International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 03, March-2018

A Survey: Detection and Prediction of Diabetes Using Machine Learning Techniques

[1] Priyanka Indoria,
M.Tech.,
Dept. of CSE,
Raipur Institute of Technology,
Raipur, Chhattisgarh, India

[2]Yogesh Kumar Rathore,
Assistant Professor,
Dept. of CSE, Raipur Institute of Technology,
Raipur, Chhattisgarh, India

***Abstract*— Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a "key" to open our cells, to allow the glucose to enter -- and allow us to use the glucose for energy. But with diabetes, this system does not work. Several major things can go wrong – causing the onset of diabetes. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. This paper focuses on recent developments in machine learning which have made significant impacts in the detection and diagnosis of diabetes.**

Keywords- Diabetes, Type 1 Diabetes, Type 2 Diabetes, Insulin, Glucose, Machine Learning.

I. INTRODUCTION

In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source – everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to 'doors' on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body can - not use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycemia) and diabetes develops. Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

Signs or Symptoms of Diabetes: Frequent Urination, Increased thirst, Increased hunger, Tired/Sleepiness, Weight loss, Blurred vision. Mood swings, Confusion and difficulty concentrating, frequent infections / poor healing. Type 1 diabetes : In Type 1 diabetes the beta cells of the pancreas have been injured or attacked by the body's own immune system (auto - immunity). As a result of this attack, the beta cells die and are therefore unable to make the required amount of insulin to move glucose into the cells, causing high blood sugar (hyperglycemia). Type 1 diabetes occurs in about 5 - 10% of those with diabetes and usually in people less than 30 years of age, but can occur at any age. The signs and symptoms have a rapid onset and are usually intense in nature. As Type 1 diabetes is caused by a lack of insulin, people need to replace what the body cannot produce itself.

According to the latest American Heart Association's Heart Disease and Stroke Statistics, about 8 million people 18 years and older in the United States have type 2 diabetes and do not know it. Often type 1 diabetes remains undiagnosed until symptoms become severe and hospitalization is required. Left untreated, diabetes can cause a number of health complications. That's why it's so important to both know what warning signs to look for and to see a health care provider regularly for routine wellness screenings [1]. Computer Aided Diagnosis is a rapidly growing dynamic area of research in medical industry. The recent researchers in machine learning machine learning promise the improved accuracy of perception and diagnosis of disease. Here the computers are enabled to think by developing intelligence by learning. There are many types of Machine Learning Techniques and which are used to classify the data sets [1, 2]. They are Supervised, Unsupervised, Semi-Supervised, Reinforcement, Evolutionary learning, and deep learning algorithms [2].

II. LITERATURE SURVEY

1. *Machine Learning Techniques for Classification of Diabetes And Cardiovascular Diseases. Berina Et Al. [3]:*

Abstract: the overview of machine learning techniques in classification of diabetes and cardiovascular diseases (CVD) using Artificial Neural Networks (ANNs) and Bayesian Networks (BNs). The comparative analysis was performed on selected papers that are published in the period from 2008 to 2017. The most commonly used type of ANN in selected papers is multilayer feedforward neural network with Levenberg-Marquardt learning algorithm. On the other hand, the most commonly used type of BN is Naive Bayesian network which shown the highest accuracy values for classification of diabetes and CVD, 99.51% and 97.92% retrospectively. Moreover, the calculation of mean accuracy of observed networks has shown better results using ANN, which indicates that higher possibility to obtain more accurate

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 03, March-2018

results in diabetes and/or CVD classification is when it is applied to ANN.

In this paper authors designed to perform a review of Artificial Neural Network and Bayesian Network and their application in classification of diabetes and CVD diseases. The purpose is to show the comparison of machine learning techniques and to discover the best option for achieving the highest output accuracy of the classification.

Methods: This paper represents the comparison of application of two machine learning techniques, Artificial Neural Network and Bayesian Network in classification of diabetes and cardiovascular diseases. Guided by experience of researchers from the papers [3, 4] that also reviewed machine learning techniques but in different field of studies, the literature review was done using 20 published papers in order to obtain the relevant results about diabetes and CVD classification in the period from 2008 to 2017.

Table 1 Ann Types for Classification of Diabetes and Cvd

Paper	Type of ANN
DIABETES	
[5]	Multilayer feedforward neural network with sigmoid transfer function
[6]	Feedforward neural network using Levenberg-Marquardt method
[7]	Multilayer perceptron with backpropagation learning algorithm and genetic algorithm
[8]	Two-laver feedforward neural network with sicrmoid function
[9]	Probabilistic neural network
CVD	
[10]	Multilayer neural network with statistical backpropagation of error
[11]	Backpropagation neural network with sigmoid transfer function
[12]	Feedforward neural networks with sigmoid transfer function using Levenberg -Marquardt learning algorithm and SCG
[13]	Feedforward multilayer perceptron with sigmoid activation function trained with backpropagation algorithm
[14]	MLP neural network with sigmoid transfer function

The overview of Artificial Neural Networks used for classification of diabetes and CVD (Table 1) shows that the most commonly used type of network in both diseases is multilayer feedforward neural network. As training algorithm, most of authors of selected papers [5-11] have decided to use Levenberg-Marquardt learning algorithm. Each network uses error backpropagation algorithm to compare the system output to the desired output value, and uses the calculated error to direct the training. The difference in the architectures of these Networks is in transfer function where sigmoid transfer function is the most commonly used one.

Table Ii. Bn Types For Classification Of Diabetes And Cvd

Paper	Type of BN
DIABETES	
[15]	Naive Bayesian Network
[16]	Naive Bayesian Network
[17]	Naive Bayesian Network
[18]	MLP + Naive Bayesian Network
[19]	Naive Bayesian Network
CVD	
[20]	Markov blanket estimation
[21]	Dynamic Bayesian network
[22]	Naive Bayesian network
[23]	Naive Bayesian network
[24]	Naive Bayesian network

The overview of Bayesian Networks used for classification of diabetes and CVD (Table II) shows that the most commonly used type of network in both diseases is Naive Bayesian network. Naive Bayesian networks are very simple BNs which are composed of directed acyclic graphs with only one unobserved node and several observed nodes. This type of BNs applies Bayes' theorem with strong independence assumptions between features and does not need a long computational time for training which is its major advantage.

Results: In the comparison of application of Artificial Neural Network and Bayesian Network for classification of diabetes and CVD, different values for the network accuracy have been achieved. The results of trained ANN and BN for classification of diabetes from selected papers [5-9, 15-19]. Authors observed that the accuracy of diabetes classification using ANN varies between 72.2% and 99%. And the accuracy of diabetes classification using BN varies between 71% and 99.51 %. According to compared results, the highest accuracy was achieved in Bayesian Network but also the smallest accuracy was shown in Bayesian Network.

Conclusion: One of the biggest causes of death worldwide are diabetes and cardiovascular disease. The early classification of these diseases can be achieved developing machine learning models such as Artificial Neural Network and Bayesian Network. In comparison of mean accuracy of 10 scientific papers about diabetes classification and 10 papers about CVD classification it was concluded that the higher accuracy was achieved with ANN in both cases (87.29 for diabetes and 89.38 for CVD). The used Naive Bayesian network, due to the assumption of independence among observed nodes, might be less accurate than ANN approach. So, in accordance to obtained result it can be concluded that the higher possibility to obtain better accuracy in classification diabetes and/or CVD is when it is applied to Artificial Neural Network.

2. *Automatic Diagnosis of Diabetic Retinopathy, Dinu A.J Et Al. [25]:*

DME is one of the largest causes of visual loss in diabetes. There are various machine learning algorithms that can be used to improve the accuracy of diagnosis of diabetic retinopathy.

Iyer has performed a work to predict diabetes disease by using decision tree and Naive Bayes. J48 shows 76.95% accuracy by using Cross Validation and Percentage Split Respectively [26]. Naive Bayes presents 79.56% correctness by using PS.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 03, March-2018

Algorithms show the highest accuracy by utilizing percentage split test.

Sen and Dash developed Meta-learning algorithms for diabetes disease diagnosis. CART, Adaboost, Logiboost, and grading learning algorithms are used to predict that patient has diabetes [27]. From experimental results CART offers 78.64% accuracy. The Adaboost obtains 77.86% exactness. Logiboost offers the correctness of 77.47%. Also Misclassification Rate of 21.35%, which is smaller as compared to other techniques.

R. Catherine Silvia introduced a feature extraction technique. This technique is used to capture the global characteristics of the fundus images and separate the normal from DME images. Diabetic macular edema detection is carried out via supervised learning. Disease severity is assessed using a rotational asymmetry metric by examining the symmetry of macular region [28]. A microaneurysm is identified using Circular Hough Transform. The detection performance has specificity between 74% and 90%.

3. A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type Ii, Zahed Soltani Et Al. [29]

Abstract: Diabetes is one of the major health problems as it causes physical disability and even death in people. Therefore, to diagnose this dangerous disease better, methods with minimum error rate must be used. Different models of artificial neural networks have the capability to diagnose this disease with minimum error. Hence, in this paper authors has used probabilistic artificial neural networks for an approach to diagnose diabetes disease type II. Authors took advantage of Pima Indians Diabetes dataset with 768 samples in their experiments [30]. According to this dataset, PNN is implemented in MATLAB. Furthermore, maximizing accuracy of diagnosing the Diabetes disease type II in training and testing the Pima Indians Diabetes dataset is the performance measure in this paper. Finally, authors concluded that training accuracy and testing accuracy of the proposed method is 89.56% and 81.49%, respectively.

Method: In order to identify diabetes and other diseases such as heart diseases, Parkinson's disease, and lung cancer, having a data set is very important and necessary, since ANNs are trained by these data sets and they can perform the diagnosis task. Therefore, in this paper, authors used Pima Indians Diabetes with 768 data sample for diagnosing diabetes type 2. This data set consists of 9 features for each data sample. Table (III) shows these 9 features.

According to Table III, there are 9 features for each data sample. The first 8 features are inputs, and the last feature is the only output. In order to classify the 768 data samples, 9th feature is used as it is classified into two classes: class zero (healthy) and class 1 (patient).

Table Iii. Features of Pima Indians Diabetes for Diagnosing Diabetes Disease Type 2 [30].

No. of Attributes	Attributes	Descriptions and Attributes values
1	Number of Times Pregnant (NTP)	Numerical values
2	Plasma Glucose Concentration (PGC)	Numerical values
3	Diastolic Blood Pressure (DBP)	Numerical values in (mm Hg)
4	Triceps Skin Fold Thickness (TSFT)	Numerical values in mm
5	2-Hour Serum Insulin (2-HSI)	Numerical values in (mu U/ml)
6	Body Mass Index (BMI)	Numerical values in (weight in kg/(height in m)^2)
7	Diabetes Pedigree Function (DPF)	Numerical value
8	Age	Numerical values
9	Diagnosis of type 2 diabetes disease	Yes=1 No=0

Table Iv. Statistical Analysis for Mean And Standard Deviation In Pima Indians Diabetes Data Set [30].

No. of Feature	Feature Name	Mean	Standard Deviation
1	Number of times pregnant	3.8	3.4
2	Plasma glucose concentration	120.9	32.0
3	Diastolic blood pressure	69.1	19.4
4	Triceps skin fold thickness	20.5	16.0
5	2-Hour serum i insulin	79.8	115.2
6	Body mass index	32.0	7.9
7	Diabetes pedigree function	0.5	0.3
8	Age	33.2	11.8

The average age of this data set is between 21 and 81 years. In addition, according to Pima Indians Diabetes data set which has 768 data samples, Table 2 shows the Mean and standard deviation of the data set.

ANNs consist of different models such as PNN, MLP, RBF, and GRNN. In this paper, Authors uses PNN model for diagnosing diabetes type 2. PNN model has a parallel structure and is special for information classification. In contrast to other ANNs such as MLP, PNN has a higher speed in training the data, and it finds answers faster than MLP. This model consists of 3 layers: input layer, hidden layer, and output (competitive) layer. The hidden layer is also called radial base layer, as PNN model is a mode of RBF model.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 03, March-2018

Hidden layer units uses Gaussian transmission function, and number of neurons in this layer is same as number of rounds in training data set. This layer computes distance between input vector and training inputs, and provides a vector where its elements determine the distance between the input and training inputs. Hidden layer generates a vector of probabilities as output. Finally, this layer selects probability values from probabilities vector and generates value 1 for it and 0 for other probabilities. Gaussian transmission function which is used in hidden layer calculated.

Results: all traditional studies with Pima Indians Diabetes data set except Bayesian Regulation perform the task of identifying diabetes type 2. It is observed by authors that their method using PNN model outperforms other models such as back-propagation, Bayesian Regulation, ANN, and GRNN in terms of accuracy of diagnosing diabetes type 2. MLP is the only model which has a higher accuracy than our PNN model. Back-propagation with 82% accuracy is the worst approach. This value is close to the 82.99% accuracy which belongs to GRNN. Furthermore, MLP with 97.61% accuracy is the best approach in training phase. Also, accuracy of ANN and PNN are too close to each other.

Conclusion: in this paper PNNs are applied to identifying diabetes disease type 2. Authors of this paper implemented the PNN model in MATLAB. The Pima Indians Diabetes data set was used for diagnosing diabetes type 2, which consists 768 data samples with 9 features. 90% of these 768 samples are used as training set and 10% used as testing set. The method achieved 89.56% of diagnosis accuracy in training phase, and 81.49% in test phase. Both training and testing measures could identify the diabetes disease type 2 with a good accuracy. As a future work, authors can use the combination of fuzzy and artificial neural networks or combination of genetic and artificial neural networks for diagnosing diabetes type 2.

4. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier. Eka Miranda Et Al. [31].

Objectives: The number of deaths caused by cardiovascular disease and stroke is predicted to reach 23.3 million in 2030. As a contribution to support prevention of this phenomenon, this paper proposes a mining model using a naïve Bayes classifier that could detect cardiovascular disease and identify its risk level for adults.

Methods: The process of designing the method began by identifying the knowledge related to the cardiovascular disease profile and the level of cardiovascular disease risk factors for adults based on the medical record, and designing a mining technique model using a naïve Bayes classifier. Evaluation of this research employed two methods: accuracy, sensitivity, and specificity calculation as well as an evaluation session with cardiologists and internists. The characteristics of cardiovascular disease are identified by its primary risk factors. Those factors are diabetes mellitus, the level of lipids in the blood, coronary artery function, and kidney function. Class labels were assigned according to the values of these factors: risk level 1, risk level 2 and risk level 3.

Results: The evaluation of the classifier performance (accuracy, sensitivity, and specificity) in this research showed that the proposed model predicted the class label of tuples correctly (above 80%). More than eighty percent of respondents (including cardiologists and internists) who participated in the evaluation session agree till strongly agreed that this research followed medical procedures and that the result can support medical analysis related to cardiovascular disease. Conclusions: The research showed that the proposed model achieves good performance for risk level detection of cardiovascular disease.

REFERENCES

[1] www.diabetesresearch.org/document.doc?id=284
[2] D. Yu, and L. Deng, 2011, “Deep learning and its applications to signal and information processing,” IEEE Signal Process. Mag., vol. 28, no. 1, pp. 145-154.
[3] Habibi, N., Hashim, S. Z. M., Norouzi, A., & Samian, M. R. (2014). A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. BMC bioinformatics, 15(1), 134.
[4] Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. Acta Informatica Medica, 24(5), 364.
[5] Olaniyi, E. O., & Adnan, K. (2014). Onset diabetes diagnosis using artificial neural network. International Journal of Scientific and Engineering Research, 5(10).
[6] Jayalakshmi, T., & Santhakumaran, A (2010, February). A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. OSDE, 159-163. (2010)
[7] Pradhan, M., & Sahu, R. K. (2011). Predict the onset of diabetes disease using Artificial Neural Network (ANN). International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004).
[8] Sejdinovic, Dijana, et al. "Classification of Prediabetes and Type 2 Diabetes using Artificial Neural Network." Springer. CMBEBIH 2017.
[9] Soltani, Z., & Jafarian, A (2016). A New Artificial Neural Networks Approach for diagnosing Diabetes Disease Type II. International Journal of Advanced Computer Science & Applications, 1(7), 89-94.
[10] Atkov, O. Y., Gorokhova, S. G., Sboev, A G., Generozov, E. Y., Muraseyeva, E. v., Moroshkina, S. Y, & Cherniy, N. N. (2012). Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. Journal of cardiology, 59(2), 190-194.
[11] Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart diseases diagnosis using neural networks arbitration. International Journal of Intelligent Systems and Applications, 7(12), 72.
[12] 12] Colak, M. C. et. al., Predicting coronary artery disease using different artificial neural network modelslkoroner arter hastaliginin degisik yapay sinir agi modelleri lie tahmini. The Anatolian Journal of Cardiology (Anadolu Kardiyoloji Dergisi), 8(4), 249-255, (2008).
[13] Can, M. (2013). Diagnosis of cardiovascular diseases by boosted neural networks.
[14] Sayad, A T., & Halkarnikar, P. P. Diagnosis of heart disease using neural network approach. In Proceedings of IRF International Conference, 13th April-2014, Pune, India, ISBN (pp. 978-93).
[15] Guo, Y, Bai, G., & Hu, Y (2012, December). Using bayes network for prediction of type-2 diabetes. In Internet Technology and Secured
Transactions, 2012 International Conference for (pp. 471-472). IEEE.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 03, March-2018

[16] Kumari, M., Vohra, R., & Arora, A (2014). Prediction of Diabetes Using Bayesian Network.
[17] N. Sarma, S. Kumar, AK. Saini, A Comparative Study on Decision Tree and Bayes Net Classifier for Predicting Diabetes Type 2, 2014, ISSN: 2278-0882, ICRTIET-2014.
[18] Dewangan L. A., & Agrawal, P. Classification of Diabetes Mellitus Using Machine Learning Techniques.
[19] Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science, 69,132-142.
[20] Elsayad, A, & Fakr, M. (2015). Diagnosis of cardiovascular diseases with Bayesian classifiers. 1. Comput. Sci., II (2), 274-282.
[21] K. P. Exarchos, et al. Prediction of coronary atherosclerosis progression using dynamic Bayesian networks. IEEE EMBC, 2013.
[22] D.S. Medhekar, M.P. Bote & Deshmukh, S. D., Heart disease prediction system using naive bayes. Int. J. Enhanced Res. Sci. Technol. (2013).
[23] Patil, R. R., Heart disease prediction system using naive bayes and jelinek-mercer smoothing. International Journal of Advanced Research in Computer Science and Communication Engineering, (2014).
[24] E. Miranda et. al., Detection of CYD Risk's Level for Adults Using Naive Bayes Classifier. Healthcare Informatics Research, (2016).
[25] Dinu A.J., Ganesan R, Felix Joseph and Balaji V, "A study on Deep Machine Learning Algorithms for diagnosis of diseases." International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 6338-6346.
[26] R. Catherine Silvia, R. Vijayalakshmi, 2013, "Detection of Non-Proliferative Diabetic Retinopathy in fundus images of the human retina", International Conference on Information Communication and Embedded Systems (ICICES).
[27] Iyer A., Jeyalatha S. and Sumbaly R, 2015, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), 5, 1-14.
[28] Sen S.K. and Dash S, 2014, "Application of Meta Learning Algorithms for the Prediction of Diabetes Disease", International Journal of Advance Research in Computer Science and Management Studies, 2, 396-401.
[29] Zahed Soltani and Ahmad Jafarian, "A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II." International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016
[30] Pima Indians Diabetes Data Set, https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes [Last Available: February 2016].
[31] E. Miranda et. al., Detection of CYD Risk's Level for Adults Using Naive Bayes Classifier. Healthcare Informatics Research, (2016).

AUTHOR'S PROFILE

1. Mrs. Priyanka Indoria is pursuing Master in Technology in Computer Science & Engineering from Raipur Institute of Technology, Raipur, Chhattisgarh, Affiliated from Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India. Her area of interest includes Digital image processing and Computer Grphics,

2. Mr. Yogesh Rathore is Assistant Professor in Department of Computer Science & Engineering, Raipur Institute of Technology, Raipur, Chhattisgarh, India. He is M. Tech. in Computer Science & Engineering. His area of interest include Digital image processing and Computer Graphics.