International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 1160-1168

PREDICTION OF DIABETES USING NEURAL NETWORKS

P. Santhi[1], S.Lavanya[2]
[1]*Associate Professor, Department Of Computer Science and Engineering*
M. Kumarasamy College Of Engineering, Karur.
[2]*Associate Professor, Department Of Computer Science and Engineering*
Muthayammal Engineering College, Rasipuram

Abstract

The disease will produce the increased level of glucose which causes inadequate production of insulin in the body. This disease is called diabetes disorder. This disease is not a fatal disease but sometimes it will cause the serious problem of body parts removal especially legs in the body. This will be similar to fatal cause in the body. The removal of body parts will be done only in the extreme level of diabetes. Its incidence rates are increasing alarmingly every year. These serious issues can be prevented if the prior symptoms of the disease are identified. The dataset of the patient will be collected in the hospital. The dataset will have the entire information about the patient. The information about the patient in the report will have the hemoglobin content, plasma glucose, blood pressure, skin thickness and all other details of the patient. The existing system does not provide the prior intimation to the patients as well as to the doctors regarding the future prediction and serious level of diabetes. The idea we have used here is feature selection methods. The feature selection algorithm which we have selected is deep neural networks, coded on Python, which will gather the particular details regarding the patient and also provide more accuracy in the process of predicting the diabetes in the initial stage itself.At the end, we can provide voice based results for disease diagnosis based on the collected data and also intimate the patients by sending SMS, about the seriousness and the tablets need to take for their issues, to the patients

Keywords: Medical

1. Introduction

1.1. DIABETES

Diabetes mellitus is the disease which is persisting for the long time in the human body. It is otherwise called as chronic disease that occurs when the pancreas is no longer. Pancreas is the gland which is present behind the stomach which is responsible for segregating the digestive enzyme into the duodenum. The term diabetes was first coined by Apollonius around the year of 250 BC after that in the year of 1675 the term mellitus was added to the diabetes by Thomas Willis and it was called as diabetes mellitus. The patients will have the symptoms of tiredness, frequent urination, hungry, increased thirst, slow healing causes damage in the skin. There are two types of diabetes of T-1 and T-2. The people with T-1 don't produce insulin and people with T-2 don't response to insulin and this type of diabetes will not produce enough insulin in the human body. Nearly 95 percent of people were affected by diabetes in the world. In India diabetes affects nearly 62 million people which is more than 7.2 percent of the population. Type-1 diabetes will be common for the people. Metformin is the first medication prescribed for the type-2 diabetes patients. The latest research on diabetes will reduce the segregation of fat build-up in the pancreas and in liver so that the insulin segregation will be more in the pancreas.

1.2. MACHINE LEARNING

ML is a subset of AI which helps the software application to become more accurate in predicting the outcome without any human interaction. It uses the previous historical data to predict the output. Comparison is done between the historical data and the current input data from the user. Machine Learning is a set of algorithms which is used by the computer to predict the outcome of

ISSN: 2005-4238 IJAST

the given input without any explicit instructions. It acts as the statistical tool to perform the specific task. It consists of four types

- Supervised
- Unsupervised
- Semi supervised
- Reinforcement

Supervised learning uses the specific function of input to produce the output and it is a function from labeled data. Unsupervised learning helps the computer to predict the outcome by using pattern matching and the similarities of the data. In Semi supervised learning, the learning of both labeled and unlabeled data is used by using supervising and un supervising methods.

1.3. DEEP LEARNING

Deep learning which uses neural networks between several layers of inputs and outputs. The Neural network algorithm uses different mathematical expressions to produce the output from the certain inputs. The outputs produced are of linearor non-linear relationship.

2. LITERATURE SURVEY

2.1. Survey: Diabetes Analyses for Pregnant Ladies[1]

This Survey says that there are several stages in analyzing diabetes for Pregnant Ladies,

1. Preparation
2. Exploration
3. Cleaning
4. Model Selection

2.1.1. Stage 1: Data Preparation

In this survey, the own dataset is created,instead of using the already existing dataset called the Pima India Diabetes Dataset which was given by the unique client identifier machine learning repository .

2.1.2. Stage 2: Data Exploration

The Pregnancy data set was collected to analyze the particular data of the patient to predict the diabetes. The dimensions of the data set were calculated by using the Panda Data Frame. In that the Shape attribute has been used. From this data set the prediction of diabetes for the pregnant ladies has been analyzed. From the result if the column is one the patient is with diabetes or if the patient is with result of column is zero then the patient is not with the diabetes.

2.1.3. Stage 3: Data Cleaning

In the process of data cleaning they have used the Better Data Beats Fancier Algorithm which have produced the best result. There were some of the factors to be considered in the process of data cleaning.

1. Duplicate values in the dataset
2. Bad labeling in the dataset
3. Missing value or null data point
4. Unexpected outliers

In all these factors the unexpected outlier is the most important one because in the dataset the value for the blood pressure of the patients is zero. If the living person's blood pressure is of zero then the data will be considered under wrong analysis because person cannot have a diastolic BP of zero level. In the analyses of the same dataset the plasma glucose level was zero for the patient and the skin thickness for the normal patient will not be less than 10mm but the analyses for the dataset will have the skin thickness as zero. In the rare cases the insulin for the patient will be zero but in the analyses of the dataset it results as zero.

2.1.4. Stage 4: Model Selection

The important stage in the data analyses is that algorithm selection. They have used totally of seven classifier of KNN, Random Forest, SVM, Gaussian NB, LR, Gradient Boost and DT. Among these seven algorithms they have chosen the best algorithm of Logistic Regression. In the logistic algorithm they have achieved the accuracy of 77.64 percent. This algorithm was considered as the vital concept for the next phase.
The accuracy of the above mentioned algorithm from this survey is,

1. KNN - 0.71 -71%
2. SVM - 0.65 - 65%
3. LR - 0.77 - 77%
4. DT - 0.68 - 68%
5. GNB - 0.75 - 75%
6. RF - 0.74 - 74%
7. GB - 0.76 - 76%

From the analyses the more and best accuracy for the algorithm was considered as Logistic Regression.

2.2. Survey: Hyperglycemia in Pregnancy (HIP) [3]

The Main objective of this analyzes is to reduce the complication of pregnancy.
Hyperglycemia is one of the most complications in the pregnancy. This analyzes was done to reduce the complication of pregnancy in the upcoming year of 2030 and 2045.

2.2.1 Methods

The international diabetes federation (IDF) had used many methods to reduce the complication in the pregnancy which is projected in the year of 2030 and 2045.

1. Carrying the age adjusted prevalence rates in the year by SVM as mentioned in Figure 2
2. Applying the Linear Regression to the past four edition of the IDF as mentioned in Fig. 3
3. Applying the Linear Regression to the previous edition of the IDF with the most consistent trends followed by the extrapolation as mentioned in Figure 4

Hyperglycemia is one of the metabolic changes during the pregnancy. Hyperglycemia in pregnancy (HIP) was defined by the world health organization. The WHO had described the HIP, diabetes first detected at any time during the time of pregnancy. It was defined as pre-existing diabetes and it was further classified into two types.There are two types of Hyperglycemia. They are as follows,

1. Diabetes in pregnancy
2. Gestational diabetes mellitus

2.3. Survey-3: 52 week observational study using the four inhibitors [4]

In this observational study the effectiveness of the two distinct inhibitors is compared and results were discussed in the table 4. The two inhibitors are in the following,

1. Sodium glucose co-transporter 2(SGLT-2)
2. Empagliflozin and Dapagliflozin

The second inhibitor performs as the oral anti diabetic agents. These inhibitors were the controlling remedy for the type-2 diabetic in patient.

2.3.1 Methods

The observational study was first done the patient with Glycated Hemoglobin (HbA1c). The Glycated Hemoglobin is the content of glucose in the blood. The presence of glucose in the red blood cells is Glycated Hemoglobin. This presence of glucose will be there for about 120 days or for about 3 months. This glucose in red blood cells will be in the range of 7.2 to 12.0 %. It will be present along with the other inhibitors of Metformin, Glimepiride, and Dipeptidyl Peptidase-4. The patients will be classified into two types of categories.
1. Patient with Empaglinflozin (25 mg/day)
2. Patient with Diapaglinflozin (10 mg/day)

The results from these observational studies depend on the changes in the HbA1c, and the fasting plasma glucose (FPG) in the blood.
From the above the Research Design proves that the person with adult age of 18-80 will have the common diabetes of type-2. These persons will have the Glycated Hemoglobin in the range of >= 7.5 to < 12.0 % at the baseline along with the hree inhibitors.

1. etaformin - 2000 mg/Day
2. Glimepiride – 8 mg/Day
3. Dipeptidyl

Peptidase - (local stay for > 12 eeks)

2.3.2 Endpoints Design

The primary endpoints were designated to two things of:

1. HbA1c Mean changes
2. Fasting plasma glucose

The secondary endpoints were focus on the following parameters.

1. Changes in body weight
2. Systolic (SBP)
3. Diastolic blood pressure
4. Lipid profile

The symptoms of the hypoglycemia are sweating the one type of removal of waste water from the body from the skin of the body, tremors the muscle contraction in the body which leads to rhythmic movement in the body or the occurrence of shaking in the hands and in legs, palpitation the excess fasting of heart beat in the patient.
Patient with type-2 with oral anti diabetic drug for 12weeks (n=393). The patients classified into two types of Empagliflozin (n=180) and patient with Dapagliflozin (n=213). In both the types the patients with Empagliflozin shows the greater reduction in the HbA1c with the least production of GUT in the number of three (n=3) along with the volume depletion in the number of two (n=2).
In this observational study the total analyses was done with 362 patients. Some patients with the Empagliflozin (n=180) and some patients with Dapagliflozin (n=182). The analyses were done for 52 weeks. After the weeks, the final outcome result produces the reduction in the HbA1c and FPG. Both the types of patients will have reducing of HbA1c and in FPG in the final results. But, the reduction of Empagliflozin was greater than the Dapagliflozin in the patients. Along with this reduction the patients had the decrease in their blood pressure, body weight and in lipoprotein cholesterol. From these inhibitors the sodium glucose co-transporter-2 (SGLT-2) acts as the effective remedy for the patients with type-2 diabetic. At the same time the Empagliflozin acts as greater remedy for reducing the HbA1c than the Dapagliflozin.

3. Existing System

Nowadays Machine learning algorithms widely used in various fields especially of medical field. In medical field Machine Learning had been used for disease diagnosis and its treatment. Two types of machine algorithms had been used are correlation and associations for finding different diseases. In present days some people are dying because of serious disease called diabetics. This diabetes mellitus had been Predicted and diagnosed at the various stages of the patients and this had becomeone of the challenging factors which is faced by many doctors and hospitals in everywhere.To reduce the seriousness of the diabetes in the patients we have predicted this disease at the initial stage itself of the patients. In this project we have used machine learning algorithms and deep learning algorithms to predict this disease. Many researchers have been developing the software to help the doctors for predicting and treating of many diseases. We have also discussed about the data mining techniques which is used to predict the diabetes disease in advance. The diagnostic system in this process is used to determine the presence of diabetes is not. For this diagnostics system, machine learning algorithms are widely used. We have used machine learning techniques in the medical field because of its characteristics i.e., high performance to deal with missing data, irrelevant data and noisy data, and the ability to explain decisions. As everyone is using more date everywhere and there will be a need for some

classifier to classify the newly generated data. The classifiers are used for its accuracy and efficiency. The already existing system had mainly focused on the supervised learning technique called the Random forests by changing the values of different parameters.

3.1. EXISTING BLOCK DIAGRAM

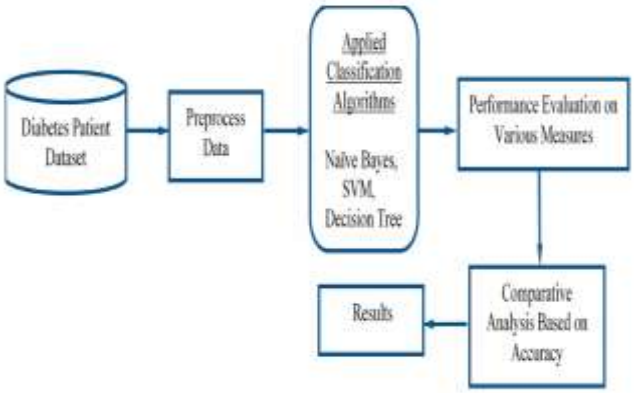

Figure 1 – Existing Block Diagram

4. PROBLEM IDENTIFIED

From all these survey, problem occurred both in T-1 and in T-2. In T-1 diabetic, the patients will have beta cells destruction in the pancreas. This will lead to the insulin deficiency in the body. In T-2 diabetic, the patients will have progressive damage, dysfunction and various failures of organs including the kidney, nerves, eyes, blood vessels. This occurrence of risk in patient is due to poor prior notification and proper treatment for their diabetes.

- Labeled data based disease classification
- Provide high number of false positive
- Binary classification can be occurred
- Computational complexity
- Prior notification

5. PROPOSED SYSTEM

An ML technique provides multiple opportunities for predicting the diseases in the medical fields. The chronic diseases like heart disorder, other infectious diseases can be predicted by using already used models of ML. Several researches had been used in machine learning models to diagnosis and predict the non-communicable diseases, which have more advantage in the medical field. Upcoming researchers have been working on deep learning models to predict specific disease especially of diabetes in the patient more effective in the prevention of the diabetes diseases. So this hospitalization of patients will get reduce. This will help the medical organization by providing more beneficial transformation. Diabetes is a chronic disease which reduces the insulin level and increases the glucose level in the body. The insulin hormone does not produce adequate insulin in the diabetes patient's body.This leads to increased glucose level and false segregation and metabolism of the carbohydrates in the body. To prevent this initial prediction of diabetes plays a vital role. This disease causes increased level of glucose in the body. Glucose will get generated in the body after eating food. The Insulin hormone helps to maintain the balancedstate of glucose level and the blood sugar level in the body. The patients with T-1 diabetes don't have sufficient insulin production to balance the sugar level whereas patients with T-2 don't have proper respond to the insulin segregation. Both T-1 and T-2 will leads to increased blood sugar level. T-1 diabetes is most common one. Some certain cases called pre-diabetes. The patients who are all suffers from pre-diabetes will have high glucose level in their blood but theywere not considered under diabetes patients. But the person who suffers from

this disease can have the possible attack of T-2 diabetes. This pre-diabetic disease can cause serious damagesin the major body organs of kidneys, eyes, legs heart and nerves. If the pregnant ladies suffer from pre-diabetes then the disease is called gestational diabetes. So implement deep learning based neural network algorithm can be used to predict the diabetic diseases with improved accuracy. Neural Networks has emerged as an important method of classification. Back-propagation has been employed as the training algorithm in this work. This project proposes a diagnostic system for predicting heart disease with improved accuracy. The propagation algorithm has been repeated until minimum error rate was observed. And also provide the diagnosis information to patients through SMS alert and also provide voice information to patients based on trained diabetic datasets.

5.1 ADVANTAGES

- Accuracy is high
- Parallel processing
- Multiple diabetic diseases are predicted
- Reduce number of false positive rate
- Prior notification to the patients

5.2 PROPOSED BLOCK DIAGRAM

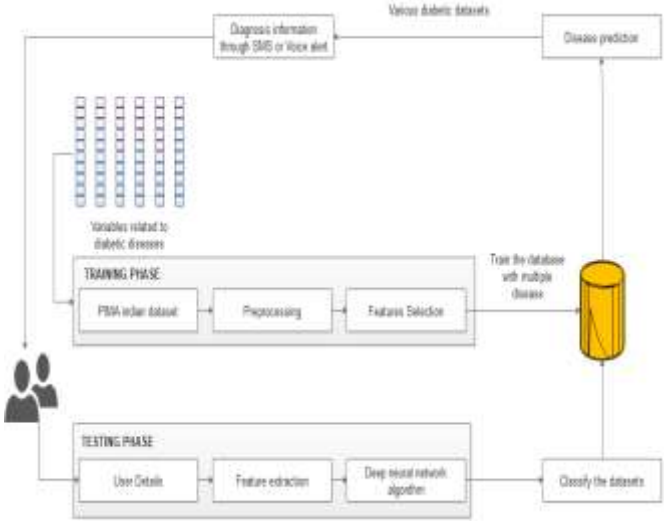

Figure 2 – Proposed Block Diagram

5.3 MODULES DESCRIPTION

- Datasets Acquisition
- Preprocessing
- Features Selection
- Classification
- Disease diagnosis

5.3.1 DATASETS ACQUISITION

Collection of datais called as dataset. Usually a dataset contains the information in the form of tables and statistical data matrix, where every column of the table represents a variable and every row represent the particular member of the data set. The variables represented in the dataset produces the information about the height and weight of an object or particular person. Every value represented in the dataset called datum. The data set refers to collection of data which are closely related to the particular events. In this module, we can upload the cardiovascular datasets related to diabetic diseases which includes the attributes such as glucose, insulin level, blood pressure, BMI and so on.

5.3.2 PREPROCESSING

Removing the missing data and irrelevant data is an important step in the process. This is called preprocessing. The two phrases are used in data mining and machine learning techniques are "garbage in and garbage out". The results produced are in out of range and several combinations of missing values are often produced under data gathering methods.If the data analysis is not done properly then it leads to problems which cause major issues in the medical field. To avoid this major issues the representation and quality of data should be analyzed carefully. If the dataset contains more irrelevant and redundant information, it will cause more difficulties during the training phase. During preprocessing the considerable amount of time will be taken by two processes of data preparation and data filtering. In this module, we can eliminate the irrelevant values and also estimate the missing values of data. Finally provide structured datasets.

5.3.3 FEATURES SELECTION

The major process consider during feature selection is to reduce the inputs for processing and analyzing. The useful information or features from already existing data will be extracted by the process of feature engineering. The statisticalmeasuresare applied to filter the feature selection according to each feature. The features are removed from the dataset and are prioritized by their ranks. The methods used in feature selection are independent and dependent variables according to the variable. It can be used to construct the multiple diabetic diseases. In this module, select the multiple features from uploaded datasets. And train the datasets with various disease labels such as T-1 diabetics with diagnosis information. T-2 diabetics with diagnosis information

5.3.4 CLASSIFICATION

In this module implement classification algorithm to predict the diabetic diseases. To predict the disease we have used one of the deep learning algorithms such as back propagation. The ANN provides back propagation along with feed forward will be assigned to sets of input data to produce the outputs. The directed graphs will have multiple layers of nodesand each node will be connected to each other. Each and every node is considered as a neuron with a function of nonlinear activation except for the input nodes. Supervised learning technique uses the back propagation methods for network called back propagation for training the network. Back propagationdistinguishes the data which is not linearly separable and it is entirely modified form of linear perceptron. If back propagationuses some mechanism i.e., if all the neurons contains linear activation function,to determine whether the neurons fires or not, then it canbe easily reducible with linear algebra withmany number of layers into twostandard layers of input-output model. To optimizethe methods to adjust the heavy weights by reducing the loss function in the network. This can be done by using gradient techniques. To compute this gradient technique of loss function the algorithm must requires anappropriate output for all types of inputs. Usually, the delta rule is used to produce the gradient for every layer is used along with feed forward networks.The activation function for different neurons can be found by using back propagation algorithms. Back Propagation algorithms are currently implemented as the major concepts on the researches of parallel and distributed computing. This Back Propagation algorithm plays a vital role in the pattern recognition domains. These algorithms are used in solving complex problems and produces appropriate results even when the complex predictions. The architecture of feed forward back propagation for supervised training is same as the architecture of back propagation in the neural network. The back propagation is the most known and most frequently used type of neural network. User can provide the features and automatically predict the diseases.

5.3.5 DISEASE DIAGNOSIS

The disease diagnosis done by the physicians and other health specialists is done by the decision making tasks. Medical decision support system is a set of decision making support program. In this module, provide the diagnosis information based on predicted diabetic diseases. Proposed system provides improved accuracy in diabetic disease prediction. Some of the conditions or habits are considered as the risk factors of the patients which make them to develop the disease. In this module, provide the diagnosis information based on predicted diabetics diseases.

Information sends to user in the form ofSMS or Voice information. Proposed systems provide improved accuracy in heart disease prediction.

5.4 DATASET USED FOR THIS PROJECT

Dataset used for this project is Indian PIMA Dataset for diabetes.

Table 1 Indian Pima Dataset For Diabetes

>>>	Pregnancies	Glucose	...	Age	Outcome
0	6	148	...	50	1
1	1	85	...	31	0
2	8	183	...	32	1
3	1	89	...	21	0
4	0	137	...	33	1
..	...	...	...	...	...
763	10	101	...	63	0
764	2	122	...	27	0
765	5	121	...	30	0
766	1	126	...	47	1
767	1	93	...	23	0

6. PERFORMANCE ANALYSIS

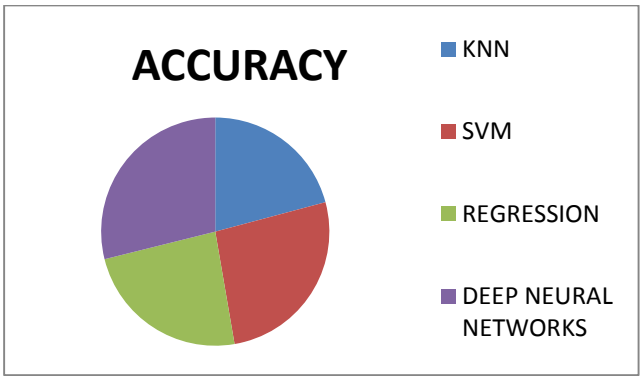

Figure 3 Performance Chart

6. CONCLUSION

In this project certain different algorithms are considered to face the problems of data mining techniques and also used in the medical field for disease prediction. The major focus is on combinations of different algorithms for the effectiveness and accuracy in the diabetes prediction using data mining. The extraction of medical data from the valuable medical rules is done by data mining techniques. This technique is also used for disease prediction and clinical diagnosis. There is an increasing interest in using classification to identify disease which is present or not. In this model target samples are taken from the patients of hospitalization.These samples were considered under many classifications. Classification algorithm is very sensitive to noisy data. The process of classification will be so critical and difficult in the presence of noisy data. This process reduces the performances and task of classification.The noisy attributes should be removed earlier before applying the classification algorithm in the datasets. In this project work, we can implement preprocessing steps and implemented the classification rule algorithms namely deep neural network algorithm are used for classifying datasets which are used by the users. By the experimental analyzes the deep learning technique provides the better results and accuracy than the other techniques.

REFERENCES

1. Lahiru Liyanapathirana, **Machine learning workflow in diabetes in the article of towards data science**, A Medium publication Feb-26, 2018.

2. Norbert Freinkel, **Classification and diagnosis of diabetes: Standards of medical care in diabetes**, The Journal of Clinical and Applied Research and Education, Volume 41, Supplement 1, January 2018.
3. Lili Yuen, PouyaSaeedi, MusarratRiaz, **Projection of prevalence of Hyperglycaemia in pregnancy in 2019 and beyond**. International Diabetes Federation Diabetes Atlas,Volume 157, 107841, November 01, 2019.
4. EuJeong Ku, **Empagliflozin versus Dapagliflozin in patients with type-2 diabetes inadequately controlled with metformin, glimepiride and dipeptidyl peptide 4 inhibitors,** Volume 151, P65-73, and May 01, 2019.
5. Ayesha A. Motala, Jonathan E. Shaw, **Global and regional diabetes prevalence estimates for 2019 and projection for 2030 and 2045**, International Diabetes Federation Diabetes Atlas, Volume 157, 107843, November 01, 2019.
6. P.Santhi, **Implementation Of Classification System Using Density Clustering Based Gray Level Co Occurrence Matrix (DGLCM) For Green Bio Technology**, International Journal of Pure and Applied Mathematics,Vol.118, No.8,PP. 191-195,2018.
7. S. Thilagamani, **Novel Recursive Clustering Algorithm for Image Oversegmentation**, European Journal of Scientific Research, Vol.52, No.3, pp.430-436, 2011.
8. Mohammed Akour1 , Osama Al Qasem2 , Hiba Alsghaier3 , Khalid Al-Radaideh4, **The Effectiveness of Using Deep Learning Algorithms in Predicting Daily Activities**, Advanced Trends in Computer Science and Engineering, pp. 2231- 2235, Volume-8, No.5, September – October 2019.
9. Ahmad al-Qerem1 ArwaAlahmad 2 , **Human Body Poses Recognition Using Neural Networks with Data Augmentation**, Advanced Trends in Computer Science and Engineering, pp. 2117 – 2120,Volume-8, No.5, September – October 2019, ISSN 2278-3091.
10. P.Santhi, G.Mahalakshmi, **Classification of Magnetic Resonance Images Using Eight Directions Gray Level Co-Occurrence Matrix Based Feature Extraction**, International Journal of Engineering and Advanced Technology, ISSN: 2249-8958, Volume-8 Issue-4, April 2019.
11. Dr. Navneet Malik, Nilesh N Wani, Jimmy Singla, **Complications of Sight Threatening Diabetic Retinopathy,** International Journal of Engineering and Advanced Technology, ISSN 2278-3091, Volume 8, No 4, July – August 2019.