International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 01 | Jan 2021 www.irjet.net p-ISSN: 2395-0072

A Fusion Approach for Prediction of Diabetes using Machine Learning Techniques: A Survey Paper

Ansuyaba Vaghela[1], Dr. Gayatri S. Pandi[2]

[1]*Department of computer Engineering, L.J Institute of Engineering & Technology (Gujarat Technology University), Ahmedabad, Gujarat, India*
[2]*Professor, L.J Institute of Engineering & Technology (Gujarat Technology University), Ahmedabad, Gujarat, India*

Abstract - *Diabetes mellitus has become a leading concern in the modern society. The number of factors that cause diabetes can be unhealthy lifestyle, lack of exercise, deficient diet, age, genetics, obesity etc. In this research, the proposed methodology tries to improve the accuracy and specificity of predicting diabetes using different machine learning techniques. Experiments observed on Pima Indian Diabetes Dataset have validated the effectiveness and ascendancy of the proposed method.*

Key Words: **Diabetes, Classification, Machine Learning, Pima Indian Diabetes (PID)**

1. INTRODUCTION

Diabetes is a chronic disease that occurs when the **pancreas** is no longer able to make **insulin**, or when the body cannot make good use of the insulin it produces. Some of the important organs for regularization of blood glucose is Small Intestines (also known as digestive system and it responsible for broken down the food and absorbed into the blood streams as glucose), Pancreas (it helps in regulating blood sugar by producing insulin in the beta cell and producing glucagon in the alpha cells), Liver (it store glucose in glycogen and also produce glucose in the process called gluconeogenesis), Muscles (it absorbs the glucose). When the blood sugar is high, the pancreas produce insulin that tells liver and muscles to absorb the glucose and when the blood sugar is low, the pancreas produce glucagon that tells liver to made the new glucose.

Diabetes is classified as-

Type-1 Diabetes also known as Insulin-Dependent Diabetes Mellitus (IDDM) is the failure of human's body to produce sufficient amount of insulin and hence it is needed to inject insulin to a patient.

Type-2 Diabetes also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM) is seen when the body cells are not capable to use the insulin efficiently.

Type-3 Diabetes also known as Gestational Diabetes is increase level of blood sugar in pregnant woman where the diabetes is not detected earlier.
A person with diabetes is tends to have a intense complications like nerve damage, kidney failure, stroke and heart attack. Therefore Detecting diabetes in early stage is essential as it is preventing the person from getting excessive complications by taking sufficient care.

Machine learning is a remarkable tool in classification. Applying machine learning and data mining methods in diabetes research is a pivotal way to utilizing plentiful available diabetes-related data for extracting knowledge.[5]

The key purpose of this research is to develop a prophecy tool for early prediction of diabetes with improved accuracy. There have been many datasets are available on diabetes. In this paper the PIMA Indian Diabetes dataset is used from UCI repository for the classification of diabetes. The comparison of different machine learning algorithms is represented in an organized manner.

The remaining part of the paper is organized in following manner: "Motivation" is described in sec. 2, "Literature survey" is described in sec. 3, "Problem Statement" is described in sec. 4, "Proposed method" is described in sec. 5 and "Conclusion" is described in sec. 6.

2. Related Work

There are many methods which have been implemented for the classification of diabetes using Pima Indian diabetes dataset and other datasets. Some of this method's work is discussed below.

Akm Ashiquzzaman et al. (2017) used deep neural network approach for the classification of diabetes. They use 3 layers with ELU as a activation function. Each layer of DNN consists of a dropout function in learning process. The first two layers of proposed neural network has a low 25% probability in dropout, but the final layer has a 50% dropout rate to reduce over fitting.[1]

Qian Wang et al. (2019) integrate principal component analysis (PCA) and k-means techniques, and then apply logistic regression for the classification of Pima Indian diabetes dataset.[2]

Dilip Kumar Choubey et al. (2019) have used several classification methods, namely Logistic Regression, K-Nearest Neighbor (KNN), Iterative Dichotomizer3 Decision Tree (ID3 DT), C4.5 Decision Tree (C4.5 DT), on several datasets, namely Pima Indian Diabetes Dataset, Localized Diabetes Dataset for classification. They have used Principal Dimensionality Reduction (PCA), Particle Swarm Optimization (PSO) as feature reduction or feature selection or attribute selection method.[3]

Qian Wang et al. (2019) used naïve bayes method to compensate missing values through prediction then oversampling method is adopted to synthesize strongly similar samples through k-nearest neighbors. Finally, the RF method is adopted to obtain the classification results through the decision tree combination voting mechanism.[4]

Yukai Li et al.(2018) used a SMOTE algorithm which is used to create one more dataset and approximately make the ratio 1:1 for the problem of class-imbalance. They selected 4 algorithms to test decision tree, support vector machine (SVM), Bagging, and Adaboost.[5]

Swapna al. (2018) employed deep learning networks of Convolutional neural network (CNN) and CNN-LSTM (LSTM = Long Short Term Memory) combination to automatically detect the abnormality in heart rate signals from Electrocardiograms for the classification of diabetes.[6]

Huma Naz et al. (2020) used four data mining algorithms i.e. Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT) and Deep Learning (DL) for the classification of Pima Indian diabetes dataset. They used Shuffled sampling that split the dataset randomly and builds subsets from the applied dataset and then data is selected arbitrarily for assembling subsets. The model trained with stochastic gradient descent using back-propagation.[7]

Sushant Ramesh et al.(2018) use six classification techniques, ANN, SVM, DT, RF, LR and NB by applying the 10-fold cross validation method.[8]

3. Problem statement

To predict Diabetes requires a lot of information about bio-medical and medical field. Predicting if the Diabetes diagnosis is positive or negative based on several observations/features. There are 8 features are used, examples:- Number of times pregnant, Plasma glucose concentration a 2 h in an oral glucose tolerance test [Milligram per deciliter(mg/dL)], Diastolic blood pressure [Mille meter per mercury(mm Hg)], Triceps skin fold thickness [Mille meter (mm)], 2-h serum insulin [Micro unit per mille liter(mu U/ml)], Body mass index [Kilogram per square meter(weight in kg/(height in m)2)], Diabetes pedigree function, Age [Years]. Datasets are linearly separable using all 8 input features.
Target class: - Positive – Negative

4. Proposed System

Following describes the steps of the proposed system for classification:

Step 1: Select dataset
Input the Pima Indian diabetes dataset from the UCI machine repository.

Step 2: Data pre-processing
In this step the irrelevant data and noise are identified and removed unrelated data which is not important for the classification.

Step 3: Standardize data
In this we *Standardization* typically means rescales data to have a mean of 0 and a standard deviation of 1.

Step 4: Dimensionality reduction using PCA
In this step we apply principal component analysis for the feature reduction.

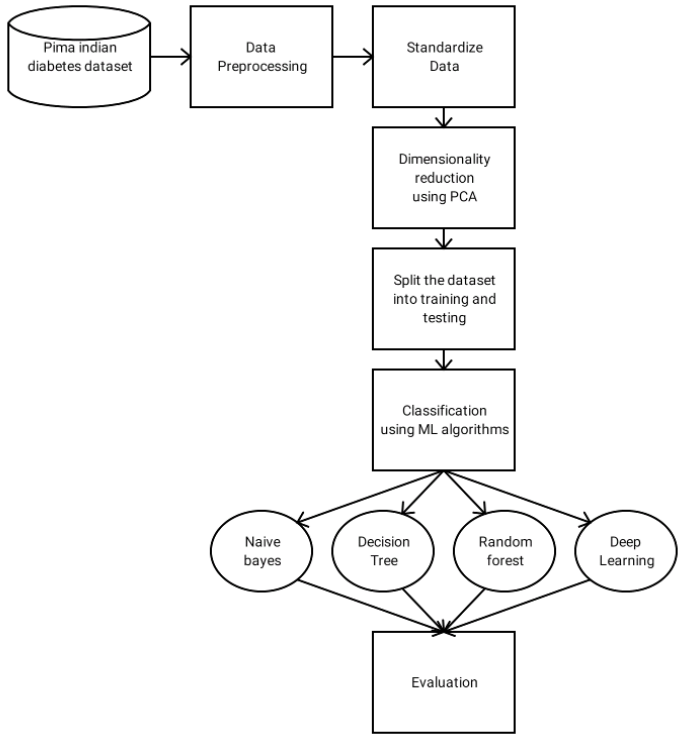

Fig -1: Proposed solution

Step 5: Split the dataset
In this step the dataset is spitted into training and testing sets and then we apply the algorithms.

Step 6: Classification
Classification of Pima Indian diabetes dataset using Naïve bayes, Decision tree, Random forest and Deep Learning methods are applied.

Step 7: Evaluation
Evaluation of algorithms based on accuracy, precision and sensitivity.

5. Conclusion

Survey of research papers give me an insight of techniques and algorithms used in the prediction of Diabetes. A performance comparison between different machine learning algorithms: Naive bayes, Decision Tree, Random forest, Artificial Neural network, Deep learning on the Diabetes datasets are conducted. Main parameters for the comparison were Accuracy, precision, Sensitivity, Specificity. By using mentioned techniques in the proposed system will

definitely help in better prediction of Diabetes and providing higher accuracy.

REFERENCES

[1] Akm Ashiquzzaman, Abdul Kawsar Tushar, Md. Rashedul Islam, Dongkoo Shon,Kichang Im, Jeong-Ho Park, Dong-Sun Lim, and Jongmyon Kim. "Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network" Springer: 31 August 2017 , p1.

[2] Changsheng Zhu, Christian Uwa Idemudia, Wenfang Feng. "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques" SCIENCE DIRECT: 4 April 2019 , Vol. 17, p1.

[3] Dilip Kumar Choubey, Prabhat Kumar, Sudhakar Tripathi, Santosh Kumar. "Performance evaluation of classification methods with PCA and PSO for diabetes" SPRINGER: 17 december 2019, p1-2.

[4] Qian Wang, Weijia Cao, Jiawei Guo, Jiadong Ren, Yongqiang Cheng, Darryl N Davis "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data with Missing Values" IEEE Access: 19 July 2019 ,vol 7,p2.

[5] Yukai Li, Huling Li, and Hua Yao "Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017" Hindawi: 10 July 2018 , vol 2018,p2.

[6] Huma Naz, Sachin Ahuja "Deep learning approach for diabetes prediction using PIMA Indian dataset" Springer:14 April 2020,p2.

[7] Swapna G, Soman KP, Vinayakumar R . "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals " Elsevier: 2018, p2.

[8] Sushant Ramesh, H. Balaji, N.Ch.S.N Iyengar and Ronnie D. Caytiles "Machine Learning Based Unified Framework for Diabetes Prediction" ACM: August 2018.

[9] S. rajasekaran, G.A. vijayalakshmi pai "Neural Networks, Fuzzy Logic and Genetic Algorithms: synthesis and applications" prentice-Hall of India Pvt. Ltd, new Delhi, 2004.