Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020)
IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2

Survey on Type 2 Diabetes Prediction Using Machine Learning

P. Moksha Sri Sai [1]
M. Tech student, Dept of CSE

G. Anuradha[2]
Associate Professor, Dept of CSE

VVNV Phani kumar[3]
Assistant Professor, Dept of CSE

[1, 2,3]Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, AP, India.

mokshasrisai1997@gmail.com **ganuradha@vrsiddhartha.ac.in** **phanikumar.venna@gmail.com**

Abstract **-As consistent with world medical research diabetes is observed as the fastest growing diseases. In 2050, the world's diabetes patients will reach to 700 million, which means one of the 20adults in future is suffering from diabetes. Diabetes is India's fastest developing disease, where 70 million instances recorded in 2017, it may doubly in 2025. It may occur due to various causes such as viral infection, chemical contents mix with the food, environment pollution, eating habit, changes in food, etc. With the rapid development of present technology, it has implemented many components of scientific research. Many machine learning researches forecast the Type 2 Diabetes without need of medical exams. In this research, Data analysed by different types of algorithms to avoid risk factor of type 2 diabetes. Moreover, this paper explores the accuracy in diabetes prediction the use of medical reports with machine learning algorithms and methods.**
Keywords: Type 2 Diabetes, Machine Learning, world health organisation (WHO), Data Mining.

I. INTRODUCTION

Knowledge learning is a subtitle of intelligence retrieval. The predominant objective of device getting to know is to analyze the shape of facts and relevant the information to models that can be accepted and easily analyzed by the persons.

Knowledge learning is a subject in laptop device, it comes from some rapid applications, in traditional computational, Methods are sets of explicitly designed used to solve difficult task. Knowledge learning allow test and trains an input data and uses numerical analysis into final values with in a specific range.

Machine gaining knowledge of generally categorized in to two categories, these categories are based on how gaining knowledge of is obtained to the system developed. Two of the most widely used device getting to know techniques are Supervised studying, Unsupervised learning.

In supervised gaining knowledge of, pc is supplying inputs that are labelled with preferred outputs. The cause of the algorithm is in a position to 'learn' via comparing real output with "taught" output to discover errors.

In unsupervised learning, records are unlabeled; Algorithm is left to locate its input statistics. Unlabeled data are more accrue than labelled ones.

It commonly used for transactional data like large datasets of customers to purchase it.

Machine learning algorithms are good at health care to prevent it. In this paper, we are discussing about insulin-dependent and non insulin-dependent diabetes as follows.

Insulin-dependent diabetes is a common disease that can be characterised by excessive increase in blood glucose level. It mainly occurs in adults. In insulin dependent diabetes, pancreas not releases insulin due to the damage of pancreatic beta germ. High blood glucose ends in each acute and persistent complication bring about failure of various organs.

There are many challenges to manipulate blood glucose in Type 1 diabetes. One of them is that glucose kinetic process is complex. To keep away from this we are the use of some machine mastering algorithms like reinforcement gaining knowledge of, feed ahead controller. By using those we control blood sugar in insulin dependent diabetes without affecting to the organs.

Non insulin-dependent diabetes (diabetes mellitus) is a metabolism sickness that causes glucose to build up in the boby. The severity level of diabetes vary too high, some patients must make small adjustments to their life-style after they are diagnosed. Reducing a weight and getting workout may additionally enough for them to control their diabetes. Other people who have non insulin-dependent diabetes want durable remedy that entails taking drugs or injections. This is especially significant for them to have an excellent understanding of their illness and follow what they are able to do to be more active.

Things that are extraordinary in non insulin-dependent diabetes, in which sebum is made of pancreas but the body cells continuously lose the future to consume and use the sebum. In the past, non insulin-dependent diabetes becomes cited as "adult-onset" diabetes because it is commonly identified in

adults only. There are 90% of people who have suffering from non insulin-dependent diabetes.

If non insulin-dependent diabetes is natural, blood glucose levels are always be high. It is not substantive at starting time. Non insulin-dependent diabetes can evolve continuously over numerous decades without any great indications. Blood glucose levels might be constantly high may occurs the following indications, though

1. High thirst
2. Continuous urine
3. Weariness and laziness
4. Vomiting
5. Faintness

People who have non insulin-dependent diabetes, the pancreas simply give sufficient sebum; however now it is not an effect at the frame material and germs. Doctors refer to as “insulin resistance”. Pancreas can give amends by producing extra insulin. At few point we cannot maintain up, and then blood glucose levels start to rising up.

In this paper, we are discussed about Type 1and Type 2 diabetes. By using some methos we can prevent it early without affecting to the organs, they are shown below

1.1 Motivation

The main motivation of type2 diabetes prediction is to discover out the suffer is sufferening from sugar or not and checking the severity level of the patients. We are using some machine learning techniques like support vector machine, convolutional neural networks etc.

1.2 Scope

Non insulin-dependent diabetes forecast is used to reduce the risk factor before effecting to organs. The venture proposes SVM and RF algorithms for a patient dataset and trying to improve the attributes accuracy rate. In future, we can also predict the diabetes in children by using machine learning algorithms. In this model, we can use more efficient datasets for prediction of type 2 diabetes.

II. METHODS

A. *K-Means Algorithm*

Cluster study pursuits at dividing the statements into disparate parts so that statements within the equal cluster are higher nearly related to each aside the special parts [5]. K manner is the highly liked cluster algorithms. It is a typical distance-primarily based cluster methods, and the space is used as a degree of likeness i.e., the shorter distance between items specifies the more likeness. Indicates a compelling system of the K approach set of rules and strategies of the k manner Cluster method are suited for type 2 diabetes predictions. This can be used in first phase of non insulin-dependent diabetes

B. Logistic Regression

The class set of rules aimed to set up a method that may plan information recorded to a given class, primarily built at the current data. It altered to enhance meaningful facts entries from the method or to await the leaning of records. In logistic relapse method is paired-category. In ending up, we determined to apply the logistic relapse as a part of our current model. Logistic relapse method is based totally on linear relapse model. In any phase we can use this method

C. Support Vector Machine

Support vector machine (SVM) is used in each category of relapse. In this model, the statistics dots are pointed on the sphere and labelled into businesses and the dots with same things drops in equal cluster. In small SVM statistics group is taken into consideration as p length magnitude vector that can be spited by p-1 levels referred to as hyper-levels. The level that has most margins among the two training is known as the most-edge hyper-level. This can be used in both starting and final stage of diabetes

D. K-Nearest Neighbour (KNN)

K-Nearest neighbour is an easy method to perform better results. It is a idle, statistic and example based learning method. This method uses both relapse and cluster. In this category, KNN is carried out to locate the magnificence, belongs to new unlabeled objects. KNN forecast the result with better accuracy. Only in starting phase KNN can be used

E. Random Forest

The random forest approach is an adjustable, speed, and easy gadget getting to know method that is a merger of tree calls. Allocation is a great mission of system learning. In irregular forest, a irregular subset of features offers extra correct consequences on massive data items, and more irregular trees developed by solving a irregular edge for all features, rather of locating maximum correct edge. The set of rules additionally solves the over fitting issue. In the final stage of the prediction we are using random forest.

F. Decision Tree:

In the previous decades, a high range of methods had been established for category based records digging. This is a critical class method in statistics digging. The main favour of choice tree methods is that they are smooth to construct and the resulting bushes are effortlessly interpretable popular choice tree algorithms along with ID3, CART, C4.five, C5.0, and J48 etc. C5.zero and J48 are the improved variations of C4.five algorithms. In the WEKA information digging tool, J48 set of policies in an unlock supply Java execution of C4.5 method. By using this technique, a tree is built to version the class process. Once the tree is built, it's far carried out to each tuple inside the database and outcomes in category for the tuple.

G. Naive Baye's

The Naive Bayes method is used for allocation problems. This is well organized type method in records digging can manage misplaced values throughout allocation [8] [9]. This method is quite rapid dynamic model. Mostly, this type is used for unsolicited mail leak and emotional evaluation. The name of Naive is known as for it is kind of capabilities different for an occurrence of one more item. It is used in last phase of non insulin-dependent diabetes.

In this paper, Discussed about some algorithms and their accuracies as shown below "TABLE1"

TABLE1. TECHNIQUES WITH ACCURACY

S.No	Techniques	Accuracy
1.	K-Means Algorithm	75%
2.	Logistic Regression(LR)	81%
3.	Support Vector Machine(SVM)	93%
4.	K-Nearest Neighbour(KNN)	84%
5.	Random Forest(RF)	77%
6.	Decision Tree(DT)	78%
7.	Naive Bayesian(NB)	79%

III. LITERATURE SURVEY

This section explains comparison of different data mining and machine learning algorithms, comparision table is shown in Table2.

TABLE 2. COMPARISON OF DIFFERENT MACHINE LEARNING (ML) ALGORITHMS

S. No	Year	Author	Methodology	Advantage	Limitation
1.	2018	Hon wu, Shergqi Yang, et.al.,	k-means algorithm, Logistic regression	We solve better accuracy of forecast model and creating model adapt to distinct data items.	Data item should be huge enough for both training and testing.
2.	2018	Basharat naqui, Arshid ali.et.al.,	Decision tree, ID3, Random forest	It contains two labels yes or no, yes indicates diabetes and no indicates absence of diabetes.	It must provide OFPS and OFNS scenario for the better accuracy of an diabetes prediction.
3.	2018	Priyanka indira, Yogesh kumar rathore	Artifical neural networks, Bayesian network	This method uses PNN model out performs other models such as ANN etc.	In future, writers can use both neural networks and computational methods.
4.	2018	Samart kumar, Ashraf hossian	KNN, Support vector machine, Naive bayes	ANN provides highest accuracy with highest and lowest scaling method on hospital pima data items.	In future focus on expert system model and maintain local data items.
5.	2018	Quan zou, Kalyan qu	Decision tree, Random forest, Neural networks.	In paper by using random forest and decision tree we get an highest	It works on only single patient data not on whole hospital data

				accuracy	
6.	2018	Swapna. g, Vinaya kumar.	Recurrent neural networks, Long brief term memory, convolution al neural networks.	By using deep learning concepts we can early detection of diabetes is extremely crucial	We can take the fall further doctors and patients datasets for better prediction of diabetes.
7.	2018	Harlem Kaur, Vinita Kumar	KNN, ANN, Multifactor Dimensional Reduction.	Using dimensional Reduction, the dataset used to measure several values at a time.	It works only few peoples. It cannot take bulk data.
8.	2018	Javson Weston	Support Vector Machine, Logistic Regression.	Author conveys about the dataset used in the paper was to predict early stage of diabetes.	Svm provides better accuracy than any other algorithms.
9.	2018	Narges Razavia n, Squal Blecker.	AUC, Random Forest.	By using AUC we can predict diabetes very easily without affecting to the organs.	Random forest contains an better performance accuracy than AUC.
10.	2017	Rahul joshi, min yechil.	Random forest, KNN.	Finally selects the organized data items for better outcome.	Doing with one algorithm may occurs bad results. So provide multiple algorithms.
11.	2017	Wenqia n chen, Shuvu chen.et.a l.,	K-means, Decision tree	By using confusion matrix we apply J48 decision tree algorithm for an better accuracy.	In future is required is to access the effectivenes s of proposed method with large amount of data.
12.	2017	Loannis Kavakio tis, Yasin Ali.	ANN, Random Forest, K-Means.	By using ANN we get the several neural networks to predict the diabetes patients.	In further extension looking for large dataset form patients for early prediction.
13.	2017	Jianferg Zhang, Jiatuoxu .	PCA, SVM.	PCA is used for measuring an single patient data and specifies the patient suffering form it or not.	Support vector machine contains less accuracy than the principal component analysis.
14.	2017	David J. Abbas, Lejia Alic.	Dynamic Gaussian.	By using Gaussian filter we can enhance the missing data recovery for duplicate values.	It contains better performance than other papers by using ECG signals.
15.	2017	Khalil. M, Adel Jamaily.	Support vector machine, K-Means, Probabilistic neural networks.	By using probabilistic neural networks, we can predict early stage of diabetes using PIMA datasets.	Support vector machine contains more accuracy than the probabilistic neural networks and k-means.
16.	2017	Jongeh Kim,	Logistic regression,	Using deep neural	Deep neural networks are

		Junmo Kim.	Deep neural networks.	networks, we can predict early of type 2 diabetes.	the one of the best machine learning technique for diabetes.
17.	2017	Francesco Mercado, Vitoria Nardonr.	Random Forest, SVM.	SVM enhances better performance accuracy than other models.	Random Forest uses a less dataset values than other ones so better to use other models.
18.	2017	Asir Anatomy, Gnana Singh.	Random Forest, Convolutional neural networks.	By using random forest we get diabetes patients list clearly rather than others.	Random forest contains better accuracy than neural networks.
19.	2018	Muhammad Azeem, Nasir Kamal.	K-means, Random forest.	K-means the data is divided into number of clusters to enhance the diabetes patients easily.	Here in this paper, K-Means contains better performance to predict diabetes in early stage.
20.	2019	Talha Mahboob, Yasin Ali	Artificial Neural Network, Random Forest, K-Means.	Artificial neural network works on special attributes to avoid the blood sugar level.	By using artificial neural networks we get better performance matrix.
21.	2019	Hasan T. Abbas, Lejia Alic.	Support Vector Machine.	By using an Support Vector Machine we can enhance hyper line for two classes.	It reduces an diabetes patients count and gives better performance accuracy.
22.	2019	Binh P. Nguyen Hung n. Pham.	SMOTE,AUC	By using smote methods we can have less level of attributes to check whether the patient is diabetes or not.	Auc is a better modelling one which contains an all the types of attributes without excluding it.
23.	2019	An Dinh, Stacey Miertuchin.	Decision tree(DT), Logistic Regression algorithm.	Logistic regression algorithm is a good algorithm for non insulin-independent diabetes.	It contains a high accuracy values than other machine learning techniques.
24.	2019	N.Sneha, Tarun Ganglia.	Naive Bayesian, Artificial neural networks	Naive Bayesian model extracts the dataset with ECG signals and explores about the diabetes patients.	By using an Naive Bayesian (NB) we get better performance accuracy with 84%.
25.	2019	Sajida Perveen, Muhumad Shabhaz.	K-Means, Random Forest.	By using k-means we have cluster to explore the patient's data and controls diabetes.	Using K-means get better accuracy than other ones.

IV. CONCLUSION

Non insulin-dependent diabetes imposes unstoppable and notable problems on the public in time period of misplaced efficiency, temporality and untouchable prices for poor first-class of living. Now-a-days risk factor for developing diabetes has been increasing day by day. The important part of this experiment be made up in developing SVM to take out divining statistics from clinical size in way to decide a-previous 20 year of possibility to develop it. The advanced research has the capability to use in hospital settings to verify possible sensitive folks who are maximum probably good from protective remedy, whilst escape needless surgery for people who are at low risk.

V. FUTURE WORK

For later work, it is far vital to conduct in hospitals actual and newest sufferers' information for regular instructing and development of our present model. The amount of the data item has massive sufficient for education and forecasting. Some further methods and portraits need to take a look of DM (data mining). To expand a chain of regulation and quality approach to stop human beings from come out of data mining. It assists to minor the increase price of blood glucose and subsequently reduce the danger of forecasting data mining.

REFERENCES

[1] Francesco Mercaldoa, Vittoria Nardoneb, Antonella Santoneb, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques", Procedia Computer Science, Elesevier 112 (2017) 2519–2528.
[2] Loannis Kavakiotis, Olgatsave, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural biotechnology, Elsevier 15(2017)104-116.
[3] Rahul Joshi, Minyechi lAlehegn, "Analysis and Prediction of Diabetes Diseases Using Machine Learning Algorithm: Ensemble approach", International Research Journal of Engineering and Technology (IRJET) 10(2017)2395-0072.
[4] Wenqian Chen, Shuyu Chen, Hancui Zhang, "A Hybrid Prediction Model for Type 2 Diabetes Using Machine learning and Decision Tree algorithms", National herbal basis of technology in china, IEEE (2017)5386-0497.
[5] Asir Antony Gnana Singh, Jebamalar Leavline, "Diabetes Prediction Using Medical Data", Journal of Computational Intelligence in Bioinformatics, ISCN 0973-385X Volume 10, Number 1(2017).
[6] Khali, jumaily, "Machine learning based prediction on depression among type 2 diabetes patients", International conference on intelligence system and knowledge engineering, Science Direct 12(2017).
[7] Jongoh Kim, Junmo Kim, Min Ji Kwak, "Genetic prediction of type 2 diabetes using deep mach neural network", Computational and Structural biotechnology 109(2017)10-111.
[8] H. Kaur, V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach", Applied computing and informatics, Science Direct(2018).
[9] Han Wu, Shangri Yang , Zhangqin Huang, Jian He, Xiaoyi Wang, "Type2 diabetes mellitus prediction model based totally on information mining", Informatics in Medicine Unlocked, Elsevier 10 (2018) 100–107.
[10] Swapna G., Vinayakumar R., Soman K.P, "Diabetes detection the usage of deep gaining knowledge of algorithms", ICT Express, Science Direct 4 (2018) 243–246.
[11] Basharat Naqvi, Arshad Ali, Muhammad Adnan Hashmi," Prediction Techniques for Diagnosis of Diabetic Disease using Machine learning", IJCSNS, VOL.18 No.8, August (2018) 118.
[12] Priyanka Indoria, Yogesh Kumar Rathore, "A Survey: Detection and Prediction of Diabetes Using Machine Learning Techniques", International journal of Engineering research and technology 03(2018)2278 0181.
[13] Samrat Kumar Dey, Ashraf Hossain, "Implementation of a Web Application to Predict diabetes Disease using Machine Learning Algorithm", International Conference of Computer and Information Technology (ICCIT), 12(2018).
[14] Quan Zou*, Kaiyang Qu, Yamei Luo, "Predicting Diabetes Mellitus with Machine Learning Techniques", Frontiers in genetics, 06 NOV (2018)00515.
[15] Muhammad Azeem Sarwar, Nasir Kamal, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", International Conference on Automation & Computing, Elsevier, 07 sep (2018).
[16] Binh P. Nguyen, Hung N. Pham, Hop Tran, "Predicting the onset of type 2 diabetes the use of huge and deep Neural Network with electronic health records", Computer Methods and Algotithms in Biomedicine, Elsevier 182 (2019) 105055.
[17] Talha Mahboob Alama, Muhammad Atif Iqbala, Yasir Alia, "A model for Fast prediction of diabetes", Informatics in Medicine Unlocked Systems, Elsevier 16(2019)100204.
[18] Jason Weston, "Diabetes Prediction Using Machine Learning", NEC Labs America,4 Independence Way 142(2019) 105015.
[19] David J. Albers, Matthew Levine, "Personalized Glucose Forecasting for Type2 Diabetes Using Data Assimilation", Computational biology, PCBI, 27 APR (2017)1005232.
[20] Sajida Perveen, Muhammad Shahbaz, "Prognostic Modelling and Prevention of Diabetes Using Machine Learning Technologies", Science Reports, 24 SEP (2019)41598-019-49563.
[21] Narges Razavian, Saul Blecker, "Population Level Prediction of Type2 Diabetes from Claim Data and Analysis of Risk Factors", Original Article, Volume 3 Number 4, 10-1089(2015)0020.
[22] An Dinh, Stacey Miertschin, "A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease With Machine Learning", BMC Medical Information and Decision Making, Springer, 06 NOV (2019)19-211.
[23] Jiafeng Zhang, Jiatuo Xu, "Diagnosing Method of Diabetes Based on Support Vector Machine", Biomed Research International, 04 Jan (2017)7961494.
[24] N. Sneha, Tarun Gangil, "Analysis of Diabetes Mellitus for Early Prediction Using Optimal Feature Selection", Journal of Big data, Springer, 06 Nov (2019)13.
[25] Hasen T. Abbas, Lejla Alic, "Predicting Long-Term Type 2 Diabetes With Support Vector Machine Using Oral Glucose Tolerance Test", NPRP 10-1231-160071, 11 Dec (2019)0219636.
[26] Waqas Samil, Tahir Ansari, "Effect of Diet on Type 2 Diabetes Mellitus: A Review", Qassim University, IJHS, volume 11, Apr (2017).
[27] Nita G Forouhi, Anoop Misra, "Dietary and Nutritional Approaches For Prevention and Management of Type 2 Diabetes", BMJ, 13 June (2018)10.1136.
[28] NP Steyn, J Monn, "Diet, Nutrition and the Prevention of Type 2 Diabetes", Public health nutrition, volume 07(1A), Nov (2017)147-165.
[29] Franziska Jannasch, Janime Kroger, "Dietary Patterns and Type 2 Diabetes: A Systematic Literature Review and Meta-Analysis of Prospective Studies", Journal of Nutrition, 19 Apr (2017) volume 147.
[30] Abdulfatai B, Okolona, "Type 2 Diabetes Mellitus review of Current Trends using Machine Learning", OMAN medical journal, 08 may (2012)269-273.