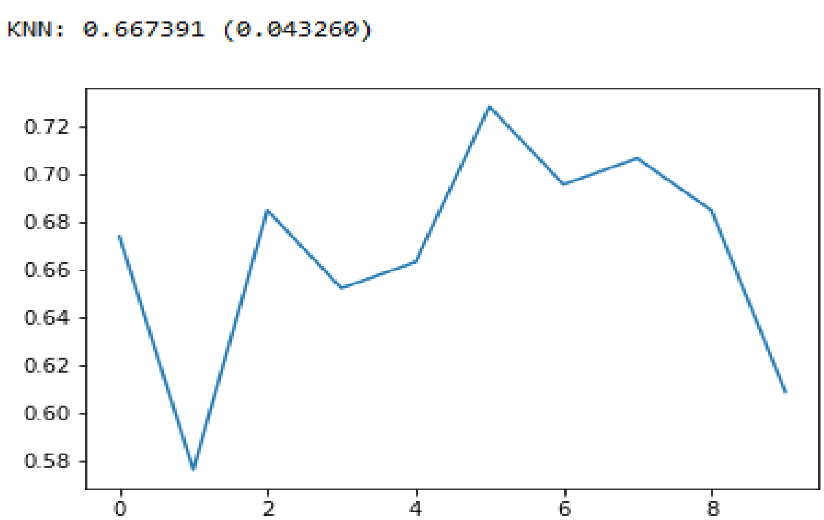

Figure 4.2: Training accuracy of KNN

4.3 Training Accuracy of Random Forest Algorithm

For Random Forest our training accuracy is 66.09%.Random Forest can also be used for both classification and regression like SVM. We see that the accuracy result of KNN and Random Forest is quite close. Random forest gives an accuracy of 66.09% which also can be considered good for our work.

```
RF: 0.660870 (0.063529)
```

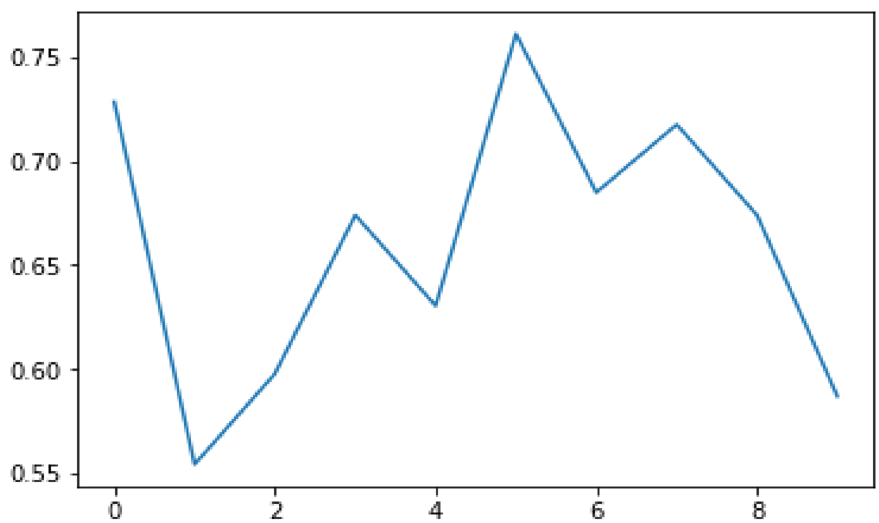

Figure 4.3: Training accuracy of Random Forest

4.4 Training Accuracy of NNET Algorithm

For NNET algorithm our training accuracy is 72.61%. NNET stands for neural networks which is one of the most efficient algorithms among all of them. As we are getting the values of our dataset from the retinal images we are using neural networks. From the above graph we see that we get the best result using this algorithm. It gives us 72.61% accuracy for the training set which is highest among all the previous algorithms we used.

So, we used total number of four algorithms KNN, NNET, Random Forest and Support Vector Machine (SVM). Among them NNET gives the best accuracy which is 72.61%.

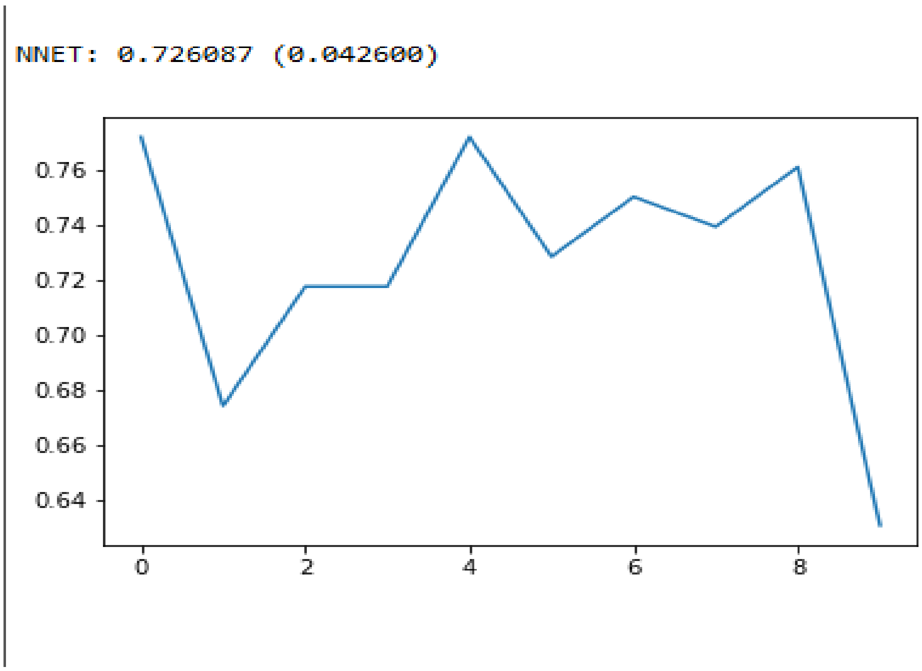

Figure 4.4: Training accuracy of NNET

4.5 Comparison between Algorithms

Figure 4.5 shows a comparison between the algorithms we used for our training dataset. Here, the tall line indicates standard deviation and the rectangular box indicates median value and the brown line in the box indicates the mean value. From here we can understand which algorithm is good for our model.

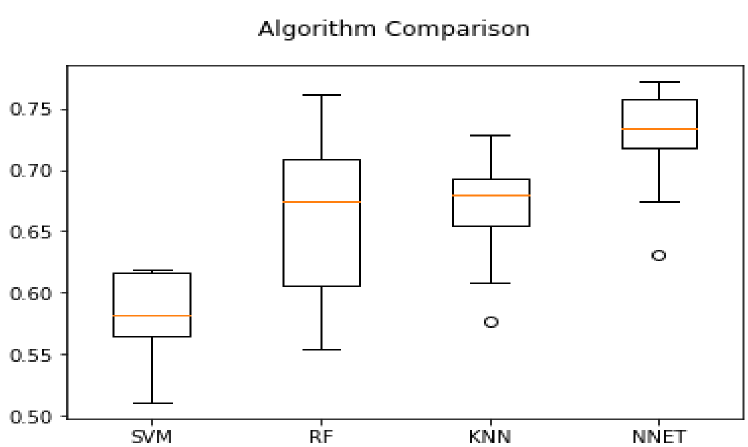

Figure 4.5: Comparison between algorithms

After training the model we test the model with the testing dataset. We have 20% data for testing in the testing set. Table 4.1shows the testing accuracy, precision, recall and F1 score. The detailed information of the test data evaluation with unigram model is as follows-

Table 4.1: Accuracy of test dataset

Models	Accuracy	Precision	Recall	F1 Score
SVM	57.07%	62%	57%	53%
KNN	64.50%	65%	65%	65%
RF	63.63%	64%	64%	64%
NNET	75.32%	78%	75%	75%

In experimental result, we observe that the accuracy of the both training and testing set is quite similar and for both training and testing dataset NNET algorithm is giving higher accuracy rate which is around 75%.So, we can say that this algorithm will give us more accurate prediction about the disease. As our main purpose of the thesis is to build a model which will classify the diabetic retinopathy as accurate as possible, we hope that this final model will give us proper and appropriate results.

We have also determined our train and test model accuracy and loss. For this visualization model we have used keras package for obtaining this train and test -loss and accuracy. We have also used History callback for this purpose. One of the default callbacks that are registered when training all deep learning models is the History callback. It records training metrics for each epoch. This includes the loss and the accuracy (for classification problems) as well as the loss and accuracy for the test dataset, if one is set.

The history object is returned from calls to the fit function used to train the model. Metrics are stored in a dictionary in the history member of the object returned.

Figure 4.6 shows accuracy on the training and test datasets over training epochs.

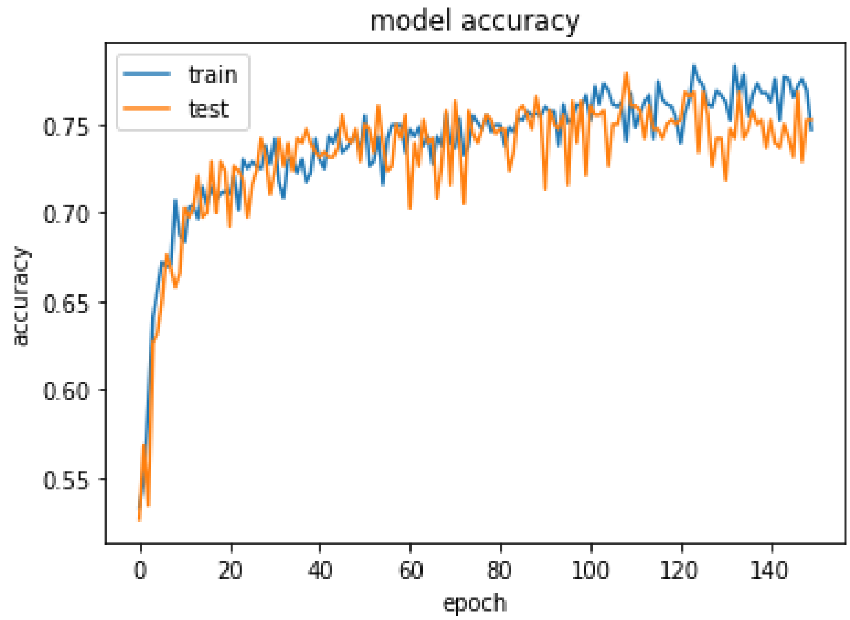

Fig 4.6: Train-Test model accuracy

From the plot of accuracy we can see that the model could probably be trained a little more as the trend for accuracy on both datasets is still rising for the last few epochs. We can also see that the model has not yet over-learned the training dataset, showing comparable skill on both datasets. Figure 4.7 shows a plot of loss on the training and test datasets over training epochs.

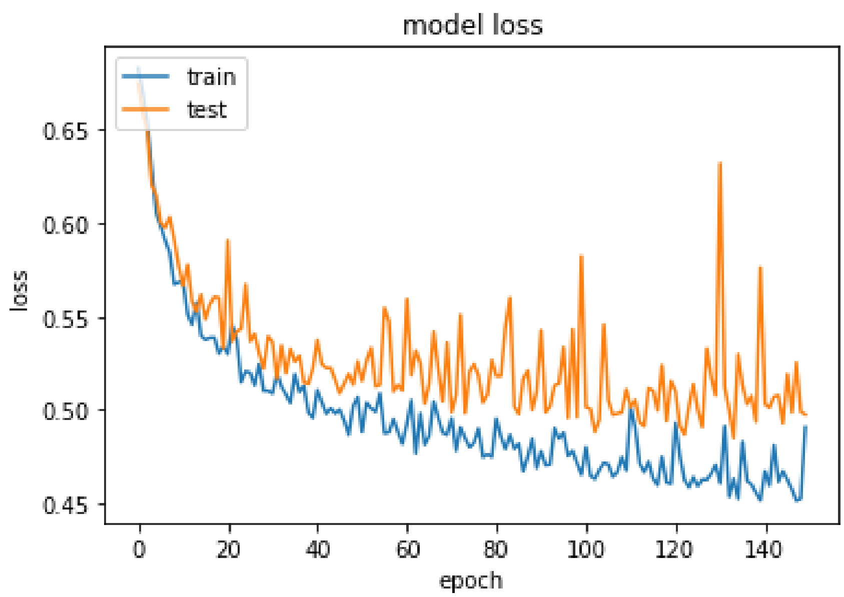

Fig 4.7: Train-Test model loss

From the plot of loss, we can see that the model has comparable performance on both train and test datasets. If these parallel plots start to depart consistently, it might be a sign to stop training at an earlier epoch.

If the lines of train-test loss seem to converge to the same value and are close at the end, then the classifier has high bias. If on the other hand the lines are quite far apart, and then we have a low training set error but high validation error, then your classifier has too high variance.

From these we can conclude that our train-test loss model training set loss is low and our test set error is not too high. So, from this it can be said that we have a good train-test accuracy model.

CHAPTER 5

CONCLUSION

This chapter contains the difficulties, future works and concluding remarks, which will give the summary of our thesis work and also give the indication of our future plan with our thesis project.

5.1 Difficulties

There are many difficulties we faced while working. First of all, there's a lot more to do before an algorithm like this can be used widely. For example, as a classifier algorithm KNN and NNET was remarkable in percentage but we had to work more on binary classification. Secondly, if we could manage more of our training data we could train our algorithm more to achieve more accuracy. Furthermore, we also faced some problems while choosing algorithms. It was quite difficult for us to choose some specific machine learning algorithms that would give accurate classification of the disease. In addition we used simple techniques for feature selection and scaling and possibly we could arrive at better results by introducing more complex techniques for selecting and generating features. We looked at small subspaces for model parameters. Possibility there be other parameter spaces that would yield better performing models. For these few amount of data in our dataset we faced difficulties in implementing the classification.

5.2 Future Work

For any research, there is always room for improvement. Ours is not an exception of that. We have found some areas where this system can be improvised:

1. **Work on more Categories**: This can be improvised with a lot more categorized such as according to ages, genders, background studies, working facilities and so on. As an example, A matured man from the IT background has different eye condition that a matured women from Teaching background.

2**. Work on more classes**: As we working on only two classes whether it is good or bad. In future we are going to add more classes like low, medium, severe condition. In this way patients can know about their condition more accurately

3. **Different Algorithms**: CRF (Conditional Random Field), maximum entropy and other probabilistic graphical model can also be used to train our dataset in order to improve the algorithm.

4. **More Analysis:** To achieve more accuracy we could use more dataset. If we use huge amount of dataset, machine will train more and it would give us more accurate prediction and accuracy.

5. **Hardware Implementation**: A hardware product can be the best solution for patient. So, we are looking forward to build a hardware system where we can use our model to implement results on diabetic patients easily. We can then input the data of the patient and wait for the machine to create a new prescription integrated with Doctor's suggestion.

6. **Software Implementation**: 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.

5.3 Concluding Remarks

We have tried to construct an ensemble to predict if a patient has diabetic retinopathy using features from retinal photos. After training and testing the model the accuracy we get is quite similar. For both sets NNET is providing higher accuracy rate for predicting DR. Despite the shortcomings in reaching good performance results, this work provided a means to make use and test multiple machine learning algorithms and try to arrive to ensemble models that would outperform individual learners. It also allowed exploring a little feature selection, feature generation, parameter selection and ensemble selection problems and experiences the constraints in computation time when looking for possible candidate models in high combinatorial spaces, even for a small dataset as the one used. The structure of our research has been built in such a way that with proper dataset and minor alternation it can work to classify the disease in any number of categories.

References

[1] Gandhi M. and Dhanasekaran R. (2013). Diagnosis of Diabetic Retinopathy Using Morphological Process and SVM Classifier, IEEE International conference on Communication and Signal Processing, India pp: 873-877

[2] Li T, Meindert N, Reinhardt JM, Garvin MK, Abramoff MD (2013) Splat Feature Classification with Application to Retinal Hemorrhage Detection in Fundus Images, IEEE Transactions on Medical Imaging, 32: 364-375

[3] 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

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

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

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

[7] Saimadhu P. How the Random Forest Algorithm Works in Machine Learning. Published on May 22, 2017

[8] Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493–507.

[9] Jason B, Boinee P.“Machine Learning Algorithms”2(3), 138–147. Published on 15, 2016.

[10] Boser B. E, Guyon I. M.,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

[11] Wikipedia.http://en.wikipedia.org/wiki/Support_vector_machine.

[12] Ben-Hur.A, Weston.J (2009) ."A User's Guide to Support Vector Machines". Data Mining Techniques for the Life Science.Humana Press. On Page(s): 223-239.

[13] Akara S. ,BunyaritU.,SarahB.,Tom W.,Khine T. (2009)“Machine learning approach to automatic exudate detection in retinal images from diabetic patients” volume 57-issue 2.

[14] Rajendra Acharya U.,E. Y. K. Ng, Kwan-Hoong Ng, Jasjit S. Suri (2012) “algorithms for the automated detection of diabetic retinopathy using digital fundus images” volume 36, Issue 1, pp 145–157

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

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

[17] 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

[18] 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

[19] Gandhi M and Dhanasekaran R (2013) Diagnosis of Diabetic Retinopathy Using Morphological Process and SVM Classifier, IEEE International conference on Communication and Signal Processing, India pp: 873-877

[20] Rocha A,Carvalho T, Jelinek HF, Goldenstein S, Wainer J (2012) Points of Interest and Visual Dictionaries for Automatic Retinal Lesion Detection. IEEE Transactions on Biomedical Engineering 59: 2244 - 2253.

[21] Lazar I and Hajdu A (2013) Retinal Microaneurysm Detection Through Local Rotating Cross-Section Profile Analysis. IEEE Transactions On Medical Imaging32: 400-407.

[22] Deepak KS, Sivaswamy J (2012) Automatic Assessment of Macular Edema from ColourRetinal Images, Medical Imaging. IEEE Transactions31: 766-776

[23] http://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set

[24] http://archive.ics.uci.edu/ml/machine-learning databases/00329/messidor_features.arff

[25] Breiman. Pasting small votes for classification in large databases and on-line. Machine Learning, 36(1-2):85–103, 1999. (Cited on pages 169, 170, and 187.)

[26] 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(3):556–64.

[27] Cheung N, Mitchell P, Wong TY. Diabetic retinopathyLancet. 2010;376(9735):124–36. doi: 10.1016/S0140-6736(09)62124-3