JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Optimization Accuracy of Diabetes Prediction using Different Machine Learning Technique

Nahid Noor Hussan[1], Prof. Suresh. S. Gawande[2]

M. Tech. Scholar, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal[1]

Guide, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal[2]

Abstract: Changing the info information into the arrangement of highlights is called include extraction if the highlights extricated are precisely picked it is normal that the highlights set will separate the pertinent data from the information so as to play out the coveted errand utilizing this diminished portrayal rather than the full size info. In this paper, gradient boosting machine learning technique to train the Diagnosis diabetes to classify the diabetes patients is two class values. The positive diabetes patients are defined by class '0' value and negative diabetes patients are defined by class '1'. The total Diagnosis diabetes dataset is 768. All dataset applied to the gradient boosting machine learning technique and get the 500 dataset is not diabetes and 268 dataset is diabetes. In proposed algorithm we used an ensemble of gradient boosting to achieve an accuracy of 81.95%. The Majority vote-based model as demonstrated which comprises of Naïve Bayes, Decision Tree and Support Vector Machine classifiers gave an accuracy of 76.56%, sensitivity of 79.16% and specificity of 77.476% for diabetes disease dataset.

***Index Terms* – Diabetic dataset, Classification, Machine Learning, Gradient Boosting**

I. INTRODUCTION

The mining of medical data is a valuable job in medical data mining. Most clinical DM algorithms are based on general knowledge or DMT. Medical expertise is used by most algorithms. Medical details are complicated and large. The sizes are minimized according to the requirements or the appropriate factors are only taken into consideration. When it is necessary to use medical data mining to reduce the process time and price and improve the efficiency of the result, it is important to choose appropriate features. Complex medical DMTs consisting of eleven feature selection methods and three fluorescent modelling techniques are challenging to find. Owing to the high cost focus, certain approaches are not available on the market. The best technology for mining using an effective combination of characteristics and fluorescent simulation is therefore crucial. The DMT is commonly used in the identification, retrieval and medical interpretation of patterns [1, 2].

Blood Sugar Level (BSL) for a certain time is known as a metabolic disorder. Data mining and deep learning allow us to remove hidden trends from big data in the medical sector. They can be used to examine essential therapeutic criteria, prevent outbreaks, approximate tasks in the pharmaceutical sector, facilitate care plans and to track patients. For the prediction and analysis of diabetes, many algorithms are predicted. The current approaches have more accurate knowledge than traditional systems. DM algorithms can contribute to sustaining decision-making in a variety of areas, including the medical sector [3, 4].

True health care data has been gathered, pre-processed, classified, predicted and clustered. In addition, DM time series is analyzed to provide DM models for diabetic patients with health-related results. Test results show that the built-in DM model will help healthcare providers develop affordable scientific choices for the classification of diabetic patients. The model can also be improved to take into account patient safety. In the future, the findings will be used to develop a diabetes treatment strategy since people with diabetes are not normally identified until a later stage of the disease is reached. 'METABO' is a diabetes surveillance and management device to monitor and decode the condition of a patient and to provide guidance on the collection for each person concerned as well as for the doctor [5].

Diabetes patients have to show blood glucose continuously and adjust the dosage of insulin, to conserve their blood glucose levels as regularly as possible. The phases of blood glucose which differ in general vary between direct and serious problems for the short and long term. An automatic prediction framework that warned people of making near changes in blood glucose phases would help them achieve anticipatory action using a well-known physiological blood glucose model to generate helpful elements of a vector support regression model that is informed on patient data. The latest dummy predicts diabetes experts at blood glucose level which can be used almost one-quarter of the times 30 minutes in advance to forecast hypoglycemia. Since the present resulting accuracy is only 42%, almost all hypoglycemic areas contain the most false warning. It is also sufferers who are not wounded by mediation in the response to these hypoglycemia markers [6, 7].

II. DIABETES DATA AND PRE-PROCESSING

Neural network procedures have been effectively pertinent to the conclusion of a few restorative issues. In this study we dissect the diverse neural system strategies for the determination of diabetes. The Pima Indian informational index is helpful to contemplate the

characterization exactness of the neural system calculations. The different information pre-preparing strategies are assessing to enhance the speculating exactness of the neural system calculations.

Neural network preparing can be made more proficient by executing certain pre-handling ventures on the system data sources and targets.

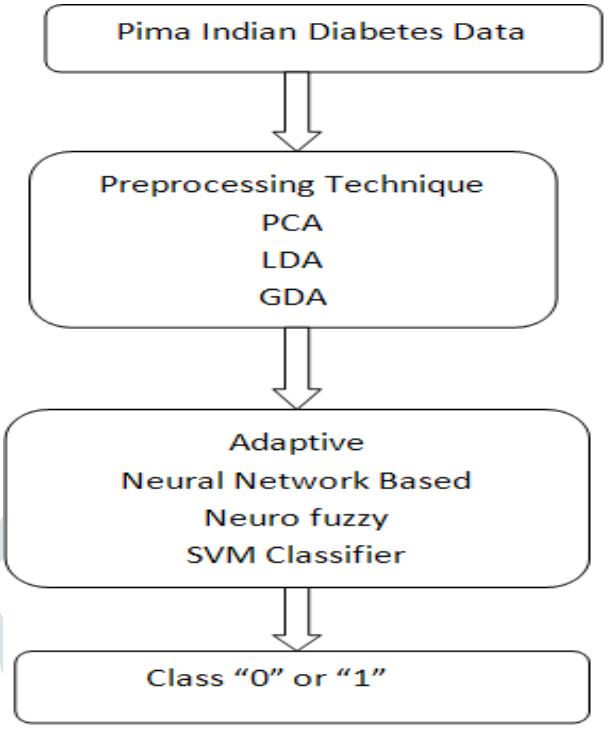

Fig. 1: Data pre-processing of Pima Indian data set

Diabetes Mellitus is a persistent elevated blood glucose (BG) condition. Almost half of all people with diabetes have a family inheritance which is one of diabetes mellitus most essential characteristics. Inadequate insulin in the pancreas and the ineffectual use of insulin by the body are both pathological causes of diabetes mellitus. Generally, two kinds of diabetes mellitus occur. Type 1 Diabetes Mellitus (T1DM) pathogenesis means that the pancreas secretes β-cells that have been damaged and prevents BG levels from declining. The disorders of Type 2 Diabetes Mellitus (T2DM), also named non-insulin-dependent diabetes mellitus, are insulin resistance and insulin secretion dysfunction. In the past 30 years, people have begun to realize that this chronic disorder has profoundly influenced every family and everyone's daily lives in China, with an increasing number of diabetes people. The percentage of diabetics in the general population increases and male diabetics are rising higher than female diabetics. According to government figures, in 2017, there were nearly 110 million diabetics in China.

Information pre-handling can likewise accelerate preparing time by beginning the preparation procedure for each element inside a similar scale. It is particularly valuable for displaying application where the sources of info are for the most part on generally extraordinary scales. This area comprises of two sub areas: diabetes informational collection and information pre-preparing. The neural system display for type II diabetes is created utilizing Pima Indian Dataset and the target of the information handling is to set up the informational index for neural system calculations as given in Fig. 1.

III. PROPOSED METHODOLOGY

Supervised machine learning classifiers can be categorized into multiple types. These types include naïve Bayes, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), generalized linear models, stochastic gradient descent, support vector machine (SVM), linear support vector classifier (Linear SVC) decision trees, neural network models, nearest neighbours and ensemble methods. The ensemble methods combine weak learners to create strong learners. The objective of these predictive models is to improve the overall accuracy rate. This can be achieved using two strategies. One of the strategies is the use of feature engineering, and the other strategy is the use of boosting algorithms. Boosting algorithms concentrate on those training observations which end up having misclassifications. There are five vastly used boosting methods, which include AdaBoost, CatBoost, LightGBM, XGBoost and gradient boosting.

Dataset was divided into two datasets (70%/30%, training/testing) to avoid any bias in training and testing. Of the data, 70% was used to train the ML model, and the remaining 30% was used for testing the performance of the proposed activity classification system. The expressions to calculate precision and recall are provided in Equations (2) and (3).

Precision provides a measure of how accurate your model is in predicting the actual positives out of the total positives predicted by your system. Recall provides the number of actual positives captured by our model by classifying these as true positive. F-measure can provide a balance between precision and recall, and it is preferred over accuracy where data is unbalanced.

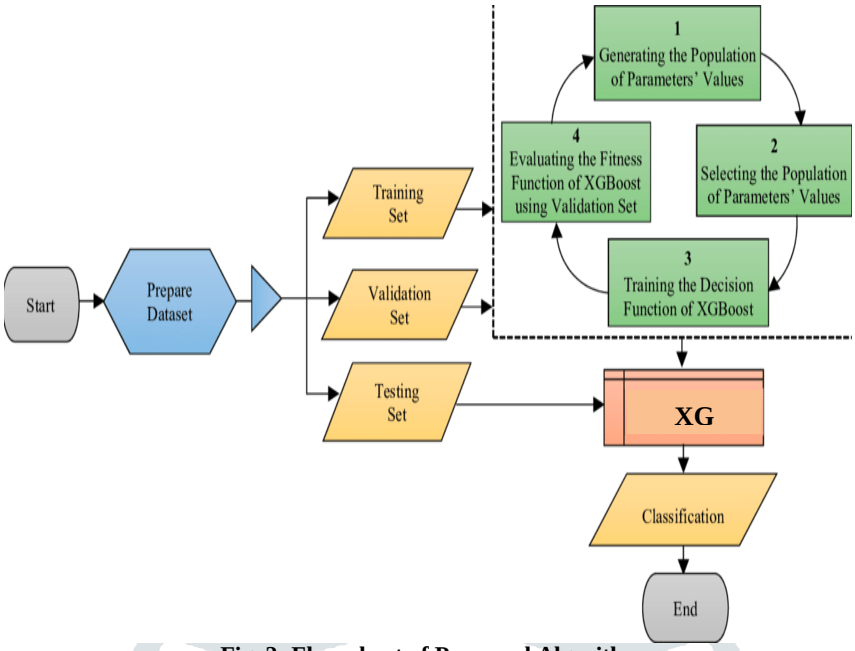

Fig. 2: Flow chart of Proposed Algorithm

Algorithm steps:

Input: D= $\{(\boldsymbol{x1},), (\boldsymbol{x2},), \ldots, (\boldsymbol{xN},\boldsymbol{yN})\}$, $\boldsymbol{L}(\boldsymbol{y},\boldsymbol{O}(\boldsymbol{x}))$

Where: $(\boldsymbol{y},(\boldsymbol{x}))$ is the approximate loss function.

Begin

Initialize: $(\boldsymbol{x}) = \frac{\boldsymbol{argmin}}{\boldsymbol{w}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{L}(\boldsymbol{yi},\boldsymbol{w})$

$\boldsymbol{for\ m}$=1:$\boldsymbol{M}$

$$\boldsymbol{r_{im}} = -\frac{\boldsymbol{\partial L(yi,O(xi))}}{\boldsymbol{\partial O(xi)}}$$

Train weak learner $\boldsymbol{C_m(x)}$ on training data

Calculate $\boldsymbol{w}$: $\boldsymbol{w_m} = \mathbf{arg\,min} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{L(yi, O_{m-1}(xi) + wC_m(xi))}$

Update : $\boldsymbol{O_m(x)} = \boldsymbol{O_{m-1}(x) + wC_m(x)}$

End for

End

Output: $\boldsymbol{O_m(x)}$

IV. EXPERIMENT RESULTS

Data Set

In this section describes the detailed analysis of experimental works carried out for our proposed model. The computational complexity of the proposed algorithm may change for different datasets depending on its size. The parametric values may vary accordingly.

Dataset description:

The source of Pima Indians dataset diabetes dataset on which the experiment is performed is UCI machine learning repository [12] with 768 data instances and 9 attributes. All patients in this dataset are Pima Indians women whose age is at least 21 years old and living near Phoenix , Arizona which denotes either "0" or "1", where '0' is tested as negative and '1' is tested as positive for diabetes.

Table 1: Description of benchmark dataset for diabetic for pima Indians

Datasets	No. of features	No. of classes	No. of patterns
Pima India Diabetic Dataset	9	2	686

Table 2: Description of parameter used for FFFNN

Datasets	Description	Considered value/Size
N	Number of input vector	768
D	Desired output vector	768
M	Number of hidden neurons	15
W	Weight vector	150
N	Number of input neuron	9
X	Input vector	768x9

Software

For implementation, we use MATLAB software with version 7.10.0. The coding for Classification using modified PSO-FFNN is executed in the command window. The operating system used is windows operating systems with 2 GB RAM. The results tabulated in the table Table 6 are carried out in MATLAB.

Table 3: Description of Diabetic Data Set

Data Set	No. of Attributes	Feature Set
Diabetic	9	No. of times pregnant Plasma glucose concentration Diastolic blood pressure Triceps skin fold thickness Serum insulin Body mass index Diabetes pedigree function Age of patient Class '0' or '1'

Parameter details:- The different significant parameter used for FFNN are center, spread and weight. The different symbols used for FFNN, PSO and IPSO are described in table 2 and table 3.

Evalution metrics: Generally, the evaluiton of a classification problem is based on a matrix calledd as a confusion matrix with the number of testing samples correctly classified and incorrectly classified represnted as follows

So, the accuracy can be measured according to Eq. 1

$$Accurancy = \frac{TN + TP}{TN + TP + FN + FP} \quad (1)$$

For a binary classification problem, the other measures include Precision, Sensitivity or Recall and Specificity. The formula to derive these measures is given in Eq. 6 and Eq. 7.

$$\Pr ecision = \frac{TP}{TP + FP} \quad (2)$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \quad (3)$$

In these relations ((1),(2) and (3) formula) TP means the number of samples that are healthy and properly diagnosed. FP indicates the number of samples that are healthy and have been diagnosed wrongly. FN indicates the number of samples that were sick but healthy wrongly diagnosed. TN contains a number of examples that have been patient and the patient is properly diagnosed.

Table 4: Comparison Result for Accuracy

Techniques	**Previous Algorithm**	**Implemented Algorithm**
SVM Technique	73.43%	77.60%
Decision Tree	72.91%	79.16%
Random Forest	74.4%	78.64%
KNN	71.3%	71.35%
Logistic Regression	72.39%	80.20%
Gradient Boosting	-	81.95%

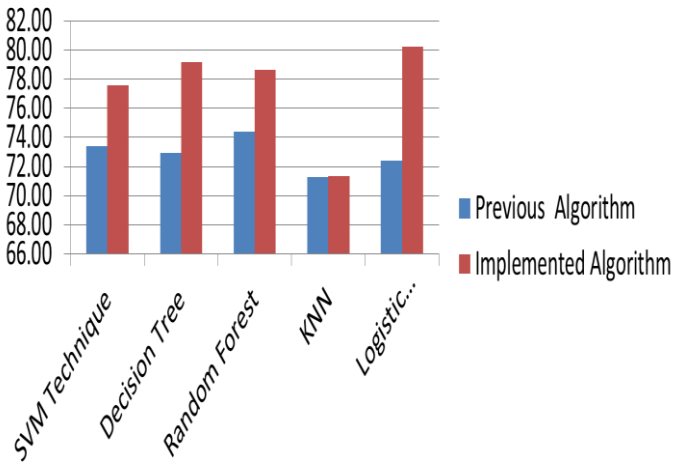

Fig. 3: Bar Graph of the Previous and Implemented Algorithm for Accuracy

V. CONCLUSION

According to the World Health Organization (WHO) 2018 Report, diabetes is one of the world's fastest-growing chronic disease threats to life that affected 422 million individuals. Due to the presence of a relatively long asymptomatic phase, early diabetes detection is always wanted to achieve clinically significant results. Approximately, 50% of all diabetic people are not diagnosed due to their long asymptomatic stage. Effective detection of diabetes can be rendered only by correct identification of typical and less typical indication signs observed at different phases from illness start to detection. Typically, diabetes disorder is caused by blood sugar (BS) rates that are greater than average. Instead, insulin output may be found inadequate. In recent days, it has been noticed that the percentage of patients with diabetes has expanded worldwide. To ensure that the total number of diabetes individuals reduces, this issue will undoubtedly be treated more seriously over the coming days. In proposed algorithm we used an ensemble of SVM, KNN and gradient boosting to achieve an accuracy of 81.95%. The Majority vote-based model as demonstrated which comprises of Naïve Bayes, Decision Tree and Support Vector Machine classifiers gave an accuracy of 76.56%, sensitivity of 79.16% and specificity of 77.476% for diabetes disease dataset.

REFERENCES

[1] Jyoti Agarwal, Anu Sharma, Namit Gupta, P.R. Lakshmi Eswari and Satyanadha Sarma Samavedam, "Diabetes Predication Analysis using Supervised Machine Learning Algorithm", 11th International Conference on System Modeling & Advancement in Research Trends (SMART), IEEE 2022.

[2] Anuj Mangal and Vinod Jain, "Performance analysis of machine learning models for prediction of diabetes", 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT), IEEE 2022.

[3] Nicole D'Souza, Kunjal Shah, Pranav Singh, "Diabetes Detection Using Machine Learning Algorithms", IEEE Bombay Section Signature Conference (IBSSC), IEEE 2022.

[4] Yogita Dubey, Pushkar Wankhede, Tanvi Borkar, Amey Borkar and Kajal Mitra, "Diabetes Prediction and Classification using Machine Learning Algorithms", IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), IEEE 2021.

[5] S.M. Tahsin Zaman, Subrata Kumer Paul, Rakhi Rani Paul and Md. Ekramul Hamid, "Detecting Diabetes in Human Body using Different Machine Learning Techniques", International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), IEEE 2021.

[6] Binhe Chen, Maosong Yan, Hongchuan Zhong and Bingwei He, "Prediction Model of Diabetes Based on Machine Learning", 5th Asian Conference on Artificial Intelligence Technology (ACAIT), IEEE 2021.

[7] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, "Prediction of Diabetes using Machine Learning Classification Algorithms", International Journal of Scientific & Technology Research, Volume 9, Issue 01, January 2020.

[8] Muhammad Azeem Sarwar, 2Nasir Kamal, 3Wajeeha Hamid,4Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2019.

[9] Rao G.A., Syamala K., Kishore P.V.V., Sastry A.S.C.S. ., "Deep convolutional neural networks for sign language recognition", 2018, International Journal of Engineering and Technology(UAE) ,Vol: 7, Issue 5, pp: 62 to 70.

[10] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", Springer, 2018.

[11] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neuro computing, vol. 237, pp. 350–361, May 2017.

[12] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

[13] Reddy S.S., Suman M., Prakash K.N. ., "Micro aneurysms detection using artificial neural networks", 2018, Lecture Notes in Electrical Engineering ,Vol: 434 ,Issue 3, pp: 409 to 417.

[14] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal 15, 104–116, 2017.

[15] Majid Ghonji Feshki and Omid Sojoodi Shijan, "Improving the Heart Disease Diagnosis by Evolutionary Algorithm of PSO and Feed Forward Neural Network", International paper on IEEE 2016.

[16] L. Hermawanti, "Combining of Backward Elimination and Naive Bayes Algorithm To Diagnose Breast Cancer", Momentum, vol. 11, no. 1, pp. 42-45, 2015.

[17] O.S. Soliman, E. Elhamd, "Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine", IEEE 2014.