# An Enhanced Method For Diabetes Prediction Using Machine Learning Approaches

**K.S.Aparna[1], Dr. Rajiv Kannan[2]**

ME-CSE Student, Department of CSE, KSR College of Engineering, Tiruchengode,
Namakkal, Tamilnadu, India
Professor & Head, Department of CSE, KSR College of Engineering, Tiruchengode,
Namakkal, Tamilnadu, India.

**Abstract**:
Most of the people in recent years from many parts of the world are affected with kidney failure, blindness, stroke, heart attack, etc. All such things are caused due to diabetes disease. Because of improper creation and release of insulin in human body, several metabolism activities may go wrong, which in turn onset of diabetes. General classification of diabetes includes Type 1, Type 2 which are commonly occurred and Gestational diabetes during pregnancy. To detect and predict the disease in earlier stages becomes necessary to safe guard the people. The machine learning approaches have been proposed to predict diabetes using PIMA dataset. With the consideration of essential features of datasets using feature selection algorithm like genetic algorithm, the performance metric like accuracy can be improved by using several machine learning algorithms. The analysis of results shows that 96.23% accuracy by using Multilayer Perceptron approach.

**Keywords-** Diabetes, Prediction, Machine Learning, Genetic Algorithm, Accuracy.

## I. Introduction

Firstly, to understand the impact of diabetes and how it develops, this section explains what happens in the body due to diabetes. When the people eat carbohydrate foods like rice, fruit, pasta, cereal, bread, dairy products, starchy vegetables, etc the body breaks them down into glucose. The glucose moves around body and brain. Also a part of glucose is used by the cells of body and the liver where is stored as energy. The hormone called 'insulin' is required to convert glucose into energy and it is produced by beta cells in pancreas. When pancreas is unable to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycemia) and diabetes develops. High levels of sugar (glucose) in the blood stream and urine is known as "Diabetes Mellitus".

Usually Type 1 diabetes occurs in people ages below 30 years, due to damage or attack of beta cells of pancreas caused by body's own immune system, known as 'auto immunity'. So the beta cells die and are therefore unable to make insulin at a sufficient level to move glucose into the cells, causing high blood sugar (hyperglycemia). The following are the 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. Often type 1 diabetes remains undiagnosed until symptoms become severe and hospitalization is required. Left untreated, diabetes can cause a number of health complications. The recent research focuses machine learning plays a vital role in promising the improved accuracy of perception and diagnosis of disease.

The categorization of machine learning algorithms is as shown in figure 1.

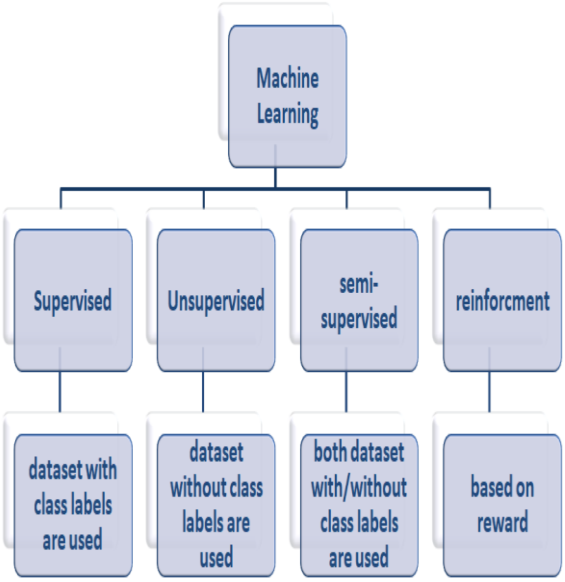
Machine Learning
Supervised
Unsupervised
semi-supervised
reinforcment
dataset with class labels are used
dataset without class labels are used
both dataset with/without class labels are used
based on reward

Figure 1. Machine Learning Algorithms

Various researches are carried out to improve the automation systems with desirable accuracy for various kinds of applications using machine learning approaches.

In general the patient must predict either to be in diabetic category or non-diabetic category. Errors in diagnosis may lead to unnecessary treatments or no treatments at all when required.

To avoid or reduce severity of such impact, there is a need to create a system using machine learning algorithm and data mining techniques which will provide accurate results and reduce human efforts.

## II. Literature Review

The researchers are developed various prediction models using variants of data mining techniques, machine learning algorithms or deep learning algorithms. The prediction models are helpful in different sectors of applications like sentiment analysis, network analysis, health care applications, stock market analysis, educational systems, fraudulent analysis, business analysis, sports analysis, entertainment, social media analysis, political analysis, weather analysis, software project development, etc.

The latest recent researchers have involved in the prediction of diabetes using machine learning approaches. The machine learning models like c4.5, KNN, Logistic regression, naïve bayes, support vector machine, decision tree, random tree, random forest, etc are used for building the diabetes prediction model. Asif Hassan Syed and Tabrej Khan (2020) developed a web application for diabetes prediction model for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia. It uses Chi-Squared test and binary logistic regression to analyze most significant diabetes risk factor for T2DM. Decision Forest model showed better average F1 score 84%.

To further improve the performance of the diabetes prediction model, the proposed system finds the most important features among all features in the dataset available in data repositories and use different machine learning algorithms. The performance metric accuracy is analysed and the results are discussed.

## III. Proposed System:

The following figure 2 shows the overview of the model used for diabetes prediction.

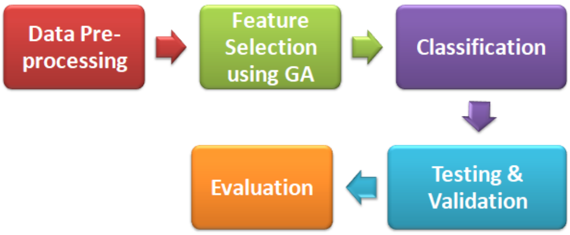

Fig. 2 Diabetes Prediction Model

The architecture diagram for diabetes prediction model includes five different modules. These modules are-

i. Data Pre-processing
ii. Feature Selection using GA
iii. Classification
iv. Testing and Validation
v. Evaluation

### i. Dataset Collection :

The diabetes dataset PIMA, available in UCI data repository is used for prediction model. The description of the dataset is as given in Table 1.

**Table 1. Dataset Information**

| Attributes | Type |
|---|---|
| Number of Pregnancies | Numerical |
| Glucose Level | Numerical |
| Blood Pressure | Numerical |
| Skin Thickness(mm) | Numerical |
| Insulin | Numerical |
| BMI | Numerical |
| Age | Numerical |
| Diabetes Pedigree Function | Numerical |
| Outcome | Class label |

i.**Data Pre-processing:** The missing data and the unknown values are reduces the performance of the prediction model. In the proposed system, the missing values are replaced with average value of that attribute. Also Minmax scaler algorithm is used to normalize the data.

**ii. Feature Selection:** More number of features leads to "overfitting" problem in the prediction model. To avoid such problem and get more accurate values, the essential attributes can only be used in the model. In the proposed system, the important features among all features have been selected by using genetic algorithm. Hence the selected four attributes from PIMA dataset are Glucose level, Skin Thickness, BMI and Age, because these attributes cannot have values zero.

**iii. Classification Algorithms used:**

**Logistic Regression (LR):** The logistic function frequently used in logistic regression is sigmoid function. It is similar to linear regression, but the calculation of the linear output is followed by a stashing function over the output. It is an S-shaped curve that can obtain any real-valued number and mapped into the values between 0 and 1.

$$1 / (1 + e\text{^-value})$$

where e denotes the base of the natural logarithms and the value is the actual numerical values that is to be transformed[5].

**K-Nearest Neighbours (KNN):** The objective of this algorithm is to identify the nearest neighbors of a given query point, such that a class label is assigned to that point. Various distance measures like Euclidean, Manhattan, Minkowski, Hamming distances are helping to form decision boundaries.

**Support Vector Machine (SVM):** This algorithm is to find a hyperplane that has the maximum margin in an N-dimensional space that distinctly classifies the data points. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane.

**Naive Bayes (NB):** This arrangement method depends on the Bayes' hypothesis. This classifier accept that the nearness of a specific component in a class is inconsequential to the nearness of some other element.

**Decision Tree (DT):** It creates a training model which can predict the class by learning simple decision rules retrieved from the training data. The splitting attribute can be selected using the criteria like entropy, information gain, gini index,etc.

**Multilayer Perception (MLP):** A multilayer perceptron (MLP) is a feedforward artificial neural network that generates outputs based on inputs. It uses backpropogation for training the network. It is widely used for solving problems that require supervised learning as well as research into computational neuroscience and parallel distributed processing. The domains like speech recognition, image recognition, machine translation, etc are using MLP.

**iv. Testing and Validation:** The model build can be used for testing with 20% testing data. The crossfold validation can be done for validating the model.

**v. Evaluation:** The performance metric accuracy is determined from the terms used in Equation 1. As given below:

Accuracy =(True Positive + True Negative) / (True Positive + True Negative + False Negative + False Positive) --(1)

**IV. Results:** The performance measure accuracy value obtained for all the classification algorithms used is as shown in Table 2.

Table 2. Accuracy comparison

| Classification | Accuracy (%) |
|---|---|
| Logistic Regression(LR) | 92.94 |
| K Nearest Neighhors (KNN) | 93.58 |
| Support Vector Machine (SVC) | 94.23 |
| Gaussian Naïve Bayes (NB) | 94.23 |
| Decision Tree (DT) | 85.89 |
| Multilayer Perceptron (MLP) | 96.79 |

In Table 1, it is shown that the accuracy of all classification methods are around 90%. K-fold cross-validation is done for all the classification approaches. The accuracy obtained in various approaches have been compared and represented in the figure 3. However, MLP classifier outperforms the other classification algorithms in the prediction of the diabetes as shown in figure 3.

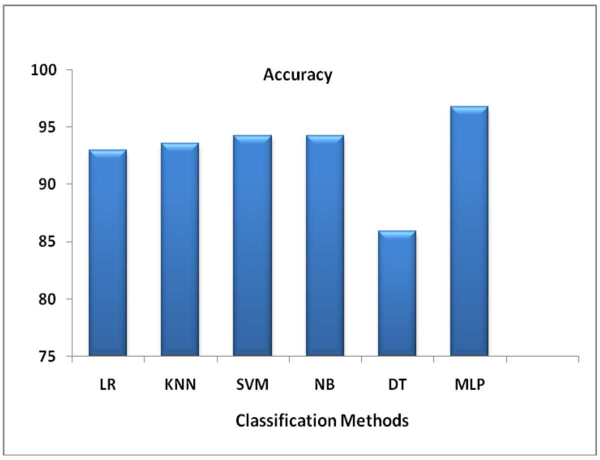

Figure 3. Accuracy for different classification methods

## VII. Conclusion

One of the most significant challenges in the health care industry is detection of diabetes in prior to severe stage. In the proposed method, data preprocessing includes scaling of data using minmax scaler and missed data is filled with average value of that feature. Using the feature reduction method, the essential features have been selected using genetic algorithms. Here four features (Glucose level, Skin Thickness, BMI and Age) out of 9 features have been selected as input features in the PIMA dataset. The proposed method used different machine learning algorithms, including LR, KNN, SVM, NB, DT and MLP for the prediction of diabetes disease and evaluated the performance using accuracy. All models show good results for accuracy. All models provided good result whereas MLP with highest accuracy of 97% outperforms other classification methods.

Further research can be carried out by using deep learning and big data algorithms to improve the performance of the system.

## VII. References

[1] https://www.kaggle.com/uciml/pima-indians-diabetes-database
[2]https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284.
[3]https://www.niddk.nih.gov/healthinformation/diabetes/overview/symptoms-causes.
[4]https://www.healthgrades.com/right-care/diabetes/is-there-a-cure-for-diabetes.
[5]https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects.
[6] S.A. Kaveeshwar, J. Cornwall, The current state of diabetes mellitus in India, Australas. Med. J. 7 (1) (2014) 45.
[7]https://www.cdc.gov/diabetes/basics/prediabetes.html.
[8] C.L. Huang, M.C. Chen, C.J. Wang, Credit scoring with a data mining approach based on support vector machines, Expert Syst. Appl. 33 (4) (2007) 847–856. [9]http://dx.doi.org/10.1016/j.eswa.2006.07.007.
[10] R Kamalraj, A Rajiv Kannan, R Ranjani, "Stability-based component clustering for designing software reuse repository", in International Journal of Computer Applications, 2011, Vol.27, Issue 3, pp. 33-36.
[11] V Vennila, A Rajiv Kannan, "Symmetric Matrix-based Predictive Classifier for Big Data computation and information sharing in Cloud" in Computers & Electrical Engineering, 2016, Vol.56, pp.831-841.
[12] R. Kamalraj, A.Rajivkannan, R. Ranjani, "Reuse Frequency Effect Classification Model (Rfec) For Classifying Reusable Software Components", International Journal of Engineering Research and Industrial Applications, 2012, pp. 299-304.
[13] J. Chaki, S. Thillai Ganesh, S.K. Cidham, S. Ananda Theertan, Machine learning and artificial intelligence-based diabetes mellitus detection and self-management: A systematic review, J. King Saud Univ. - Comput. Inf. Sci. (2020).
[15] I. Contreras, J. Vehi, Artificial intelligence for diabetes management and decision support: Literature review, J. Med. Internet Res. 20 (5) (2018) e10775.
[16] T.M. Alam, et al., Informatics in medicine unlocked a model for early prediction of diabetes, Inform. Med. Unlocked 16 (2019) 100204.

[17] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585.

[18] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Vol. 9, Frontiers in genetics, 2018, p. 515, http://dx.doi.org/10.3389/fgene.2018.00515.

[19] S. Perveen, M. Shahbaz, A. Guergachi, K. Keshavjee, Performance analysis of data mining classification techniques to predict diabetes, Procedia Comput. Sci. 82 (2016) 115–121.

[18] Asif Hassan Syed and Tabrej Khan, "Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study", IEEE ACCESS, Nov 2020, DOIL 10.1109/ACCESS.2020.3035026