IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Fusion Approach for Prediction of Diabetes using Deep Learning Techniques

[1]Ansuyaba Vaghela, [2]Dr. Gayatri s pandi

[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 and diabetes type dataset have validated the effectiveness and ascendancy of the proposed method.

Index Terms: Diabetes, Classification, Deep learning, Machine Learning.

I. INTRODUCTION

Diabetes Mellitus is a chronic disease in which no insulin is produced in body from the pancreas or if the insulin is produced then the body is not able to use it.[1] 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 cells 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 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. [2]

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 and Diabetes type dataset is used to predict that whether the person have type 1 diabetes or type 2 diabetes. The comparison of different machine learning algorithms with deep learning 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.

II. MOTIVATION

Rates of both type 1 and type 2 diabetes are increasing globally. According to the International Diabetes Federation (IDF) Diabetes Atlas, here are the overall rates including both type 1 and type 2:

- 415 million adults have diabetes (1 in 11 adults)
- By 2040, 642 million adults (1 in 10 adults) are expected to have diabetes
- 46.5% of those with diabetes have not been diagnosed
- 1 in 7 births is affected by gestational diabetes and 542,000 children have type 1 diabetes
- 12% of global health expenditure is spent on diabetes ($673 billion)[3]

III. 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.

Md. Kamrul hasan et al. (2020) used Multilayer perceptron (MLP) and other machine learning classifier for the classification of diabetes. After removing missing values and outliers they apply MLP with 3 hidden layers that contain 16, 64 and 64 neurons respectively. The MLP architecture is trained with 200 epochs and relu as an activation function.[4]

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.[5]

Changsheng Zhu 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.[6]

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.[2]

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.[7]

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.[8]

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.[9]

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.[10]

IV. 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

V. PROPOSED SYSTEM

As shown above the proposed model is divided in to seven stages:

Stage 1: Gathering of Diabetes Dataset.

The Pima Indian Diabetes dataset obtained from UCI repository of machine learning. The dataset comprised 768 sample female patients from the Arizona, USA. The dataset has a total of 8 attributes with one target class. In the dataset there are a total of 268 tested positive instances and 500 tested negative instances.[11] The diabetes type dataset is derived from data.world which contains 8 columns with class of type 1 and type 2. [12]

Stage 2: Data Preprocessing and Standardize

In Machine learning models if you feed the garbage data, then you should expect garbage result with high probability. So data preprocessing step is necessary for cleansing of data which is required for the given model. Data preprocessing can be done by normalization, dealing with the missing data. The Blood Pressure, BMI, Glucose and insulin columns contain zero values which are replaced by the median value.

The outliers are removed using the Inter quartile Range (IQR) method. IQR measures dispersion by dividing a rank-ordered dataset into four equal parts, called quartiles.[13]

The values that divide each part are denoted by Q1, Q2, and Q3, where Q1 and Q3 are the middle value in the first and second half of the rank-ordered dataset respectively; and Q2 is the median value in the entire set. IQR is then equal to Q3 minus Q1. Outliers here are data instances that fall below Q1−1.5 IQR or above Q3+1.5 IQR.[14]

Then data standardize assures that all data of attributes are of same type distribution with zero mean and unit variance.

Stage 3: Applying SMOTE to the dataset

The dataset contains more numbers of negative rows than positive cases. So the dataset is balanced using SMOTE (Synthetic Minority Oversampling Technique**)** which oversample the minority class using synthetic instance that generated using convex combination of nearest minority class neighbors.

Stage 4: Split the dataset into training and testing sets

The dataset is divided into 80:20 ration where 80 percent of the data is trained using the algorithms and the remaining 20 percent of the data is used for testing and validation of the algorithm.

Stage 5: Training of Deep Learning and Machine learning model for Diabetes Prediction
In this stage the generated preprocessed and balanced data will feed to the input layer of deep learning sequential model. The model use 2 hidden layers with 64 and 32 neurons with tanh as an activation function. The output layers contain 2 neurons with softplus as an activation function. The model is trained using 1000 epochs and the loss is counted using the binary cross entropy. The other machine learning algorithms that are logistic regression, naïve bayes, random forest and svm are trained and the output will provide, where we get the result.

Stage 6: Testing of Model
After training the model it is required to test the model for whether the output given by it is accurate or not.

Stage 7: Result
The result will give the prediction that whether the person having diabetes or not. And if the person suffers from diabetes then predict that the person having type 1 diabetes or type 2 diabetes.

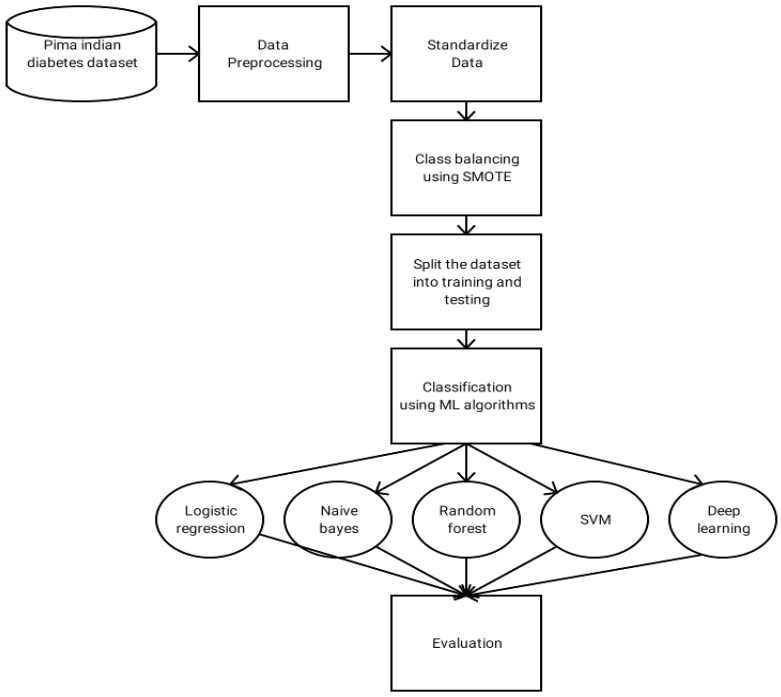

Fig 1: Proposed solution

VI. TOOLS AND TECHNIQUES

The dataset used for classification is "pima Indian diabetes dataset" which is retrieved from the UCI machine learning repository and is composed of data relented to 768 people including 8 attributes out of which 268 is tested positive and 500 tested negative. The other dataset id Diabetes type dataset which is derived from data.world and it contains 8 attributes. For achieving the required goal of the proposed method we will use python libraries. For training of Deep Learning we can use pretrained model or deep learning libraries like TensorFlow. TensorFlow is the core open source library for developing the machine learning model. TensorFlow Extended is an end-to-end platform for preparing data, validating, training and deploying models. It provides practical approach for Deep Learning with GPU support. For other machine learning methods python library sklearn is used.

VII. RESULTS

The experiments are done in a Google colab. Google colab provides virtual platform for the deep learning techniques that use tensor flow in the backend. After preprocessing of the dataset the smote algorithm is applied for the class balance. The dataset contains 500 entries each for negative and positive class.

Deep learning Results:

The deep learning method gives the best accuracy on pima Indian diabetes data that is 98.4.
The classification report that contains precision, recall, f1 score and support of pima Indian diabetes dataset are shown in the figure. After applying Smote the dataset contains 1000 entries for equally positive and negative case.

	precision	recall	f1-score	support
0	0.99	0.98	0.98	500
1	0.98	0.99	0.98	500
accuracy			0.98	1000
macro avg	0.98	0.98	0.98	1000
weighted avg	0.98	0.98	0.98	1000

Fig 2: result of deep learning on PID dataset

The confusion matrix is plotted using seaborn library. It contains four boxes that is true positive(TP), false positive(FP), True negative(TN) and False negative(FN).
TP :- the result is true and the predicted result is also true.
FP : the result is false but the predicted result is true.
TN : the result is false and the predicted result is also false.
FN : the result is true but the predicted result is false.

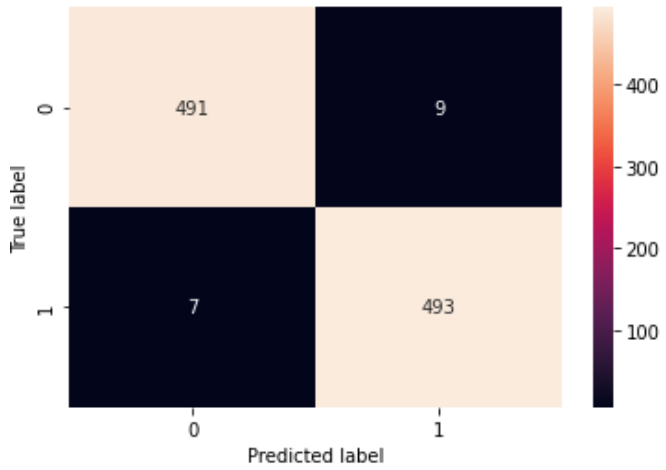

Fig 3: Confusion matrix of deep learning on PID dataset

The other dataset used is diabetes type dataset for the classification of type 1 and type 2 diabetes. Same neural network with same number of hidden layers and with smote balncing algorithm is applied on the dataset. The model can predict the class of diabetes with 100 percent accuracy.
The classification report that contains precision, recall, f1 score and support of diabetes type dataset are shown in the figure.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	205
1	1.00	1.00	1.00	205
accuracy			1.00	410
macro avg	1.00	1.00	1.00	410
weighted avg	1.00	1.00	1.00	410

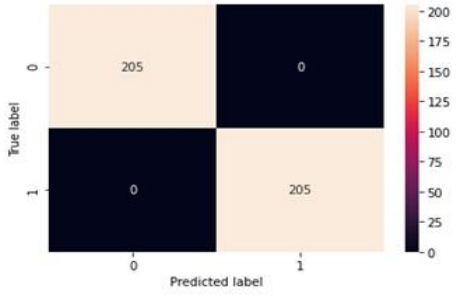

Fig 4: result of deep learning on Diabetes type dataset

After data processing and balancing and dividing it into training and testing set of 80:20 the data is also trained using different machine learning algorithms that are random forest, naïve bayes, logistic regression and svm. The machine learning model is implemented using sklearn library of python.

Classification report of random forest is:-

	precision	recall	f1-score	support
0	0.88	0.93	0.90	500
1	0.92	0.88	0.90	500
accuracy			0.90	1000
macro avg	0.90	0.90	0.90	1000
weighted avg	0.90	0.90	0.90	1000

confusion matrix of random forest:-

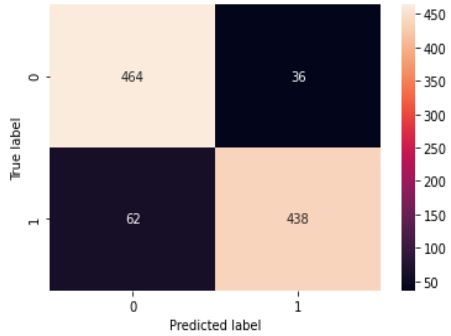

Fig 5: Random forest result

Classification report of naive bayes is:-

	precision	recall	f1-score	support
0	0.79	0.77	0.78	500
1	0.78	0.79	0.78	500
accuracy			0.78	1000
macro avg	0.78	0.78	0.78	1000
weighted avg	0.78	0.78	0.78	1000

confusion matrix of naive bayes:-

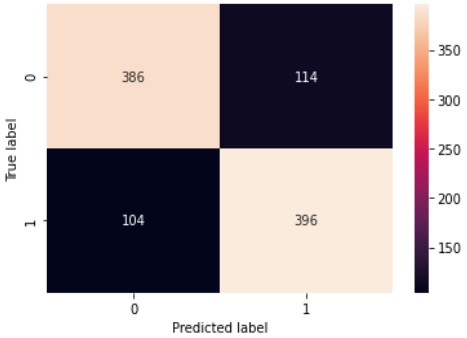

Fig 6: Naïve bayes result

Classification report of Logistic regression is:-

	precision	recall	f1-score	support
0	0.76	0.78	0.77	500
1	0.78	0.75	0.76	500
accuracy			0.77	1000
macro avg	0.77	0.77	0.77	1000
weighted avg	0.77	0.77	0.77	1000

confusion matrix of Logistic regression:-

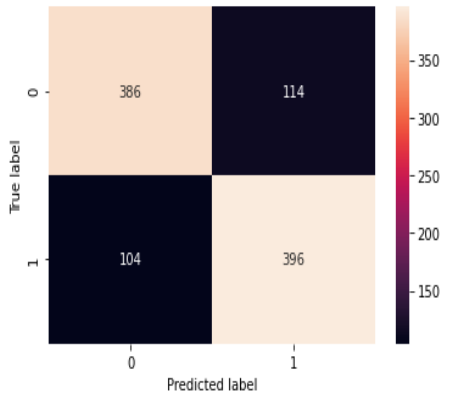

Fig 7: Logistic regression result

Classification report of SVM is:-

	precision	recall	f1-score	support
0	0.89	0.89	0.89	500
1	0.89	0.89	0.89	500
accuracy			0.89	1000
macro avg	0.89	0.89	0.89	1000
weighted avg	0.89	0.89	0.89	1000

confusion matrix of SVM:-

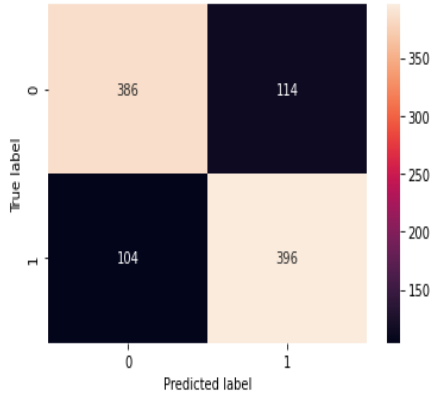

Fig 8: SVM result

Comparison of implemented algorithms

After applying different machine learning algorithms, we have conclude that the deep learning is more efficient in classification of diabetes.

Algorithm name	**Accuracy**
Deep learning	98
Random forest	90
Logistic regression	76.8
Naïve bayes	78.2
SVM	89.1

Table 1: Comparison of results of different algorithms

VIII. 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: Deep learning, Random forest, Logistic regression, Naive bayes, SVM on the Diabetes datasets are conducted. Main parameters for the comparison were Accuracy, precision, recall, support. **By using mentioned techniques in the proposed system will definitely help in better prediction of Diabetes and providing higher accuracy.**

REFERENCES

[1] Diabetes definition "https://www.idf.org/aboutdiabetes/what-is-diabetes.html/" accessed on 25 October 2020.
[2] Dilip Kumar Choubey, Prabhat Kumar, Sudhakar Tripathi, Santosh Kumar. "Performance evaluation of classification methods with PCA and PSO for diabetes" SPRINGER: 17 december 2019.
[3] Global diabetes rates "https://www.diabetesdaily.com/learn-about-diabetes/basics/what-is-diabetes/how-many-people-have-diabetes/", accessed on 25 October 2020.
[4] MD. Kamrul hasan , MD. Ashraful alam , Dola das , Eklas hossain. "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers" IEEE: 23 April 2020.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] Swapna G, Soman KP, Vinayakumar R . "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals" Elsevier:2018.
[10] Huma Naz, Sachin Ahuja "Deep learning approach for diabetes prediction using PIMA Indian dataset" Springer:14 April 2020.
[11] Pima Indian diabetes dataset "https://www.kaggle.com/uciml/pima-indians-diabetes-database" accessed on 01 October 2020.
[12] Diabetes type dataset "https://data.world/abelvikas/diabetes-type-dataset" accessed on 14 november 2020.
[13] Upton G, Cook I. Understanding statistics. Oxford University Press; 1996.
[14] Nonso Nnamoko, Ioannis Korkontzelos "Efficient treatment of outliers and class imbalance for diabetes prediction" Elsevier: 31 january 2020.