ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse13922020.pdf
https://doi.org/10.30534/ijatcse/2020/13922020

Prediction of Diabetes using Neural Networks

P. Santhi [1], N. Deeban [2], N. Jeyapunitha [3], B. Muthukumaran [4], R. Ravikumar [5]

[1] Department of CSE, M.Kumarasamy College of Engineering, India, santhip.cse@mkce.ac.in
[2] Department of CSE, M.Kumarasamy College of Engineering, India, deebankaviraj@gmail.com
[3] Department of CSE, M.Kumarasamy College of Engineering, India, punithanambi98@gmail.com
[4] Department of CSE, M.Kumarasamy College of Engineering, India, hapieeji@gmail.com
[5] Department of CSE, M.Kumarasamy College of Engineering, India, ravi99cse@gmail.com

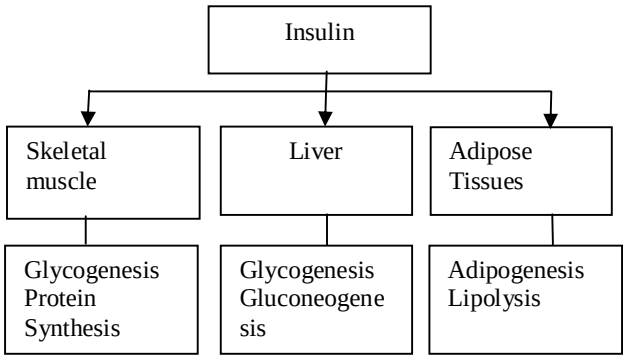
ABSTRACT

The disease will produce the high level of glucose in the blood which leads to inadequate production of insulin in the body. This disease is called diabetes mellitus. 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. 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 major idea of this project is to use the feature selection methods. The feature selection algorithm which we have selected is neural networks, 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.

Key words: Insulin, Logistic Regression, Support Vector Machine, Neural Networks, Empagliflozin, Dapagliflozin

1. INTRODUCTION

Table 1: Insulin Production throughout the parts of the body



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 frequent urination, increased thirst, feeling very tired, always feeling hungry, slow healing of the wounds and patches of dark skin in the body. There are two types of diabetes of type-1 and type-2. The people with type-1 diabetes don't produce insulin and people with type-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 was 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.

2. LITERATURE SURVEY

2.1. Survey: Diabetes Analyses for Pregnant Ladies

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

1. Data Preparation
2. Data Exploration
3. Data 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. This data seems to be wrong because a living person cannot have a diastolic blood pressure of zero. 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 Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Gaussian Naive Bayes, Random Forest And Gradient Boost. 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 prime candidate for the next phase.

The accuracy of the above mentioned algorithm from this survey are,

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.

2.1.5.Dataset

Table 2: Datasets for Pregnant Ladies

Patient	Glucose	Glucose level	Blood pressure level	Skin thickness	Insulin	BMI	Diabetes pedigree	Age of patient	Out Come
0	5	149	71	28	0	32.6	0.638	45	1
1	1	87	64	34	0	24.6	0.347	30	0
2	7	143	64	0	0	22.3	0.681	29	1
3	1	90	76	30	90	21.1	0.214	31	0
4	0	117	45	21	156	40.1	2.355	41	1

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 [Figure1]
2. Applying the Linear Regression to the past four edition of the IDF [Figure2]
3. Applying the Linear Regression to the previous edition of the IDF with the most consistent trends followed by the extrapolation [Figure3]

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

Charts were used to describe the analyses of Hyperglycemia in different years. The process of reducing the Hyperglycemia in the upcoming year of 2030 and in 2045 has discussed.

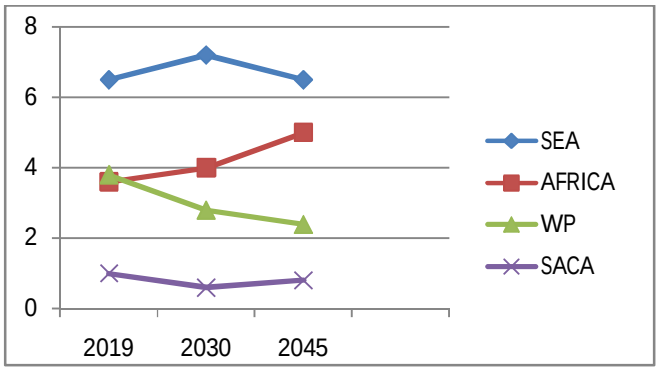


Figure 1: Future Analyses Of Hyperglycemia In 2030, 2045 Method -1 Analyzes

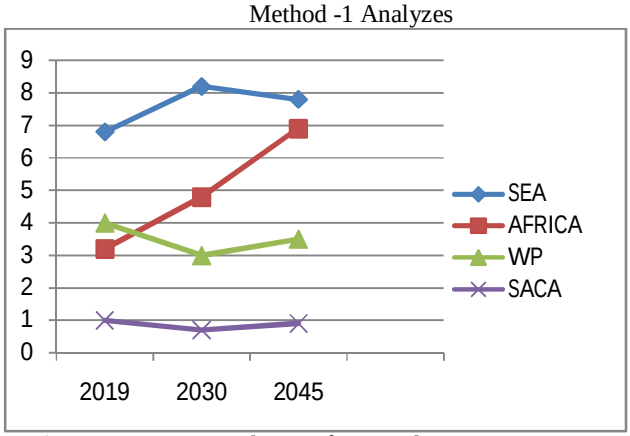


Figure 2: Future Analyses Of Hyperglycemia In 2030, 2045 Method-2 Analyzes

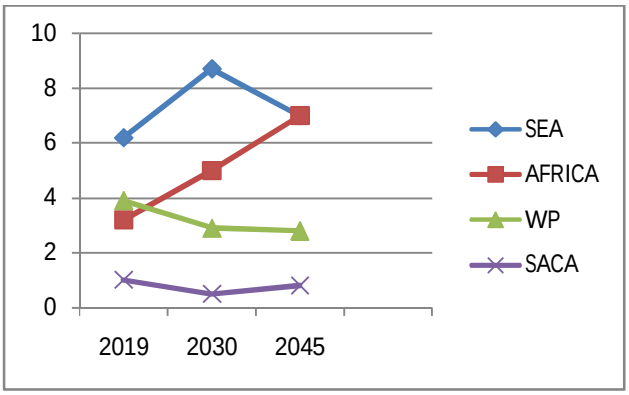


Figure 3: Future Analyzes Of Hyperglycemia In 2030, 2045 Method- 3 Analyzes

2.2.2. Dataset

Table 3: Datasets for Hyperglycemia

COUNTRY	TOTAL BIRTH 2030	HIP IN 2030	AGE ADJUSTED IN %	TOTAL BIRTH 2045	HIP IN 2045	AGE ADJUSTED IN %
SEA	25,645,284	7,487,451	26.44	23,240,547	6,421,410	26.44
AFRICA	38,341,741	4,048,784	10.24	47,471,241	4,745,466	10.22
WP	25,752,654	6,745,471	10.17	24,741,541	2,540,471	10.18
SACA	64,471,265	7,241,476	10.46	48,596,415	3,462,189	1.44
OVERALL PREVALANCE	1,30,594,412	18,456,546	14.00	135,750,047	17,993,475	13.25

From the above survey, the results for analyzes of Hyperglycemia in the year of 2030 and in 2045 are shown as percentages in the below table 4.

Table 4: Accuracy for Each Methods

METHODS	2030 (IN %)	2045 (IN %)
METHOD ONE	12	13.4
METHOD TWO	16.2	18.7
METHOD THREE	16.0	15.4

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. 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 three inhibitors.

1. Metaformin - 2000 mg/Day
2. Glimepiride – 8 mg/Day
3. Dipeptidyl
   Peptidase - (local stay for > 12 weeks)

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).s

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.

2.3.3 Flowchart

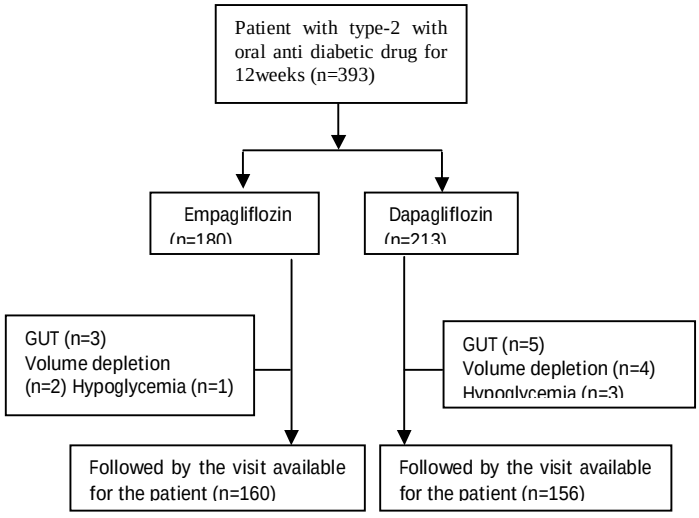


Figure 4: Flowchart for Inhibitors

3. COMPARISON TABLE

Table 5: Comparison Tables for Algorithms

S.No	Method	Accuracy in (%)
1	K-Nearest neighbor	71
	Support Vector Machine	65
	Logistic Regression	77
	Decision Tree	68
	Gaussian Naive Bayes	75
	Random Forest	74
	Gradient Boost	76
2	Support Vector Machine	65
	Linear Regression	72
	Linear Regression	68
3	Sodium glucose co-transporter 2(SGLT-2)	65
	Empagliflozin and Dapagliflozin	60

4. PROBLEM STATEMENT

From all these survey, problem occurred both in type-1 and in type-2. In type-1 diabetic, the patients will have beta cells destruction in the pancreas. This will lead to the insulin deficiency in the body. In type-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.

5. CONCLUSION

Predictive analytics which is help to healthcare organizations to evaluate data on the past behavior and predict likelihood of future behavior to enable better decisions and outcomes of their patient. Predictive models can make human decisions more effective and highly automate an entire decision-making process. Diabetes can be a chronic condition that is brought on by awkwardness inside the discharge of agent transportation regarding associate unsettling influence inside the sugar levels of the blood. Big Data mining is process of extracting hidden knowledge from large volumes of raw data. Big Data mining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans Disease Prediction plays an important role in data mining. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. Data Mining is used intensively in the field of medicine to predict diabetics which is defined as any degree of glucose intolerance with onset or first recognition. This project analyzes the diabetic's predictions using different classification algorithms. Medicinal data mining has high potential for exploring the unknown patterns in the data sets of medical domain. These patterns can be used for medical analysis in raw medical data using deep learning algorithm and number of experiment has been conducted in Python tool for analyze the large datasets in real time environment

REFERENCES

1. Lahiru Liyanapathirana, **Machine learning workflow in diabetes in the article of towards data science**, A Medium publication sharing concepts, ideas, and codes, Feb-26, 2018.
2. Norbert Freinkel, Diabetes Care,**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, Pouya Saeedi, Musarrat Riaz, **Projection of prevalence of Hyperglycaemia in pregnancy in 2019 and beyond**: results from the International Diabetes Federation Diabetes Atlas ,9th edition: Diabetes research and clinical practice, Volume 157, 107841, November 01, 2019. https://doi.org/10.1016/j.diabres.2019.107841
4. Eu Jeong Ku, Dong-Hwa Lee, Hyun Jeong Jeon, Tae Keun Oh, **Empagliflozin versus Dapagliflozin in patients with type-2 diabetes inadequately controlled with metformin, glimepiride and dipeptidyl peptide 4 inhibitors,** Volume 151, P65-73, May 01, 2019. https://doi.org/10.1016/j.diabres.2019.04.008
5. Ayesha A. Motala, Jonathan E. Shaw, **Global and regional diabetes prevalence estimates for 2019 and projection for 2030 and 2045**: Results from the International Diabetes Federation Diabetes Atlas, 9th edition, Volume 157, 107843, November 01, 2019. https://doi.org/10.1016/j.diabres.2019.107843
6. P.Santhi, R.Vikram, **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, N.Shanthi, **A 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** , of Advanced Trends in Computer Science and Enginee, pp. 2231- 2235, Volume-8, No.5, September – October 2019. https://doi.org/10.30534/ijatcse/2019/57852019

9. Ahmad al-Qerem1 Arwa Alahmad 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. https://doi.org/10.30534/ijatcse/2019/40852019
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. P. Pandiaraja, N Deepa 2019 , **A Novel Data Privacy-Preserving Protocol for Multi-data Users by using genetic algorithm** , Journal of Soft Computing , Springer , Volume 23 ,Issue 18, Pages 8539-8553.
12. K Sumathi, P Pandiaraja , **Dynamic alternate buffer switching and congestion control in wireless multimedia sensor networks** , Journal of Peer-to-Peer Networking and Applications , Springer.
13. P.RajeshKanna and P.Pandiaraja 2019, **An Efficient Sentiment Analysis Approach for Product Review using Turney Algorithm** , Journal of Procedia Computer Science , Elsevier ,Vol 165 ,Issue 2019, Pages 356-362 https://doi.org/10.1016/j.procs.2020.01.038