International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

IJIEMR Transactions, online available on 20th Jan 2022.
Link : https://ijiemr.org/downloads/Volume-11/Issue-01

Title: **An automatic RBF ANN and PSO algorithm-based diabetes detection application**

volume 11, Issue 01, Jan: 182-190

Paper Authors: **M.Dileep Kumar, Dr.Sateesh Nagavarapu, N.Pavan.**

An automatic RBF ANN and PSO algorithm-based diabetes detection application

[1]**M.Dileep Kumar, Asst.Prof.** Prof, Dept of CSE, Mallareddy Institute of Technology, Post Via Kompally, Maisammaguda, Dulapally, Secunderabad, 500100

[2]**Dr.Sateesh Nagavarapu, professor.** Prof, Dept of CSE, Mallareddy Institute of Technology, Post Via Kompally, Maisammaguda, Dulapally, Secunderabad, 500100

[3]**N.Pavan, Asst. Prof.** Prof, Dept of CSE, Mallareddy Institute of Technology, Post Via Kompally, Maisammaguda, Dulapally, Secunderabad, 500100

Abstract

Patients with diabetes mellitus have blood sugar levels that are consistently excessively high. This illness, which affects a broad variety of organs in the human body, targets blood vessels and nerves in particular. Many lives may be saved by early detection and treatment of diseases like these. In order to reach this aim, this study makes use of machine learning methodologies to analyze multiple risk indicators associated with this condition. 'Diabetes patients' diagnostic medical data sets may be utilized to develop prediction models for future outcomes, using machine learning techniques. Such data might be used to predict diabetes cases. Predicting diabetes using four well-known machine learning techniques based on adult population data: Support Vector Machine (SVM), Naive Bayes classifier (NBC), K-Nearest Neighbor classifier (KNN), decision tree (C4.5), neural network with cluster validity index (CVI), and genetic algorithm (GA)..

Keywords: Diabetes Mellitus, RBF Neural Network, Genetic Algorithm, Machine Learning

1. INTRODUCTION

The lack of insulin hormone in the human body is the root cause of diabetes, a metabolic and genetic illness. The hormone insulin is crucial in the process of converting food into usable energy. Excess sugar levels in the blood are caused by a deficiency of insulin. As a result, blood glucose levels in diabetics tend to be higher than normal. Diabetes Mellitus is the medical term for diabetes (DM). High blood sugar levels, frequent urination, and an increase in appetite and thirst are all signs that you may be suffering from this condition. Many individuals throughout the world are being affected by diabetes, which is on the rise at an alarming rate. When blood sugar levels are too high, damage to blood vessels may occur and lead to a variety of health issues, including heart disease and stroke. Diabetes is a leading cause of Non-Communal Disease (NCD) fatalities in the global population, according to World Health Organization (WHO) data. More often known as "diabetes," which is a long-term condition in which blood sugar levels are unusually high. Diabetic symptoms may be caused by either a lack of insulin synthesis (the hormone produced by the pancreas) or a lack of cell sensitivity to insulin.

Two of the most common types of diabetes that result in these two processes are insulin-dependent (type1) and non-insulin-dependent (type2).

In type 1 diabetes, insulin levels are low or non-existent. Type 2 diabetics often have adequate insulin, but the cells on which it is supposed to work aren't always sensitive to it. Both types of diabetes have the same symptoms: increased urine output, a decreased appetite, and weariness. Glycosylated hemoglobin levels, blood glucose levels, and the glucose tolerance test are all used to diagnose diabetes (hemoglobin A1C). Depending on the type of diabetes, there are many treatment options. Diabetic consequences include dangerously high or abnormally low blood sugar, as well as blood vessel disease, which may cause damage to the heart, eyes, kidneys, nerves, and other parts of the body. In this work, multiple machine learning approaches are used in conjunction with numerous risk variables related with diabetes to predict the condition.

Details of the software:

1. Python3 programming language

2. SQL, NLP and python libraries are used in this project (like Scipy, Sklearn)

Specs about the hardware

1. To begin with, choose a device.
2. Keep an eye on the results.
3. In addition, a web browser is required**.**

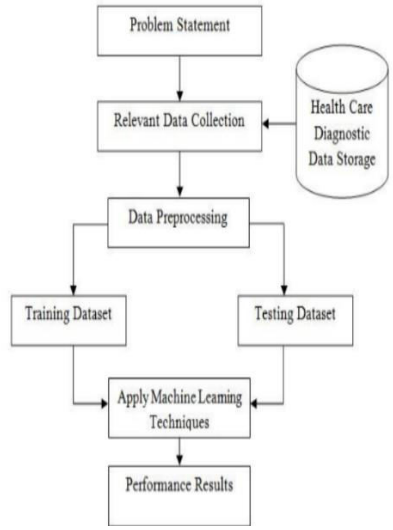


Fig.1: Block diagram for Predicting Diabetes Mellitus

Many researchers have used machine learning approaches to extract information from accessible medical data in the study of diabetes. It was shown that the J48 machine learning algorithm gives superior performance and accuracy than other algorithms before preprocessing, according to a recent research. Because of cross-validation, classification methods were not more efficient. Data mining methods such as IB1, Naive Bayes, and C4.5 on an Ulster Community and Hospitals Trust dataset were used to predict and manage diabetes (UCHT). The performance of IB1 and Naive Bayes was improved by using the feature selection approach. It has been used to predict diabetes illnesses using real-world data sets collected by the distributed questioner (ANN, Logistic regression, and J48). Finally, J48 machine learning approaches were shown to be more effective and accurate than the competition.

2. RELATED WORK

Classifier Models for Predicting Diabetes Mellitus in 2015: A Performance Analysis:

Four prediction models for diabetes mellitus based on eight significant features are compared to each other in two separate scenarios. There are two stages in the data preparation process. According to the research, the decision tree J48 classifier is more accurate than the other three classifiers, with a precision rate of 73.82%. We get more accurate findings in subsequent studies that have been pre-processed, compared to the initial one. Compared to the other three classifiers, KNN (k=1) and Random Forest are the most accurate, with a classification accuracy of 100 percent. Models did not perform well with noisy data in this technique.

Classification Algorithms on the Diabetes Dataset-2016:

It is the kind of data we utilize as input that has the most impact on algorithm output and model creation time, but the data mining techniques we use are also impacted by the dataset's architecture.

Predicting outcomes is best done using classification approaches, according to a review of several methods. False alarms and poor detection rates plague the diagnosis of illness, despite the use of a variety of approaches.

Type-2 Diabetes Detection via Electronic Health Records in 2017: A Machine Learning-Based Framework A pilot research was conducted to identify patients with and without type 2 diabetes (T2DM) based on EHR data, and this work suggested an accurate and efficient approach. T2DM patterns may be extracted using machine learning, which can then be boosted by overcoming the vast variety of examples and controls in expert methods, thereby enhancing the framework's prediction potential. It was shown that this system can identify people with and without T2DM with an average AUC of roughly 0.98, greatly surpassing the state-of-the art at an AUC of 0.71.

3. PROCEDURE AND METHODOLOGY

Data Acquisition

The Pima Indians diabetes database has been used to predict diabetes. The data comprises 768 rows and 9 columns of characteristics such as glucose, BMI, age, pregnancies, insulin, skin thickness, and diabetes pedigree function, blood pressure.

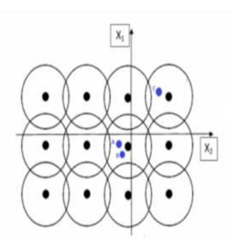
Data preprocessing

We use the term "preprocessing" to describe the process of turning raw data into something our system can comprehend and handle more quickly. It is common for real-world data to be inconclusive, inaccurate, and lacking in precise habits or trends. A proven solution to this problem is data preparation. It's common for preprocessing to include things like eliminating or replacing null values, encoding values, and deleting characteristics that are strongly linked or redundant. Before training the model, the work of standardization is accomplished by using the standard scalar as

Standard Scalar. The magnitude of variables in a dataset might vary widely. The AGE column in the Employee dataset, for example, will have values ranging from 20 to 70, whereas the SALARY column would have numbers ranging from 10,000 to 80,000. In order to develop a machine learning model with a consistent scale, the scales of these two columns are standardized.We use the term "preprocessing" to describe the process of turning raw data into something our system can comprehend and handle more quickly. It is common for real-world data to be inconclusive, inaccurate, and lacking in precise habits or trends. A proven solution to this problem is data preparation. It's common for preprocessing to include things like eliminating or replacing null values, encoding values, and deleting characteristics that are strongly linked or redundant. Before training the model, the work of standardization is accomplished by using the standard scalar as Standard Scalar. The magnitude of variables in a dataset might vary widely. The AGE column in the Employee dataset, for example, will have values ranging from 20 to 70, whereas the SALARY column would have numbers ranging from 10,000 to 80,000. In order to develop a machine learning model with a consistent scale, the scales of these two columns are standardized.

Training the Data

Algorithms are trained using training data. Both the training and testing sets typically make up a specific proportion of the total dataset. Training and testing data sets are separated. In order to make sure the algorithm used to train the machine is more precise and effective, training data is used. The algorithm or classifier often performs better when the training data is better. The activation function of the neuron processes the incoming signal to create an output. When a hidden neuron is activated by an input vector, its activation function is defined as (X), which means that given an input vector X, its output will be (X). 1-D input x generates a Gaussian neural activation function with centre (mean). Gaussian neural nodes divide the feature vector space into regions, each of which generates a signal corresponding to an input vector and whose strength is determined by the distance between its centre and the input vector. For inputs that are near in Euclidian vector space, the output signals must likewise be comparable. A neuron's reaction to X Gaussian neural nodes is referred to as a neuron's receptive field, and a neuron's receptive field is represented by the border of circles. There are 12 Gaussian nodes dividing up this 2-dimensional vector space. RBF may select how to react based on the combination of the activations of the collective system of neurons from each input vector. A and B will produce identical output signals from the neurons in the aforementioned setup, however C will provide very distinct output signals.



International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Fig 2.1(a): Gaussian Nodes

The center of each hidden neuron is represented by the weights linking the input vector to that neurons' hidden counterparts in RBF architecture. In order to train the network, weights linking hidden neurons to output neurons are set in advance such that the whole space is covered by their receptive field.

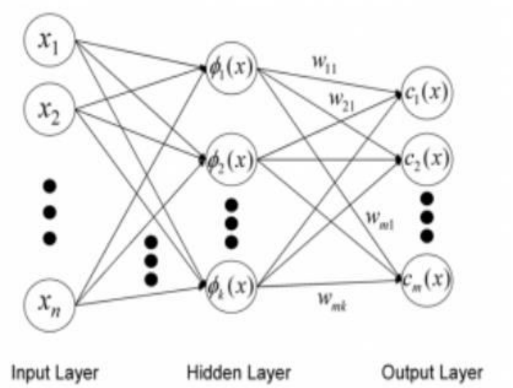


Fig 2.2(b): Radial Basis Neural Network Layers

- K Is for Select the number of cluster centers "K" in the clustering process.
- Randomly choose K locations from the dataset and designate them as the K centroids of the data.
- Find the centroid that is closest to each point in the dataset.
- For each centroid, compute the average of the points closest to the centroid.
- The associated centroids to their corresponding mean values, as seen in (4).
- Once you've made it to (3), keep going until you see a confluence.

In order to encompass the whole input vector, neurons have a receptive field that covers the complete range of their receptive fields. In other words, the greatest "d" distance between two hidden neurons is used to determine the value of sigma. Where M is the total number of hidden neurons, and d is their greatest distance from one others.

$$\sigma = \frac{d}{\sqrt{2M}}$$

Analyzing Data from Various Machine Learning Classifiers Using Trained Data

Diabetes may be predicted using a variety of machine learning approaches. However, choosing the optimal method to forecast based on these characteristics is quite challenging. So we are using trained data on SVM, Naive Bayes, decision tree, and k-means and is made accessible for testing. However, we are also using an optimized radial basis neural function as it is not relevant to the model we are implementing the entire logic, preprocessed data is followed by making clusters of data using k-means clustering logic, then applying Gaussian function as kernel function to optimize the clusters formed, and to name clusters class by class we are using Dunn, Davies Bouldin, silhouette validity indexes and computing inverse-weight matrix to train the model. We are employing a genetic algorithm to know the correctness of the model when just the best features picked from the input are delivered to the output.

Algorithm

- Use "K" to indicate the number of neurons that are concealed.

International Journal for Innovative
Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

- K-means clustering was used to determine the locations of RBF centers.
- Use the equation to determine (2)
- Use the equation to determine the RBF node's actions (1)
- Equations may be used to train the output (3)
- Training the Data

The model's performance may be assessed using a set of performance measures, which are recorded in the test set. In this instance, we're focusing on precision.

4. IMPLEMENTATION AND RESULTS

Dataset Pre – processing and Training the Data

In order to apply a model, you must first load the necessary libraries, import the dataset, deal with missing data, handle categorical data, divide the dataset into training and testing subsets, and then apply feature scaling to the dataset. The dataset for diabetes prediction has been processed and trained, as can be seen in the accompanying image.

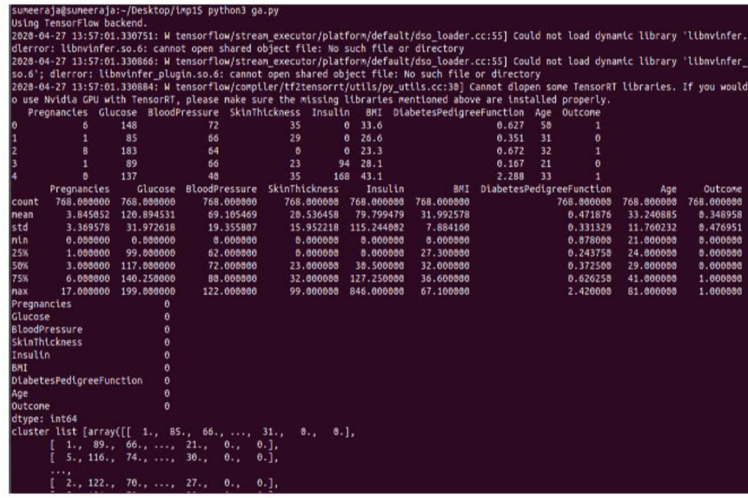


Fig.3: Description of The Diabetes Dataset

Applying Clustering Techniques

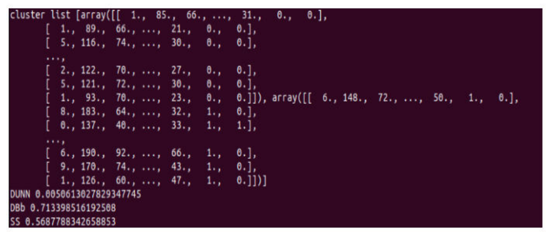


Fig 4: Indexing To Clusters

Here, clustering is utilized to name the clusters for use in RBFN, an unsupervised machine learning task. K-Means is shown here. In order to reduce the variation within a cluster, clustering is used to allocate instances to certain clusters. The Dunn Index (DI) is a statistic that is used in conjunction with K-Means to assess clustering algorithms. Using the Davies–Bouldin index, clustering techniques may be evaluated. Using the dataset's intrinsic quantities and properties, this internal evaluation method evaluates how successfully the clustering was done..

Constructing the RBFN and Predicting Its Accuracy

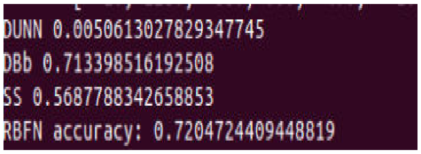


Fig 5: Accuracy of RBFNN Model

In this section, we would be predicting the RBFN accuracy by applying the K-means clustering technique implemented above.

Computing Various Machine Learning Techniques

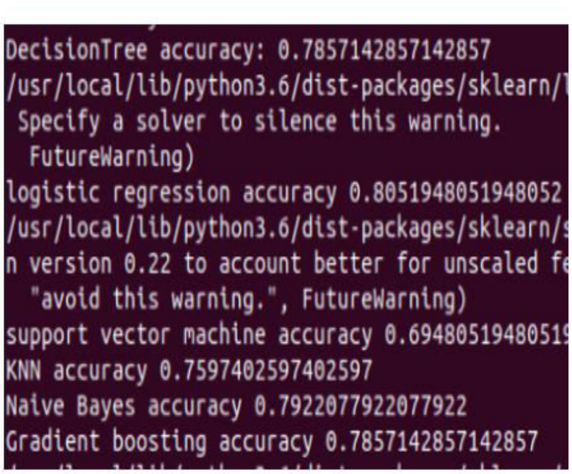


Fig 6: Accuracies of Various Machine Learning Algorithms

A variety of machine learning accuracy measures, such as gradient Boosting and Naive Bayes, SVM, Decision Tree, and Logistic Regression, are computed.

Predicting Genetic Algorithm Accuracy Using SVM Classifier as RBF kernel Function

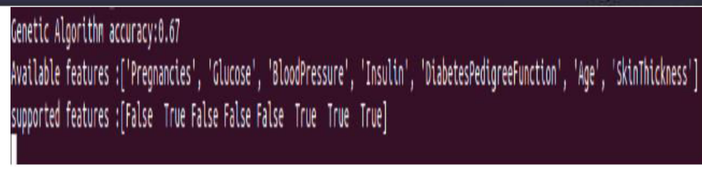


Fig 7: Accuracy of Genetic Algorithm Model

The following output predicts the Genetic Algorithm's accuracy using SVM Classifier as RBF kernel Function with gamma and cost as hyper-parameters by selecting the key features trained in the dataset.

Performance Analysis of Computed Models as a Bar Plot

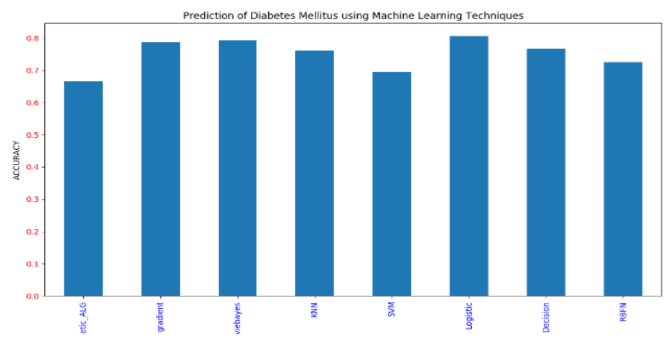


Fig 8: Bar Plot of Performance Analysis

Comparative Difference of All Computed Models

On the same dataset, different models were run, and each model produced different performance metrics. The accuracy achieved by the models is summarized in the table below.

Table 9: Performance Analysis of Various Models on Dataset

MODEL	ACCURACY
Radial Basis Neural Network Function	72%
Genetic Algorithm	67%
DUNN Index	5%
Davies Bouldin Index	71%
Silhouette Index	56.8%
Decision Tree	78.5%
Logistic Regression	80.5%
Support Vector Mchine	69%
K-Nearest Neighbors	75.9%
Naive Bayes	79%
Gradient Boosting	78.5%

5. CONCLUSION AND FUTURE SCOPE

A novel categorization model for diabetic patients was suggested in this study, which took into consideration a variety of characteristics connected to the condition. Cluster validity index and k-means clustering technique are integrated into the proposed model in Radial Basis Neural Network. For the prediction of diabetes mellitus, we have also carried out our tests on adult population data using machine learning techniques, notably Support Vector Machine (SVM), Nave Bayes (NB), K-Nearest Neighbor (KNN), and decision tree (DT). Although Naive Bayes has the best accuracy (79.2%), it presupposes class conditional independence and hence loses accuracy. Using an SVM classifier as a kernel function and an RBFNN neural model, we achieved accuracies of 67% and 72%, respectively. Diabetes patients were categorized into one of two groups (positive or negative) using our approach. K-means algorithm incorporates a cluster validity index to ensure the best possible cluster placements. The categorization process was sped up by optimizing cluster centers, which reduced network complexity. As a result, the Genetic Algorithm model used in this study is helpful in predicting diabetes with more accuracy and minimizing the likelihood of false alarms. A hybrid particle swarm optimization algorithm may be used to determine the weights of objects in the future. As a further option, we may use kernel functions like polyharmonicspline and inverse quadratic for classification.

REFERENCES

[1] Platt, John C. “12 fast training of support vector machines using sequential minimal optimization.”Advances in kernel methods (1999): 185-208
[2] John, George H., and Pat Langley. “Estimating continuous distributions in Bayesian classifiers.” Proceedings of the Eleventh Conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995.
[3] Aha, David W., Dennis Kibler, and Marc K. Albert. “Instance-based learning algorithms.” Machine learning 6.1 (1991): 37-66.
[4] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
[5] Witten, I. H. et al. (1999). Weka: Practical machine learning tools and techniques with Java implementations.
[6] Morteza, M., Franklyn, P., Bharat, S., Linying, D., Karim, K., and Aziz G. 2015. Evaluating the Performance of the Framingham Diabetes Risk Scoring Model in Canadian Electronic Medical Records. Canadian Journal of diabetes 39, 30(April. 2015), 152-156.

[7] V., A. K., and R., C. 2013. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. 3, (April. 2013), 1797-1801.

[8] Carlo, B G., Valeria, M. and Jesus, D. C. 2011. The impact of diabetes mellitus on healthcare costs in Italy. Expert review of pharmacoeconomics & outcomes research. 11, (Dec. 2011), 709-19.

[9] Nahla B., Andrew, et al. 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine, IEEE Transactions. 14, (July. 2010), 1114-20.

[10] Abdullah A. Aljumah et al., Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2, July 2013, Pages 127-136

[11] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).

[12] Zheng, Tao et al. "A machine learning-based framework to identify type diabetes through electronic health records." International journal of medical informatics 97 (2017): 120- 127.

[13] Rani, A. Swarupa, and S. Jyothi. "Performance analysis of classification algorithms under different datasets." In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pp. 1584-1589. IEEE, 2016.

[14] Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 (2015): 45-51.

[15] Y. Huang, P. McCullagh, N. Black, R. Harper, Feature selection and classification model construction on type 2 diabetic patients' data, Artificial Intelligence in Medicine 41 (3) (2015) 251–262.

[16] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or pre-diabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

[17] N V Krishna Rao,Y HarikaDevi,NShalini,AHarika,VDivyavani,Dr.N Manga Thayaru"Credit Card Fraud Detection Using Spark and Machine Learning Techniques", ICACECS2020,Machine Learning Technologies and Applications pp 163-172, Algorithms for Intelligent Systems book series (AIS),march 2021.

[18] N.V.KrishnaRao, KayiramKavitha, AshokaDeepthiManukonda, R.V.S.Lalitha, Abhishek Reddy" Semi-Equalizing Load in Multi-hop Wireless Networks" ", International Journal of Innovative Technology and Exploring Engineering' at Volume-9 Issue-1, November 2019 pp.3047-3051 (ISSN: 2278-3075).

[19] N V Krishna Rao,S. Laxman Kumar, K. Kavitha,PuduPravalika,KotteSruthi, R.V.S. Lalitha, "Fashion Compatibility using Convolutional

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Neural Networks", ACCES September,2020.