EARLY PREDICTION MODEL FOR TYPE-2 DIABETES BASED ON LIFESTYLE

Kirti Hirnak[1,3,*] , *Nikita* Chaudhari[2,* *] , *Akshay* Singh[3,* * *] , and *Deepali* Patil[1,3,* * * *]

[1]Department of Information Technology, Ramrao Adik Institute of Technology,Navi Mumbai, India
[2]Department of Information Technology, Ramrao Adik Institute of Technology,Navi Mumbai, India
[3]Department of Information Technology, Ramrao Adik Institute of Technology,Navi Mumbai, India
[4]Department of Information Technology, Ramrao Adik Institute of Technology,Navi Mumbai, India

Abstract. Nowadays diabetes has become a chronic disease that may cause many complications. There are some symptoms of diabetes such as increased appetite, blurry vision, and extreme fatigue, etc. As the increasing deformity in present years the number of diabetic patients from the whole world will reach to 642 million. Diabetes accuracy is very difficult to know so in order to cure this disease. These causes us to concentrate more there to make some changes that will reduce these numbers. So to minimize these numbers of diabetes, we researched various algorithms and methods. The proposed method focuses on extracting the attributes that gives a result in early detection of Diabetes Mellitus in patients. Various existing processes provide just a result as the patient has diabetes or not which will require the patients to visit a diagnostic centers or to a doctor. So we proposed a system based on deep learning approaches that will help to solve a serious problem. These systems take collaborative inputs from dataset to give prediction with random forest algorithm which gives more accurate results.

Keywords— Diabetes Prediction, Random forest, Deep Learning, Convolutional neural network (CNN), Support Vector Machine (SVM), Data transformation.

1 Introduction

Diabetes is a common physiological health problem among humans across gender and age. Diabetes is an illness when sugar levels are high and diabetes is particularized by high blood glucose levels. This is known as hyperglycemia. In this case, the body is incompetent to produce enough insulin. Another perspective is the body cannot respond to the insulin produced. So diabetes is irretrievable; it has to be controlled.

Various types of diabetes may happen and dealing with the circumstance of diabetes relies upon the sort. All kinds of diabetes are not the cause of a person's overweight or lifestyle. In fact, some people have diabetes even before their childhood. Diabetes types are type 1 and type 2 diabetes. Type 1 diabetes which is additionally called adolescent diabetes, this sort happens when a body neglects to create insulin do type 1 diabetes people groups are insulin-subordinate. Type 2 diabetes is the most well-known diabetes as per the National Institute of Diabetes and Digestive and Kidney Diseases. Right now diabetes, the body can't utilize insulin effectively. That invigorates your pancreas to deliver more insulin until it can never again stay aware of interest. Insulin creation diminishes, which prompts high glucose. Different sorts are Gestational diabetes and prediabetes. In Gestational diabetes, this sort happens in ladies during pregnancy when the body can turn out to be less delicate to insulin. Gestational diabetes doesn't happen in all ladies and as a rule settle subsequent to conceiving an offspring. Prediabetes is a genuine well-being condition where glucose levels are higher than ordinary, yet not sufficiently high yet to be analyzed as type 2 diabetes.

A diabetes person can develop some hard complications namely heart attack and kidney failure. Globally, an estimated 463 million adults are living with diabetes, according to the latest 2019 data from the International Diabetes Federation. Diabetes prevalence is increasing rapidly; previous 2017 estimates put the number at 425 million people living with diabetes [1]. According to statistics in 2017, an estimated 8.8% of the global population has diabetes. This is likely to increase to 9.9% by the year 2045 [2]. India had a greater number of diabetics than some other nations on the planet, as per the International Diabetes Foundation. Diabetes as of now influences in excess of 62 million Indians, which is over 7.2% of the grown-up populace.

Right now, we will use the methods of deep learning, specifically - deep learning neural systems and convolutional neural systems, to propose a model for diabetes anticipation with high precision. Deep learning alludes to machine learning or it is a sub-some portion of machine learning. It is completely subject to the degree of

[*] e-mail: kirti50hirnak@gmail.com
[* *] e-mail: nikitachaudhari863@gmail.com
[* * *] e-mail: akshaysingh.mustang@gmail.com
[* * * *] e-mail: deepatli.patil@rait.ac.in

learning spoken to, significant level ideas are characterized from the low-level ideas, compared to the chain of importance of highlights, components, or ideas and the other way around [3]. It is different degrees of learning of portrayal, deliberation and breaking down the information that assists with understanding the information to be specific sound, and content, pictures. A deep learning structure comprises a multilayer perceptron with concealed layers.

The Random forest algorithm is the most acclaimed and it is an essential AI calculation. Random forest strategies give better accuracy and force. It create distinctive decision trees. Another algorithm used is Support Vector Machine. SVM is a lot of supervised AI models that are utilized in classification and relapse. The target of the support vector machine algorithm is to discover a N-dimensional space that unmistakably groups the information focuses.

2 Review Of Literature

Numerous works have been done related to the prediction of diabetes using different data mining techniques. The datasets, algorithms, methods used by the authors and observed results along with the future scope are carried out in finding out efficient methods of medical diagnosis for various types of diabetes diseases. Here research on diabetes prediction and detection has been carried out for several years.

Nahla Barakat. [4] discussed and designed an Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. They did research on the diagnosis of diabetes mellitus by using the SVM algorithm and Rule-Extraction from SVMs. They used a dataset named as “National Survey of Diabetes data collected in Sultanate of Oman” which contains various types of samples like Fasting Venous, Oral Glucose Tolerance Test (OGTT), etc. The result of these algorithms predicts more accurate results, but there is a need for pre-prediction of diabetes to prevent or delay that disease.

Riccardo Miotto. [5] proposed the prediction of diabetes from the electronic health record. Utilized a novel framework which is called “deep patient”. It was used to represent patients by a group of their general features, from a large-scale EHR database through a deep learning approach.

Ioannis Kavaliotis. [6] in this computational and structural journal, they covered aspects such as Prediction and Diagnosis, Diabetic Complications, Generic background environment, Health care, and management. They mentioned in their conclusion as in-depth exploration towards diagnosis, etipathophysiology, and treatment of diabetes mellitus have to do to prevent such disease.

Beata Strack. [7] proposed a framework dependent on the ’Wellbeing Facts database’ a national information stockroom that gathers exhaustive clinical records. This dataset contains information efficiently gathered from taking an interest establishment’s electronic clinical records and incorporates experience information (like a crisis, outpatient and inpatient), supplier strength, socioeconomics (like age, sex, and race), and determination and in-medical clinic mortality and emergency clinic attributes. This information is an implantation of 10 years of clinical consideration at 130 medical clinics. On this dataset they set some extraction criteria for e.g., length of the patient’s stay at the medical clinic, research facility test performed, and so on as per these criteria they pick 30 days for readmission of patients to that emergency clinic. In that, they found a noteworthy connection between readmission likelihood of patients and HbA1c estimations. Toward the end they got the end as the estimation of HbA1c is required to lessen the readmission rates by giving legitimate prescriptions to every patient as per their HbA1c esteem.

Muhammad Atif Iqbal. [8] used data mining techniques that are widely used for the prediction of disease at an early stage. They performed prediction of diabetes using significant attributes with their different feature relationships are also featured. Used a variety of different tools to decide proper characteristic determination and for bunching, expectation, and affiliation rule digging for diabetes. Huge factor determination was made by the vital segment investigation technique.

Trang Pham. [9] presented DeepCare, an end-to-end deep dynamic memory neural network for personalized healthcare. This model is used to minimize extraction of required attributes from a huge dataset. DeepCare peruses clinical records, remembers ailment directions and care forms, appraises the present sickness states, and predicts the future hazard.

Akm Ashiquzzaman. [10] utilized the Neural Network and Deep Learning, an expectation framework for the malady of diabetes is introduced where the issue of overfitting is limited by utilizing the dropout strategy. A deep learning neural system is utilized where both completely associated layers are trailed by dropout layers. This model is utilized for estimating the sickness of diabetes. An epic type of profound neural system for diabetes guess with expanded precision. The yield execution of the proposed neural system appears to have beaten other condition-of-craftsmanship techniques and it is recorded as by a wide margin the best execution for the Pima Indians Diabetes Informational index.

Suyash Srivastava. [11] discussed machine learning, a part of Computerized reasoning is utilized to dissect and make the diabetes expectation model. Different scientists have likewise been done to anticipate the diabetes AI calculation, however this is an extra exertion in the exploration work dependent on a particular kind of patient in a particular network. Fake Neural System (ANN) was picked for building a model to foresee diabetes. This model is perfect for foreseeing the chance of diabetes with 92% exactness while trying with the example test information. This model can accomplish more exactness on the off chance that it trains with enormous examples preparing information later on.

But still, there is a need to predict diabetes at an early stage so that it will help us to prevent diabetes mellitus to occur in the future in the patient’s life. That motivates us to build a system which predicts diabetes earlier.

3 Proposed System

Diabetic is a condition that occurs when blood glucose is too low and sugar level is too high. We proposed a model in predicting diabetes using deep learning. Deep learning is the kind of machine learning mainly based on Artificial Intelligence. The hidden layers of deep learning system do all these totally inside itself without including incidental specialists. The short portrayal of deep learning systems is given as underneath [2].

Initially, we did an analysis of existing systems that are available, but after analysis we came to know that there is a need for some real-time system that may overcome the drawbacks of an existing system. Then we searched for the dataset on which algorithms may performed. When enough information was gathered we began preprocessing information into a format that can be utilized for training purposes. When the information has been processed we start the training of our model with 70% of the complete information staying 30% will be used for testing purposes. When training of our model is finished we begin making a prediction model which will accept a few parameters as input and dependent on learning from the training informational index will give a prediction about missing parameters.

For prediction, we have used two different Algorithms Random Forest Algorithm and Support Vector Machine. In Random Forest we follow - select random samples from a given dataset, construct a decision tree for each sample and get a prediction result from each decision tree, perform a vote for each predicted result, select the prediction result with the most votes as the final prediction. While SVM follows - Generate hyperplanes which segregate the classes in the best way.

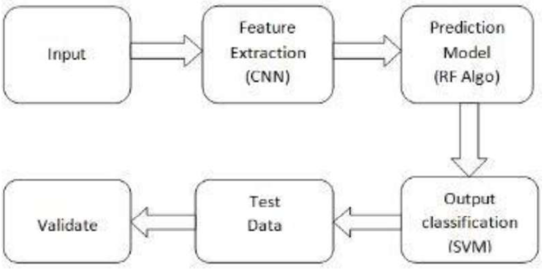

Figure 1. System flow diagram for the Deep Learning Model

Algorithms used to build the proposed system are described as follows :

3.1 Convolutional Neural Network (CNN)

A convolutional neural system (CNN) is a particular kind of artificial neural system that utilizes perceptrons, an machine learning unit calculation, for supervised learning, to break down information. CNNs have as of late been broadly utilized for some characterization and acknowledgment assignments too. In this manner, we chose these algorithm to assess their presentation for diabetes recognition to build up a choice emotionally supportive network. Extensive experiments were performed on our data set. CNN is generally made up of an input, an output layer and many hidden layers. There are 9 input layer neuron and 2 output layer neuron in proposed system. In the hidden layer 1 14 neurons and hidden layer 2 14 neurons.

3.2 Random Forest Algorithm

The Random forest algorithms is the most famous and it is a basic machine learning algorithm. Random forest techniques give better precision and power. More often than not the irregular backwoods creates better outcomes yet it is hard to enhance its exhibition and furthermore hard to deal with different sorts of information including numerical, ostensible and double information. Random forests develop different choice trees [8]. Random forest is one sort of Directed learning strategy which utilizes a gathering learning technique for characterization and relapse [16]. In Random forest, trees are run in equal so there is no collaboration between these trees while building the trees. In the Random forest technique consolidates the aftereffects of different expectations that total numerous choice trees, which is the reason it is known as a meta-estimator.

3.3 Support Vector Machine (SVM)

The Support vector machine classifies groups of data with the help of hyperplanes. It is good for binary classification of x versus other variables. SVM is a set of supervised machine learning models that are used in classification and regression. For example, if it is given two training classes, then the SVM classifier divides two clusters in a better way. For generalization purpose, one cluster should not lie closer to the data points belonging to the other cluster. It should be far away from other clusters of each type. The point occurs nearest to the margin of the classifier known as support vectors.

4 Methods and Materials

4.1 Data Preprocessing

Data preprocessing is the most important process in the data mining process. Data gathered from various sources has many missing values, null values, faulty values. These create more problems while analyzing the data. If more irrelevant and noisy data is present then it becomes difficult to apply any data mining or machine learning algorithm on such datasets. So to perform data preprocessing various tasks are performed named as: Cleaning, integration, transformation, reduction, and discretization [12].

4.2 Data cleaning

Data cleaning process comprises filling the missing qualities and expelling loud information. Loud information includes exceptions which are expelled to determine irregularities [13]. In the dataset, there are various parameters like weight, glucose, BMI, etc., which contains value as a zero. These values need to be replaced. Data cleaning replaces such values to some median values in the dataset.

4.3 Data reduction

Data reduction gets a diminished portrayal of the dataset, it is a lot liter in volume yet delivers the equivalent (or nearly the equivalent) result. Dimensionality reduction is utilized to diminish the quantity of characteristics in the dataset [14]. This task of preprocessing is used to extract required or proper attributes to the system.

4.4 Data transformation

Data transformation comprises smoothing, standardization and aggregation of information [15]. The dataset we are using has a lot of non numeric values that need to be numeric while training and model form. Converting non-numeric features into numeric. Matrix multiplication is not possible on a string, so we must convert the string to some numeric representation. Resizing inputs to a fixed size. Feed-forward neural networks and linear models have a fixed number of input nodes, so your input data always has the same size.

5 Results

The dataset used in this paper is originally taken from [7]. Wellbeing Facts is an unconstrained program offered to associations which utilize the Cerner Electronic Health Record System. The database contains parameters like age, gender, and patient ID glucose, hba1c, medical diagnosis, blood samples, etc. having total 50 parameters, which are used to predict diabetes earlier. To improve the prediction and to build a system more user-friendly we consider another dataset also which contains parameters of a patient's daily routine, diet plans, eating habits, etc. having a total 11 parameters. The combination of these parameters with previous dataset parameters gives output in a more appropriate way. This second dataset is taken from "Japan Medical Data Center (JMDC)", which is used in a medical research article [18].

Third dataset formed from the combination of above two datasets contains parameters namely Sex, Age, Blood Cholestrol, Heridatory, Smoking, Alcohol Intake, Physical Activity, Diet, Obesity, Stress, Glucose, Blood Pressure, Skin Thikness, Insuline, BMI, Medical Diagnosis and HbA1c. These are a few parameters which are really making numerous hazardous infections individuals nowadays.

Various deep learning algorithms are applied to the dataset that we used in our system and the results we get gives more accuracy than algorithms that were used in previous systems or research papers. The algorithms we ap- plied are SVM and Random Forest which build a model for training and testing separately for two separate algorithms. As a result, those models give two separate results but according to their accuracy system gives the prediction.

Table 1. CONFUSION MATRIX OF SVM ALGORITHM

Actual/Predicted	Positive	Negative
True	135	1
False	3	61

Table 2. CONFUSION MATRIX OF SVM ALGORITHM (SIGMOID MODEL)

Actual/Predicted	Positive	Negative
True	126	10
False	5	59

Table 3. CONFUSION MATRIX OF RF ALGORITHM

Actual/Predicted	Positive	Negative
True	198	0
False	0	102

The performance of SVM and RF models can be represented by the confusion matrix on the dataset. The confusion matrix shown in Table.

SVM is a supervised machine learning model. Sup port vector machine is used to evaluate the accuracy. Results obtained from these algorithms are represented in a tabular way in the following tables. Table I represents the confusion matrix of a simple SVM algorithm. It splits the dataset into 80% of the training set and 20% testing set. Thus we get an accuracy of a normal SVM algorithm is upto 97%.

Table II represents the confusion matrix of the SVM algorithm with the sigmoidal model. As this algorithm is used for classification in our system it gives the accuracy upto 97%. SVM is used for the data points that need to be classified into groups which belong to the same groups called clusters.

Then we used the Random Forest model for predicting diabetes disease. Table III represents the confusion matrix of the RF algorithm. In the confusion matrix, we included TP, TN, FP, FN in that the case of true positive, the diabetes disease is predicted and the case of true negative, diabetes has not predicted. It splits the dataset into 70% of the training set and 30% testing set. Thus we get an accuracy of an RF algorithm us upto 100%.

6 Conclusion

Diabetes mellitus is quickly developing as one of the best worldwide wellbeing difficulties of the 21st century. Diabetes is a heterogeneous gathering of ailments. It's portrayed by the constant height of glucose in the blood. We identified and reviewed machine learning and deep learning approaches applied on diabetes research. We developed a web based system which gives the result that a patient may have diabetes or not in the future. We have arranged our preparation diabetes information to be dissected; utilizing a couple of lines of Python, we have prepared a model that is fit for anticipating whether an individual is probably going to have diabetes, giving a productive way to use clinical assets to recognize and treat the most noteworthy level of patients with diabetes. Results decide the sufficiency of the structured framework with an accomplished exactness of 97% utilizing Support Vector Machine (SVM) and by using the Random Forest Algorithm accomplished accuracy upto 100%, which provides a system with more accurate results. The advancement to this work will be suggestions that help patients to consult a doctor or some improved diet provided by system will be suitable.

References

[1] Hammoudeh, Ahmad and AlNaymat, Ghazi and Ghannam, Ibrahim and Obeid, Nadim, ”Predicting Hospital Readmission among Diabetics using Deep Learning”, The journal of The 5th International Symposium on Emerging Information, Communication and Networks, 2018.

[2] Goutham, Swapna and R, Vinayakumar and Kp, Soman, “Diabetes detection using deep learning algorithms”, ICT Express. 4. 10.1016/j.icte.2018.10.005, 2018.

[3] Liu, Tianyi and Fang, Shuangsang and Zhao, Yuehui and Wang, Peng and Zhang, Jun.. “Implementation of Training Convolutional Neural Networks”, Published on research gate article, 2015.

[4] N. Barakat, A. P. Bradley and M. N. H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus," in IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 4, pp. 1114-1120, July 2010.

[5] Miotto, Riccardo and Li, Li and Kidd, Brian, “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records” Scientific Reports. 6. 26094. 10.1038/srep26094, 2016.

[6] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vla-Havas and Ioanna Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research” Computational and Structural Biotechnology Journal, 2017.

[7] Strack, Beata and Deshazo, Jonathan and Gennings, Chris and Olmo Ortiz, Juan Luis and Ventura, Sebastian and Cios, Krzysztof and Clore, John, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records”, BioMed research international. 781670. 10.1155/2014/781670, 2014.

[8] Mahboob, Talha and Iqbal, Muahammad and Ali, Yasir and Wahab, Abdul and Ijaz, Safdar and Baig, Talha and Hussain, Ayaz and Malik, Muhammad and Mehdi, Muhammad and Ibrar, Salman and Abbas, Zunish, ”A model for early prediction of diabetes. Informatics in Medicine Unlocked”, 16. 100204. 10.1016/j.imu.2019.100204, 2019.

[9] Pham, Trang and Tran, Truyen and Phung, Dinh and Venkatesh, Svetha, “Predicting healthcare trajectories from medical records: A deep learning approach. Journal of Biomedical Informatics”, 69. 10.1016/j.jbi.2017.04.001, 2017.

[10] Ashiquzzaman, Akm and Tushar, Abdul Kawsar and Islam, Dr. MD Rashedul and Shon, Dongkoo and Im, Kichang and Park, Jeong-Ho and Lim, Dong-Sun and Kim, Jongmyon, “Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network”, Published on research gate article, 2017.

[11] Srivastava, Suyash and Sharma, Lokesh and Sharma, Vijeta and Kumar, Ajai and Darbari, Hemant, “Prediction of Diabetes Using Artificial Neural Network Approach”, ICoEVCI 2018, India. 10.1007/978-981-13-1642-5-59, . 2019.

[12] Benhar H, Idri A, Fernández-Alemán J., “Data preprocessing for decision making in medical informatics: potential and analysis”, World conference on information systems and technologies.p. 1208–18, 2018.

[13] Abidin NZ, Ismail AR, Emran NA., “Performance analysis of machine learning algorithms for missing value imputation”, Int J Adv Comput Sci Appl ; 9:442–7, 2018.

[14] Liu, Huan and Motoda, Hiroshi , “Feature Selection for Knowledge Discovery and Data Mining”, Kluwer Academic, USA. 10.1007/978-1-4615-5689-3, 2000.

[15] Malley B, Ramazzotti D, Wu J T-y, “Data preprocessing. Secondary analysis of electronic health records”, Springer;. p. 115–41, 2016.

[16] VijayaKumar, K. Lavanya, B. Nirmala, I. Caroline, S, “Random Forest Algorithm for the Prediction of Diabetes”, 1-5. 10.1109/ICSCAN.2019.8878802, 2019.

[17] Kaur, Harleen Kumari, Vinita, “Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach. Applied Computing and Informatics”, 10.1016/j.aci.2018.12.004, 2018.

[18] S. U. Amin, K. Agarwal and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," 2013 IEEE Conference on Information Communication Technologies, Thuckalay, Tamil Nadu, India, pp. 1227-1231, 2013.

[19] Balaji, H. Iyenger, N Ch Sriman Narayana Caytiles, Ronnie., “Optimal Predictive analytics of Pima Diabetics using Deep Learning”, International Journal of Database Theory and Application. 10. 47-62. 10.14257/ijdta..10.9.05, 2017.