Effective Prediction System for the Growth of Diabetes Using Machine Learning Algorithms with Linear Regression Method

Jamuna.S[1], Mohan Kumar.K[2]
[1]*Research Scholar,* [2]*Research Supervisor & Head*
PG and Research Dept. of Computer Science,
Rajah Serfoji Government College (Autonomous), Thanjavur 613005,
Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.
[1]sjamunacs@gmail.com,
[1]tnjmohankumar@gmail.com

Abstract: **The chronic disease diabetes affects a large number of people worldwide and becomes a hassle to the entire world as it causes severe health problems. Accurate prediction of any disease is the first step of prevention or control the growth of it. Data mining techniques make this complicated work easier. The proposed diabetes prediction system which uses the linear regression algorithm will exactly forecast the growth of diabetic patients. Apart from medications some herbal plants play an active role in controlling the diabetes. This prediction system will also indicate the use of medicinal plants as food supplement.**

Keywords **— Prediction, Diabetes, Linear regression, Clustering, Fuzzy-C Means.**

I. INTRODUCTION

Diabetes is being one of the most dangerous chronic diseases affecting millions of people globally. The state of increased blood glucose level is termed as diabetes which causes many serious health problems quietly. It is mainly classified into three types namely Type 1 Diabetes, Type 2 Diabetes and Gestational Diabetes. The hormone insulin secreted in pancreas is responsible for moving the glucose from the bloodstream into the cells for using as energy. Due to some physiological conditions the pancreas fails to synthesize adequate insulin for the body requirement (Type 1 Diabetes) or the body cells do not respond to the synthesized insulin (Type 2 Diabetes) or hormonal changes during pregnancy (Gestational Diabetes) raises the glucose level floating in the bloodstream leading to diabetes [1].

International Diabetes Federation (IDF) provided some shocking information about diabetes in IDF Diabetes Atlas ninth edition 2019. In that IDF indicated that approximately 463 million of adult populations between the age group of 20-79 years are affected by diabetes worldwide. The prevalence of Type 2 diabetes is increasing rapidly in almost all countries of the world and 374 million people are at the risk of getting Type 2 diabetes. One fifth of the world population above 65 years old have been affected by diabetes and 79% of adults with diabetes belong to low and middle income countries. Moreover the complications of diabetes caused 4.2 million deaths in 2019 and 760 billion US Dollars was required for the health expenditure of diabetes. From this report it is clear that diabetes has become a huge burden to the whole world. As there is no way for curing diabetes, it should be prevented or controlled for the benefit of the entire human society [2].

Early prediction will be the best solution for preventing or controlling many diseases including diabetes. Accurate prediction of diabetes is a rigorous work but it can be done properly using data mining techniques. Data mining is the combination of machine learning, statistics and database systems which discovers the hidden patterns evidently in the large voluminous datasets and transforming the data into useful structures for future use. These structures can be used as input tools for machine learning and predictive analytics. Linear regression is a renowned statistical approach of machine learning technique widely used for predictive analysis due to its simplicity and more accuracy. It is more evident in finding the mathematical relationship between the input and output variables [3].

In the field of health care systems, science and technology brings a lot of latest inventions for the treatment of diseases, but they become vain or too expensive if the diseases are diagnosed at a critical stage. Here the role of prediction systems is recognized as more important to overcome many health and economic issues. This paper deals with the diabetes prediction system which forecasts the estimated prevalence of diabetes efficiently for various years. The proposed prediction system uses PHP and MYSQL codes and the datasets collected for a period of five years form a disease diagnostic centre. This work uses the linear regression algorithm to predict the growth of diabetes which will help the doctors and healthcare professionals to take necessary steps to control the rapid growth of diabetes. This paper also recommends some herbal remedies as preventive measure and for managing the blood glucose level that will be very useful to the people to protect themselves from the dangerous disease.

II. RELATED WORKS

There are many research work have been done regarding the prediction of various chronic diseases using machine learning algorithms. Desmond Bala Bisandu et al.,(2019) designed a prediction system for diabetes using Java programming language,

Weka tool and MySQL. They applied Naïve Bayes classifier approach for the diagnosis of diabetes and they found that the developed prediction system provided the result with 95% accuracy rate [4].

Deepti Sisodia et al., (2018) had chosen Decision Tree, Support Vector Machine (SVM) and Naïve Bayes machine learning approaches for designing diabetes prediction system. They used Pima Indian Diabetes Dataset (PIDD) collected from UCI repository for testing the efficiency of the three machine learning algorithms and showed that Naïve Bayes algorithm gave the best result with the accuracy rate of 76.30% [5].

Smriti Mukesh Singh et al.,(2018) collected both structured and unstructured data from a hospital for their work of improving disease prediction by machine learning and stored them in a conventional neural network based multimodal disease risk prediction (CNNMDRP) algorithm. Then they used these datasets for the prediction of chronic diseases using Naïve Bayes and SVM classifiers. Genetic algorithm was used to retrieve the missing values of unstructured data. They had also used Community Question Answering (CQA) system for improving the accuracy of the prediction system [6].

Kedar Pingale et al.,(2019) designed a system for prediction of various diseases on symptoms basis provided by the patients. They collected real life datasets of various diseases from a hospital and applied K Nearest Neighbour (KNN), Naïve Bayes and Logistic Regression for the prediction of diseases. The symptoms provided by the patients play the active role in their prediction system [7].

Vijayarani et al.,(2013) made a survey research of various data mining algorithms used for the prediction of heart diseases, breast cancer and diabetes. They deeply analysed the issues and challenges of different algorithms like ANN, SVM, Naïve Bayes, KNN, etc. in medical field for the prediction of these three diseases [8].

Mohan Kumar KN et al.,(2019) had insisted in their work about the importance of disease prediction systems for the prevention of wide spreading diseases all over the world. They had listed 23 machine learning techniques including linear regression, clustering, SVM, etc. and 14 types of datasets which are widely used for disease prediction. They had also mentioned a list of features consisting of age, gender, blood pressure, blood glucose level, etc. for the effective prediction of diseases [9].

III. METHODOLOGY

The diabetes datasets consisting of ten attributes have been collected for this work from a disease diagnostic centre for a period of five years from 2015 to 2019. The dataset contains the records of various age groups and gender who came to that centre at a stipulated period in every year for testing diabetes. The following Table I shows the diabetes dataset collected for the year 2019 from the centre which contains the details of 325 patients.

TABLE I
DIABETES DATASET

Patient's ID	Gender (G)	Age	Blood pressure	Glucose (GLS)	Skin thickness	Serum insulin	Body mass index weight/(height)2	Pregnancies
1	1	50	72	85	35	0	33.6	0
2	2	31	66	92	29	0	26.6	2
3	3	32	64	86	0	0	23.3	0
4	1	21	66	100	23	94	28.1	0
5	2	33	40	94	35	168	43.1	4
6	1	30	74	82	0	0	25.6	0
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
323	1	57	80	95	0	0	27.1	0
324	1	59	60	96	23	846	30.1	0
325	2	51	72	91	19	175	25.8	3

Linear Regression is used in this prediction system to forecast the growth of diabetes as it is easy to work and provides accurate result by comparing the dependent and independent variables statistically. It is also more reliable than many machine learning algorithms. The following Fig.1 showing the formula [10] of linear regression is applied for this work.

Y = a*X +b

Here X = Proposed year for prediction – Starting year in the sample data set

a= (n*sum(x*y)-sum(x)*sum(y)) / (n*sum(sqr of x)- sum(x) * sum(x))

where n = number of years the sample data set available

x= 0 to n-1

y = actual existing data available for the range x years

b = (1/n)*(sum(y)-(a*sum(x))

Fig.1 Linear Regression Formula

The following Fig. 2 clearly shows the prediction model of diabetes used for forecasting the number of diabetic patients.

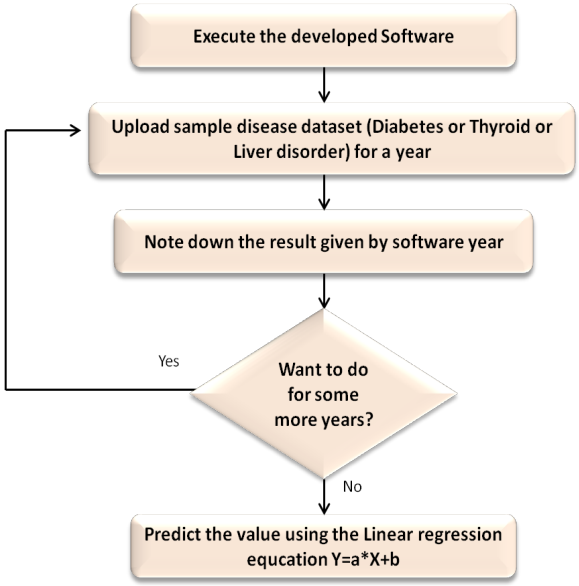

Fig. 2 Diabetes Prediction Model

IV. RESULT AND DISCUSSION

The following Fig. 3 shows the screen shot of the data uploading page in the developed software. It will be obtained by clicking the option "*Bulk Patient Upload*". This page is used to feed the dataset for any number of years. The choose file button in the page is used to select the file in the folder which contains the data file. The button "Import" is used to upload the file for processing.

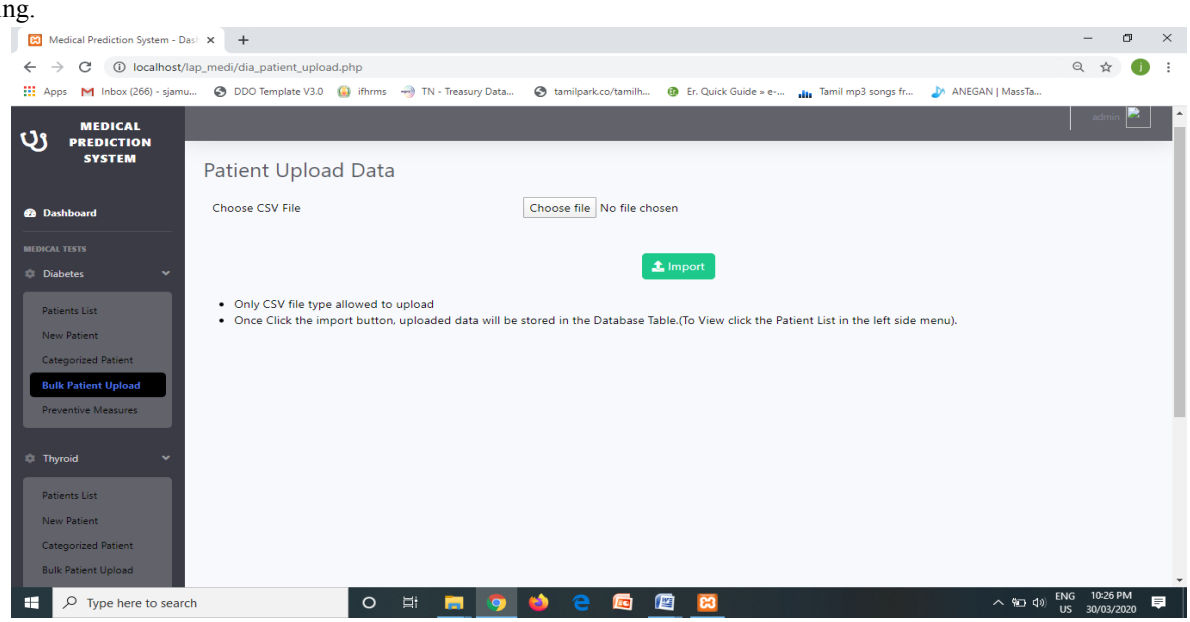

Fig. 3 Data Uploading Option Window for Diabetes

The button "*Patient list*" in the dashboard is used to list the entire detail about patients. The following Fig. 4 shows the

complete dataset which are uploaded for the year 2019.

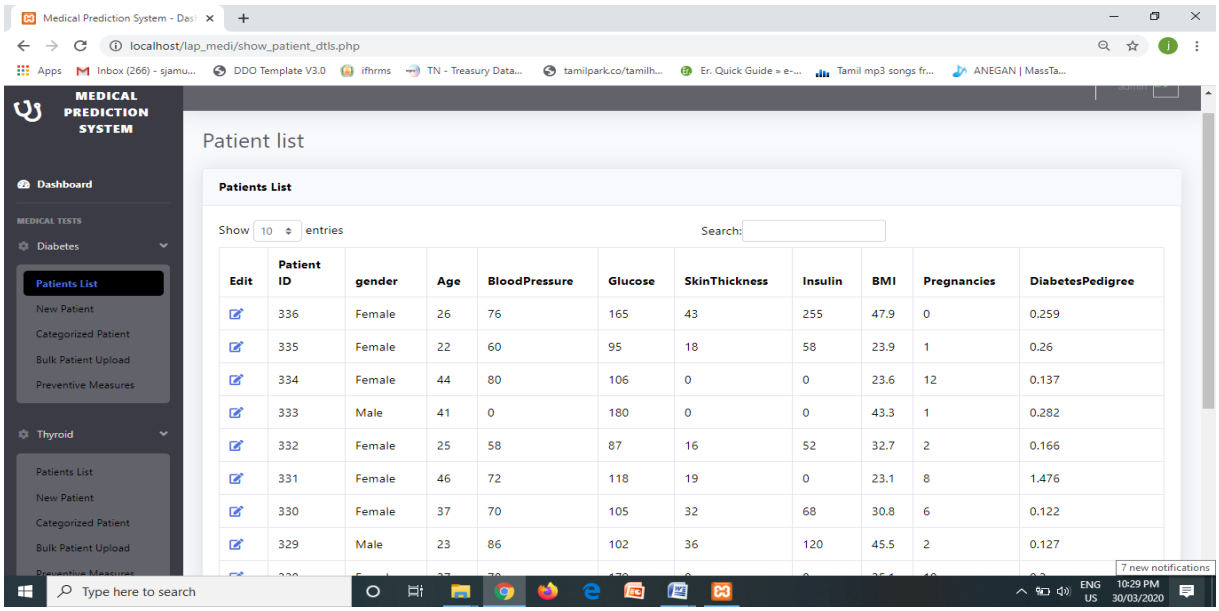

Fig. 4 Uploaded Dataset Window for Diabetes

The button "*Categorize Patient*" in the dashboard is used to classify the data into three categories namely No Diabetes, Pre Diabetes and Diabetes. The following Fig. 5 shows how the classification done for the year 2019.

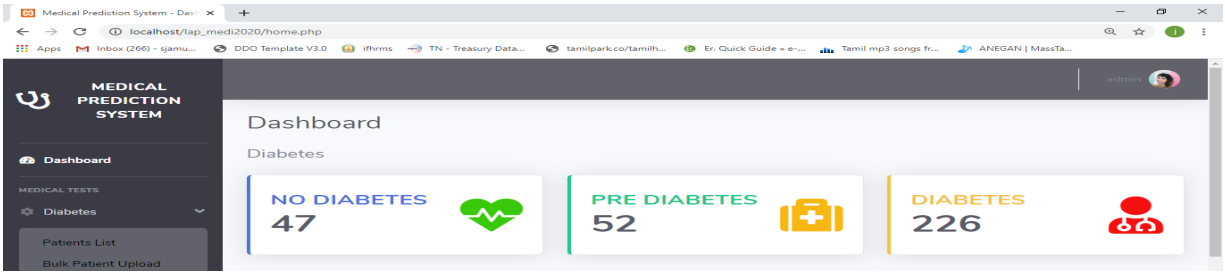

Fig. 5 Result of Diabetic Patients

Similarly the results of various years from 2015 to 2019 are obtained by using the developed software. The tabular form for those results is depicted as in the following Table II.

TABLE II
Result of Diabetic Patients (sample size: 325)

Years	**2015**	**2016**	**2017**	**2018**	**2019**
Patients	**165**	**179**	**190**	**201**	**226**

The values in the Table II are applied into linear regression method and the following Table III is arrived.

TABLE III
Linear Regression parameters

X	**Y**	**X*Y**	**SQR OF X**
0	165	0	0
1	179	179	1
2	190	380	4
3	201	603	9
4	226	904	16
10	961	2066	30

Here,

The value of a= (5*(2066)-(10*961))/((5*30)-(10*10)) =14.4

The value of b= (1/5)*(961-(14.4*10))=163.4

The number of patients in the year 2020:

Y = a*x+b

Y= (14.4*5) +163.4 = 235.4

So, the expected number of patients in the year 2020 is 235.4. Similarly for the years 2021 to 2025 are calculated and tabulated as in the following Table IV.

TABLE IV

Predictive result of Diabetic Patients (sample size: 325)

Years	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Patients	165	179	190	201	226	235.4	249.8	264.2	278.6	293	307.4

From the above table it is vividly shown that the growth of diabetic patients is steadily increasing. The table IV also shows that 307.4 of 325 individuals will be affected by diabetes in 2025 and it is almost 95% of the total sample taken for this work. The following Fig. 6 clearly shows the linear growth of diabetic patients from 2015 to 2025.

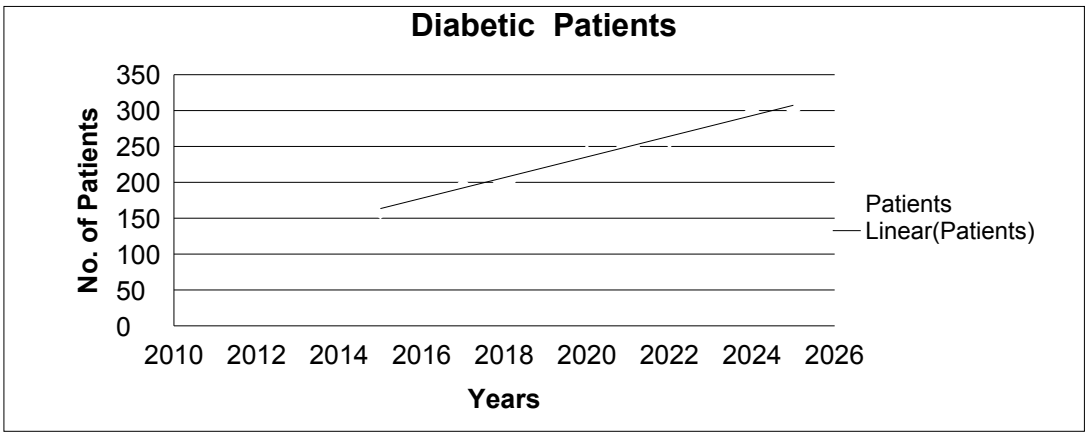

Fig. 6 Linear Growth of Diabetic Patients

V. PREVENTIVE MEASURES

The result provided by this developed software tool shows that the growth of diabetes will create major health and economic issues in India in 2025. The rapid growth should be controlled in order to save the human society from the diabetic hazards.

Some preventive measures will definitely assist the people to protect themselves from the dangerous diabetes. Food plays the key role in the process of prevention and control of diabetes. The role of medicinal plants is noticeable in the prevention and control the disease diabetes. This developed system recommends a variety of medicinal plants that can be taken as food supplement for the prevention or control the growth of diabetes. The following Fig. 7 shows the list of parts of plants and dosage given by this tool that can be used to prevent or control diabetes [11].

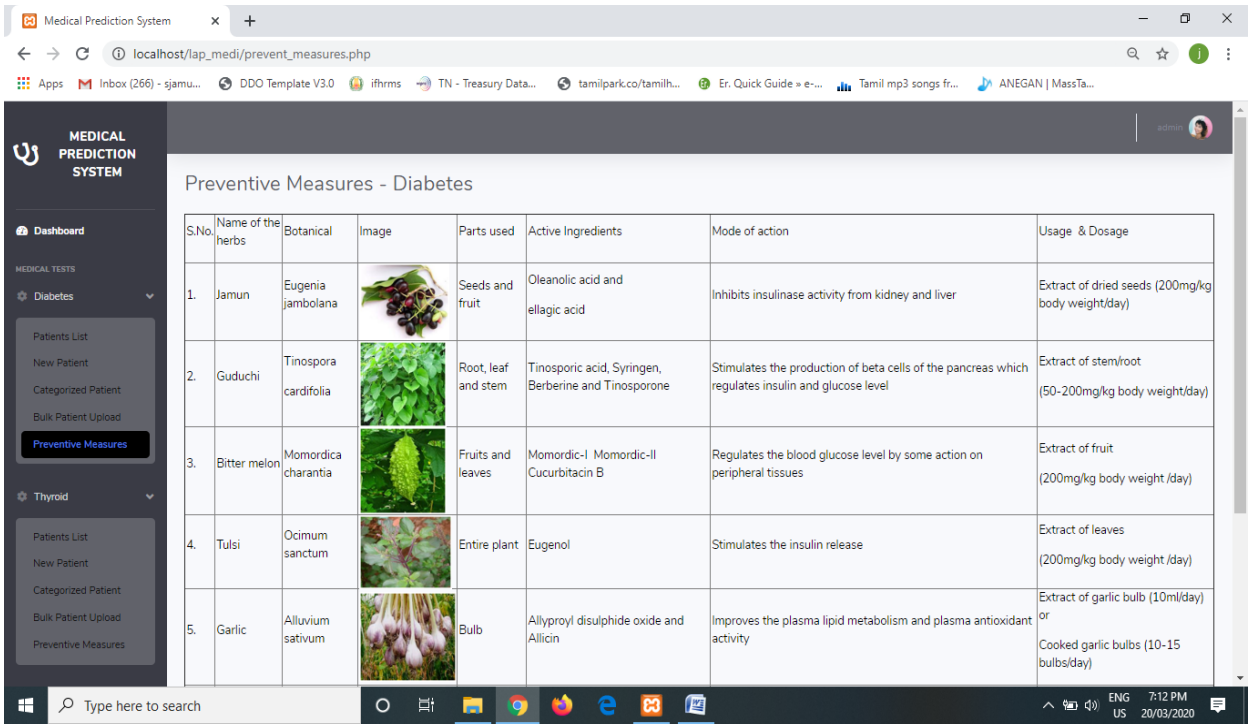

Fig.7 Herbal Remedies for Diabetes

VI. CONCLUSIONS

In this paper the refined diabetes prediction system which is very needful to the healthcare system is discussed elaborately. This developed prediction system first exactly identified and classified the diabetic patients. Then it showed the prediction result of diabetic patients using linear regression algorithm for the year 2025 and the system also recommended some medicinal plants which are capable of preventing or controlling the harmful diabetes. This diabetes prediction system will be a boon not only for the healthcare system but the entire human society.

REFERENCES

[1] Suresh Kumar.P and Umatejaswi.V (2017), "Diagnosing Diabetes using Data Mining Techniques" , *International Journal of Scientific and Research Publications,* ISSN 2250-3153 , Volume 7, Issue 6,PP.705-709.

[2] Pouya Saeedi , Inga Petersohn , Paraskevi Salpea , Belma Malanda , Suvi Karuranga , Nigel Unwin, Stephen Colagiuri (2019), " Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045 ", *International Diabetes Federation Diabetes Atlas, 9th edition,* Volume-157,PP.1-10.

[3] Vaishali , Nisha Pandey (2018),"Diabetes Prediction using Linear Regression, Decision Tree & Least Square Support Vector Machine", *International Journal of Innovative Research in Computer and Communication Engineering*, ISSN(Online): 2320-9801, ISSN (Print) : 2320-9798, Volume-6, Issue- 4, PP.3756-3763.

[4] Desmond Bala Bisandu, Dorcas Dachollom Datiri, Eva Onokpasa, Godwin Thomas, Musa Maaji Haruna, Aminu Aliyu and Jerry Zachariah Yakubu (2019), "Diabetes Prediction Using Data Mining Techniques", *International Journal of Research and Innovation in Applied Science*, ISSN 2454-6194,Volume-4, Issue-6, PP.103-111.

[5] Deepti Sisodia and Dilip Singh Sisodia (2018), "Prediction of Diabetes using Classification Algorithms", *International Conference on Computational Intelligence and Data Science*, Volume-132,PP.1578–1585.

[6] Smriti Mukesh Singh, Dinesh B. Hanchate (2018), "Improving Disease Prediction by Machine Learning" , *International Research Journal of Engineering and Technology*, ISSN(Online): 2395-0056, Volume: 05 Issue: 06,PP.1542-1548.

[7] Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, and Abhijeet Karve (2019), "Disease Prediction using Machine Learning", *International Research Journal of Engineering and Technology*, ISSN(Online): 2395-0056, Volume: 06 Issue: 12,PP.831-833.

[8] Vijiyarani.S and Sudha.S (2013), "Disease Prediction in Data Mining Technique –A Survey", *International Journal of Computer Applications & Information Technology*, ISSN: 2278-7720,Volume-2, Issue-1, PP.17-21.

[9] Mohan Kumar K N, S.Sampath, Mohammed Imran (2019), "An Overview on Disease Prediction for Preventive Care of Health Deterioration", *International Journal of Engineering and Advanced Technology*, ISSN: 2249- 8958, Volume-8, Issue-5S, PP.255-261.

[10] Douglas, Montgomery , Elizabeth A. Peck, G. Geoffrey Vining , *"Introduction to Linear Regression Analysis"*, 5th Edition, Wiley Series in Probability and Statistics,ISBN-13: 978-0470542811,2012

[11] Sonia Verma, Madhu Gupta, Harvinder Popli and Geeta Aggarwal (2018), "Diabetes Mellitus treatment Using Herbal Drugs", *International Journal of Phytomedicine,* Volume-10,Issue-1, PP.1-10.