IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Literature Review Of Hybrid DiabetesPrediction Model

Tejas Modi[1], Khyati Prajapati[2], Devyani Parmar[3]
[1]Student, [2]Assistant Prof., [3]Assistant Prof.
[1]Computer Engineering Department
[1] Sankalchand Patel College of Engineering, SPU Visnagar, India

Abstract: Diabetes mellitus is a global health crisis demanding improved early detection methods. This paper proposes a novel hybrid model for diabetes prediction that leverages the strengths of various techniques. The model employs Fuzzy C-Means (FCM) clustering with an adaptive kernel to group patients with similar character- istics. This approach goes beyond traditional clustering by handling potential data ambiguity. A Multi-Objective Ge- netic Algorithm (MOGA) then performs feature selection, simultaneously optimizing for relevance and redundancy, leading to a more focused dataset. Finally, an ensemble learning strategy combines multiple classifiers, enhancing prediction accuracy and robustness. This work investigates the effectiveness of the proposed model compared to exist- ing methods. It evaluates the model's performance using relevant metrics and compares it to established diabetes prediction models on benchmark datasets. The findings contribute to developing more accurate and reliable tools for early diabetes detection, potentially improving patient outcomes.

Keywords: Diabetes disease, classification, random forest, support vector machine, machine learning, signal processing, artificial neural network, deep learning, ensemble learning, early detection, and feature importance.

1. INTRODUCTION

Diabetes mellitus (DM), commonly referred to as diabetes, is a chronic metabolic disorder characterized by elevated bloodsugar (glucose) levels. It arises due to either the body's inabil- ity to produce sufficient insulin, a hormone responsible for regulating blood sugar, or the body's cells becoming resistant to insulin's effects. This chronic hyperglycemia can lead to a cascade of devastating complications, including cardiovascular disease, blindness, kidney failure, and nerve damage.

Diabetes is the most prevalent and deadly noncommunicable illness, affecting 537 million people worldwide. Diabetes can be caused by a variety of variables, including obesity, high cholesterol levels, family history, physical inactivity, poor eating habits, and so on. Increased urination is one of the most typical signs of this condition. Long-term diabetes patients may develop a variety of consequences, including heart disease, renal disease, nerve damage, diabetic retinopathy, and so on. However, if predicted early on, the danger can be lowered.

The global prevalence of diabetes has reached alarming proportions. According to the World Health Organization (WHO), in 2019, an estimated 422 million people worldwide had diabetes[15]. This number is projected to rise further, significantly burdening healthcare systems and individual well-being.
The symptoms include:

- **Increased thirst and urination:** When your blood sugar levels are high, your kidneys work overtime to remove the excess sugar from your blood. This can lead to you feeling thirsty more often and urinating more frequently

.

- **Excessive hunger:** Even though you may be eating more than usual, your cells are not getting the glucose they need for energy due to a lack of insulin or insulin resistance. This can cause you to feel excessively hungry.

- **Unexplained weight loss:** Although some people with diabetes may gain weight, especially in the early stages, others may actually lose weight without trying. This is because the body is unable to use glucose for energy and starts to break down muscle and fat tissue for fuel.

- **Fatigue and tiredness:** When your cells are not getting enough glucose for energy, you may feel tired and slug- gish.

While some individuals with diabetes may experience clas- sic symptoms such as excessive thirst, frequent urination, and unexplained weight loss, others may exhibit no noticeable symptoms in the early stages. This silent nature of the disease underscores the importance of early detection and intervention. Early diagnosis allows for the implementation of appropriate management strategies, potentially delaying or preventing the onset of complications.

Despite the availability of diagnostic tools, there is a con- stant need for improved methods for early diabetes prediction. Traditional screening methods often rely on a single diagnostic test, which may not be definitive in all cases. Developing more accurate and reliable prediction models can lead to earlier diag- noses, facilitating timely intervention and potentially improving long-term health outcomes for patients.

2. MACHINE LEARNING AND DEEP LEARNING MODELS

This section contains a brief introduction to machine learn- ing, how resourceful it is in the medical domain, and an insight into the important algorithms used by various researchers for detecting Diabetes disease.

Machine learning is a subset of artificial intelligence. It in- volves training the system to identify and learn patterns and thereafter make decisions without being explicitly programmed to do so. Conclusively, it's a process of training computers, with the aid of algorithms, to learn from data and improve their performance over time.

In the medical domain, machine learning has been pivotal. It has a profound impact and is used for:

- **Disease Diagnosis:** The ML algorithms can perform analysis over medical images (like X-rays, MRIs, and CT scans) to detect anomalies or patterns that might suggest the existence of any disease, such as cancer, tuberculosis, or abnormalities in organs.

- **Predictive Analytics:** Machine learning also features the ability to forecast patient outcomes, such as predicting the probability of the development of any particular disease or condition in patients based on their medical history, genetics, lifestyle, and environmental factors.

- **Personalized Treatment:** By observing large datasets, machine learning can suggest personalized treatment plans for patients based on their unique characteristics, outcomes of any historical treatments, or responses to medications.

- **Drug Discovery and Development:** Machine learning algorithms are also helpful in analyzing biological data, tracking down potential drug candidates, predicting drug interactions, and optimizing drug formulations, speeding up the drug discovery process.

These are a few of the many reasons for exploiting the power of machine learning to expand knowledge in the medical domain.

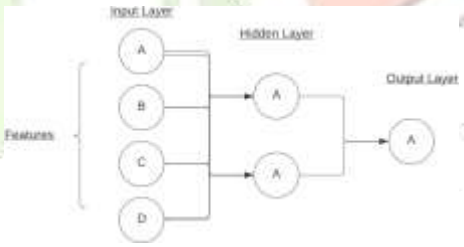
It is hoped by the researchers that once doctors can diagnose the genesis of the affected patients to a level that features prediction of the onset of the disease in a later course, those patients can be appropriately treated. At the very least, these advances could greatly delay progress.

Following is a list and a brief description of the algorithms used by the researchers for detecting Parkinson's disease:

- **Supervised Learning Algorithms:** These are the algo- rithms for which input as well as output data is well defined. The model trains itself using the input data and performance analysis as well as optimization are done after comparison with the output data. Following are a few of such algorithms:
  - Linear Regression
  - Logistic Regression
  - Support Vector Machine
  - K-Nearest Neighbour

- **Unsupervised Learning Algorithms:** Many researchers have also used unsupervised learning algorithms for preparing the machine learning model that can analyzethe signs and patterns and determine whether a person is healthy or is affected by Diabetes.
  - K-Means Clustering
  - Principal Component Analysis
  - DBSCAN (Density-Based Spatial Clustering of Appli-cations with Noise)
  - Hierarchical Clustering

- **Deep Learning Models:** The researchers have also uti- lized the power of deep learning to analyze images and medical report datasets of Diabetes-affected patients to formulate a machine learning model that predicts the existence of disease based on observable traits such as Microaneurysms, Hemorrhages, Exudates, Macular edema, to name a few
  - **Artificial Neural Networks (ANNs)** are used to cate- gorize or analyze one-dimensional inputs such as blood sugar levels or Cholestrol levels, with fully linked nodes that are changed using backpropagation. An Ar- tificial Neural Network (ANN) consists of three layers: an input layer, a hidden layer, and an output layer. In this configuration, the input layer has 4 features, the hidden layer contains 2 neurons, and the output layer has a single neuron. The network processes input data through the hidden layer to predict an output, effectively learning patterns from the input features.



[figure 1: How Neural Network Works.]

  - **Convolutional Neural Networks (CNNs)** process both two-dimensional and multi-dimensional informa- tion, such as pictures and movies. CNNs are scalable since their learning parameters are independent of in- put size. They are made up of convolutional layers that extract visual features, pooling layers that minimize feature dimensionality and computational complexity, and fully connected layers that categorize inputs based on these characteristics.

  - **Recurrent Neural Networks (RNNs)** can analyze historical blood glucose data to predict future levels. This can help identify trends and potential fluctuations, allowing for better diabetes management. RNNs are composed of CNN/ANN units, with each unit's deci- sion based on the preceding unit's condition, resulting in "short memory." Long Short-Term Memory (LSTM) networks solve the bursting or disappearing gradients problem in RNNs while preserving long-term depen- dencies in sequential data.
  - **Autoencoders** are unsupervised learning algorithms that combine an encoder and a decoder. The en- coder compresses high-dimensional input data into a lower-dimensional representation, while the decoder reconstructs the original data from the compressed form. Autoencoders are widely used for dimensionality reduction, feature learning, and data denoising. They can handle either one-dimensional data with noise, as

shown in denoising autoencoders, or multidimensional data, as seen in convolutional autoencoders. These models are particularly effective for picture reconstruc-tion, anomaly detection, and data generation.

3. LITERATURE REVIEW

Yasodhaet al.[19] classified a variety of datasets to determine whether a person is diabetic. The diabetic patient data set was gathered from a hospital warehouse, consisting of 200 instances and nine characteristics. This dataset includes two groups: blood tests and urine tests. This study implements WEKA to classify data and assesses it using 10-fold cross validation, which works well on small datasets. The results are compared. The na¨ıve Bayes, J48, REP Tree, and Random Tree algorithms are employed. J48 was found to be the most effective, with an accuracy rate of 60.2

Shubhangi M. Borkar et al.[12] explored the use of machine learning methods to identify the early onset of diabetes using the PIMA diabetes dataset. The research focused on analyz- ing the algorithms K-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forest, and XGBoost achieving an accuracy over 92%

Sachin Ahuja et al.[14] suggested the use of deep learning algorithms on the Pima India diabetes dataset to build an early prediction system. Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT) and Deep Learning (DL) scored within the range of 90–98%. Among the four of them, DL provided the best results for diabetes onset with an accuracy rate of 98.07% on the PIMA dataset.

Victor Chang et al.[5] worked upon the Pima diabetes dataset and analysed various advanced algorithms to work with IoMT effectively. In their paper the three ML algorithms that were used to analyze the Pima Indian Diabetes dataset were J48 Decision Tree, Random-Forest and Na¨ıve-Bayes. Six metrics were used to evaluate the results, including the accuracy, precision, sensitivity, specificity, F-score, and Area Under the Curve (AUC). The experimental results on the full Pima Indian Diabetes dataset shown that the Random Forest Classifier outperformed both the Na¨ıve Bayes and J48 decision tree with accuracy metric (79.57%), precision (89.40%), specificity (75.00%), f-score (85.17%) and AUC (86.24%), while the J48 had the best sensitivity (88.43%) of the three.

Jaili Gao et al. [8] in their paper modeled and predicted the Pima Indian diabetes dataset using machine learning algorithms such as KNN, decision tree, and random forest. The perfor- mance of the models was evaluated using cross-validation and confusion matrix, and the optimal diabetes prediction model was selected. The results showed that the random forest model performed the best, with an accuracy of 0.84 and an F1 value of 0.77.

Deepti Sisodia et al. [18] made systematic efforts in design- ing a system which resulted in the prediction of disease like di- abetes. During this work, three machine learning classification algorithms were studied and evaluated on various measures. These were Naive Bayes, SVM and Decision Tree. Exper- iments were performed on Pima Indians Diabetes Database. Experimental results determined the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm.

Santosh Kumar et al.[6] proposed a methodology that is implemented by Genetic Algorithm as an Attribute Selection and NBs for Classification on PIDD which has been taken from UCI machine learning repository. In Experimental studies the dataset have been partitioned between 70–30% (538–230) for training and test of NBs, GA NBs. It has been performed on PIDD and the results compared with several existing method.

Here's a brief version of the text:

Ahmad Akbar et al.[2] enhanced diabetes prediction on the Pima Indian dataset by applying several pre-processing

Measure	Training set evaluation	Testing set evaluation
Precision	0.769	0.766
Recall	0.773	0.77
F–Measure	0.769	0.767
Accuracy	77.3234%	76.9565%
ROC	0.816	0.846
Kappa statistics	0.4875	0.478
Mean Absolute Error	0.2868	0.2768
Root Mean–Squared Error	0.4157	0.3973
Relative Absolute Error	62.9039%	61.0206%
Root Relative Squared Error	87.075%	83.654%

[TABLE I: Evaluation metrics for training and testing sets]

- Santosh Kumar et al.[6]

techniques, including k-means clustering, SMOTE oversam- pling, and undersampling of minority clusters. They then utilized logistic regression for classification with 10-fold cross- validation. This approach resulted in a classification accuracy of 99.5

Md Shamim Reza et al.[17] proposed two stacking- based models for diabetes disease classification using a combination of the PIMA Indian diabetes dataset, simulated data, and additional data collected from their local healthcare facility. They used the classical and deep neural network stacking ensemble methods to combine the predictions of multiple classification models and improve accuracy and robustness. In the evaluation protocol, they used train-test and cross- validation (CV) techniques to validate the proposed model. The highest accuracy was obtained by stacking ensemble with three NN architectures, resulting in an accuracy of 95.50 %, precision of 94 %, recall of 97 %, and f1- score of 96 % using 5-fold CV on simulation study. The stacked accuracy obtained from ML algorithms for the Pima Indian Diabetes dataset was
75.03 % using the train-test split protocol, while the accuracy obtained from the CV protocol was 77.10 % on the stacked model. The range of performance scores that outperformed the CV protocol 2.23 %–12 %. Their proposed method achieved a high accuracy range from 92 % to 95 %, precision, recall, and F1-score ranged from 88 % to 96 % using classical and deep neural network (NN)-based stacking method on the primary dataset.

G. R. Ashisha et al. [7] used an ensemble voting classifier combining LightGBM, Gradient Boosting Classifier (GBC), and Random Forest (RF) to classify diabetes. The Boruta feature selection method and Random Over Sampling were applied to balance the classes and handle outliers. Tested on the Pima Indian Diabetes Dataset (PIDD) and the German dataset, the model achieved 93% accuracy on PIDD and 90% on the German dataset.

Namrata Nerkar et al. [13] explored the use of deep neural networks (DNNs) on the Pima Indian Diabetes dataset. By tun- ing hyperparameters, an accuracy of over 85% was achieved. However, they noted that DNNs could suffer from overfitting, and generalization to new datasets remains a challenge.

Pe´lagie Houngue` et al. [16] employed DNNs with 10-fold cross-validation, achieving an accuracy of 89%. The study highlighted limitations with k-fold cross-validation in the con- text of deep neural networks, suggesting alternative validation strategies for improved efficiency.

Sachin Ahuja et al. [9] proposed a deep learning model that reached an accuracy of 98.07%, outperforming other algorithms like Naive Bayes and Decision Trees. The authors suggested the inclusion of omics data to further improve the predictive power.

Fayroza et al. [11] investigated Logistic Regression, Naive Bayes, and K-nearest Neighbor (KNN) for diabetes prediction, with Logistic Regression yielding the highest accuracy of 94%. They emphasized the need for more explainable models inhealthcare applications.

Hafsa Binte Kibria et al. [4] compared Logistic Regression, Support Vector Machines (SVM), and KNN, finding Logistic Regression to be the most effective with an accuracy of 83%, followed by SVM and KNN with 82% and 79%, respectively.

G. R. Ashisha et al. [7] implemented ensemble meth- ods using Voting Classifiers, combining LightGBM, Gradient Boosting Classifier, and Random Forest. The ensemble method achieved 93% accuracy on the PIMA dataset, proving the effectiveness of feature selection via the Boruta method.

Ahmad Akbar et al. [1] improved Logistic Regression per- formance to 99.5% using SMOTE and K-Means clustering for preprocessing the Pima dataset, addressing issues like class imbalance and feature relevance.

Viloria et al. [3] applied SVMs to the Pima Indian dataset, achieving 65.6% accuracy. The study highlighted the im- portance of kernel and hyperparameter selection, suggesting further optimization with genetic algorithms for better results.

Karnika Dwivedi et al. [10] analyzed Decision Tree algo- rithms, finding J48 to be the most accurate (95.8%), compared to Naive Bayes and LAD Tree models. However, they noted the susceptibility of decision trees to overfitting, especially on
complex datasets.

4. METHODOLOGY

The development of the hybrid diabetes prediction model followed a systematic approach involving multiple stages, from data preprocessing to the application of machine learning and soft computing techniques. The following sections provide an overview of each step in the methodology.

4.1 Dataset

The dataset utilized in this study is the Pima Indians Diabetes dataset, sourced from the UCI Machine Learning Repository. The dataset comprises 768 instances and 8 features, including medical diagnostic measurements such as glucose concentration, blood pressure, body mass index, and insulin levels, along with the outcome variable indicating whether a patient has diabetes. The dataset was split into training and testing sets in a ratio of 70:30.

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to under- stand the dataset's underlying structure, distribution of features, and relationships between variables. Initial analysis included the assessment of missing values, outliers, and skewness in the data. Visualizations such as histograms, pair plots, and correlation matrices were generated to reveal patterns and insights. The target variable was carefully analyzed to ensure a balanced understanding of the outcome classes.

4.3 Polynomial Feature Engineering

To capture non-linear relationships between features, poly- nomial feature engineering was applied to the dataset. Higher- order interaction terms were generated for the original features, allowing the model to learn more complex patterns within the data. This transformation enriched the feature space and enabled the subsequent steps to explore potential interactions between the attributes more effectively.

4.4 Genetic Algorithm for Feature Engineering

A Genetic Algorithm (GA) was employed to perform feature selection and engineering. The GA optimizes the feature set by iteratively evolving a population of potential solutions. Each solution consists of a subset of features that contribute to model performance. The objective of the GA is to minimize a predefined fitness function, which in this case is based on classification accuracy. The algorithm retains only the most relevant features, reducing dimensionality and enhancing model performance.

4.5 Principal Component Analysis (PCA)

Following the application of the Genetic Algorithm, Prin- cipal Component Analysis (PCA) was used to further refine the feature set. PCA is a linear dimensionality reduction technique that transforms the selected features into a set of orthogonal components, retaining the maximum variance. By applying PCA, we ensured that the most important features are preserved while minimizing redundancy. This step reduced the computational complexity of the model and facilitated more efficient training.

4.6 Machine Learning Models and AdaBoost

In this study, we employed four weak classifiers: Random Forest, Logistic Regression, Support Vector Machine (SVM), and Artificial Neural Network (ANN). These classifiers were trained on the feature set derived from the PCA-transformed data. Each model was evaluated individually for its perfor- mance in predicting diabetes.

To enhance the overall model's predictive capability, we applied the AdaBoost algorithm, which is an ensemble tech- nique that combines the predictions of multiple weak classifiers to form a stronger, more accurate model. AdaBoost assigns higher weights to misclassified instances, enabling the ensem- ble model to focus on harder-to-predict cases. This approach leverages the strengths of individual classifiers to improve the accuracy and robustness of the prediction model.

4.7 Evaluation Metrics

The performance of each model, as well as the AdaBoost ensemble, was evaluated using several metrics, including ac- curacy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (AUC). These metrics provided a comprehensive evaluation of the models' ability to correctly classify diabetic and non-diabetic patients. Cross-validation was also employed to ensure the generaliz- ability of the model.

The methodology outlined above represents a hybrid approach to diabetes prediction, leveraging the strengths of machine learning and evolutionary algorithms to develop a robust model. Each step is designed to maximize predictive accuracy while maintaining computational efficiency.

5. CONCLUSIONS

In the realm of healthcare, diabetes stands as a critical concern, impacting millions globally with its chronic nature and potential complications. Early prediction and diagnosis of diabetes can significantly reduce its severity, enabling timely interventions that improve patient outcomes. With the rapid advancements in machine learning and soft computing, new avenues have emerged to tackle this pressing issue with greater accuracy and reliability.

The literature review of hybrid diabetes prediction models highlights the immense potential of combining various machine learning techniques and soft computing methods, such as Fuzzy C-Means Clustering, genetic algorithms, and ensemble learning. These hybrid approaches have demonstrated superior performance in terms of precision, recall, and overall predictive accuracy compared to traditional methods. The integration of adaptive kernels and multi-objective optimization further enhances the model's ability to identify intricate patterns within complex datasets, making it a valuable tool for healthcare professionals.

However, as we conclude our review, it is evident that while current models perform impressively, there is still room for further exploration and enhancement. The challenge lies not only in refining the algorithms but also in addressing issues related to data quality, interpretability, and real-world implementation. Future work must focus on developing more robust models, incorporating real-time data, and ensuring that these predictive tools are accessible and applicable across diverse populations.

Thus, this review sets the stage for ongoing research and innovation in hybrid diabetes prediction models, with the ultimate goal of improving early detection and personalized care, paving the way for a healthier future.

6. REFERENCES

[1] Rochmat Husaini Ahmad Akbar and Hari Prapcoyo. Prepro- cessing using smote and k-means for classification by logistic regression on pima indian diabetes dataset. Telematika: Jurnal Informatika dan Teknologi Informasi, 20, 2023.

[2] Ahmad Akbar, Rochmat Husaini, and Hari Prapcoyo. Prepro- cessing using smote and k-means for classification by logistic regression on pima indian diabetes dataset. Telematika, 20:238, 06 2023.

[3] Danelys Cabrera Omar Bonerge Pineda Amelec Viloria, Yaneth Herazo-Beltran. Diabetes diagnostic prediction using vector support machines. Procedia Computer Science, 170:376–381, 2020.

[4] Kibria Hafsa Binte, Matin Abdul, Jahan Nusrat, and Islam Sanzida. A comparative study with different machine learning algorithms for diabetes disease prediction. In 2021 18th Interna- tional Conference on Electrical Engineering Computing Science and Automatic Control (CCE), pages 1–8, 2021.

[5] Victor Chang, Jozeene Bailey, Qianwen Xu, and Zhili Sun. Pima indians diabetes mellitus classification based on machine learning (ml) algorithms. Neural Computing and Applications, 35, 03 2022.

[6] Dilip Choubey, Sanchita Paul, Santosh Kumar, and Shankar Kumar. Classification of pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. pages 451– 455, 11 2016.

[7] Ashisha G R, Anitha X, and Mahimai J. Classification of diabetes using ensemble machine learning techniques. Scalable Computing: Practice and Experience, 25:3172–3180, 06 2024.

[8] Na Hu and Jiali Gao. Research on diabetes prediction model based on machine learning algorithms. In 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE), pages 200–203, 2023.

[9] Naz Huma and Sachin Ahuja. Deep learning approach for dia- betes prediction using pima indian dataset. Journal of Diabetes and Metabolic Disorders, 19, 2020.

[10] Dwivedi Karnika, Sharan Hari, and Vishwakarma Vinod. Analy- sis of decision tree for diabetes prediction. International Journal of Engineering and Technical Research (IJETR), 2019.

[11] Fayroza Alaa Khaleel and Abbas M. Al-Bakry. Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80:3200–3203, 2023.

[12] Pradnya Sumit Moon, Shubhangi M. Borkar, and Shubhangii S. Shambharkar. Machine learning approach for diabetes prediction using pima dataset. In Proceedings of the 5th International Conference on Information Management & Machine Intelligence, ICIMMI '23, New York, NY, USA, 2024. Association for Com- puting Machinery.

[13] Likhit Kajrolkar Namrata Nerkar, Vaishnavi Inamdar and Rohit Barve. Diabetes prediction using neural network. International Research Journal of Engineering and Technology (IRJET), 08, 2021.

[14] H. Naz and S. Ahuja. Deep learning approach for diabetes prediction using pima indian dataset. Journal of Diabetes & Metabolic Disorders, 19(1):391–403, Apr 2020.

[15] World Health Organization. Diabetes [fact sheet], 2019.

[16] Annie Ghylaine Bigirimana Pe´lagie Houngue`. Leveraging pima dataset to diabetes prediction: Case study of deep neural network. Journal of Computer and Communications, 10, 2022.

[17] Md Reza, Ruhul Amin, Rubia Yasmin, Woomme Kulsum, and Sabba Ruhi. Improving diabetes disease patients classification using stacking ensemble method with pima and local healthcare data. Heliyon, 10:e24536, 01 2024.

[18] Deepti Sisodia and Dilip Sisodia. Prediction of diabetes using classification algorithms. 06 2023.

[19] P Yasodha and M Kannan. Analysis of a population of diabetic patients databases in weka tool. International Journal of Scien- tific Engineering Research, 2, 01 2011.