Analysis and High Accuracy Prediction of Diabetes using Gradient Boosting Algorithm

Gaurav Goswami, Prof. Manish Saxena
M. Tech. Scholar, Head of Dept.
Dept. of Computer Science & Engineering
Bansal Institute Science and Technology, Bhopal

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.

Keywords—Diabetic dataset, Classification, Machine Learning, Gradient Boosting

I. INTRODUCTION

Diabetes is a extreme growing health problem in all over the world which causes death, especially in industrial and developing countries. This chronic disease may over to long term complications and death. It can cause high risk of kidney failure, nervous system damage, blindness and heart diseases. In this disease the body does not growth or rightly use the insulin, the hormone that release the cells of the body, grant the glucose to get in and ammunition them [1].

In the absence of insulin the cells become starving of glucose energy against the being of abundant glucose in blood stream. Complications of diabetes are linked to blood vessel diseases and generally classify into bitty vessel disease such as those affect the eyes, called as diabetic retinopathy. It occurs in the patients who have diabetes for minimum five years with the drainage of protein in small blood vessels at the back of eyes and the blood in retina. Disease in blood vessel also causes the formation of small aneurysms and new but brittle blood vessel leaded to retinal scarring and retinal detachments thus impairing vision [2].

Kidney damage from diabetes is called diabetic nephropathy. Initially disease blood vessel in the kidney causes the leakage protein in the urine. Later on kidneys lose their capacity to scrub and channel blood. The collection of dangerous waste item in blood prompts the requirement for dialysis.

Nerve harm from diabetes is called diabetic neuropathy and it is additionally caused by ailment of little veins. The blood stream to the nerves is restricted leaving the nerves without blood stream and they get harmed or kick the bucket subsequently. The symptoms of nerve damage are numbness burning, aching of feet and lower extremities [3]. There are two major types of diabetes type I and type II. Type I is basically diagnosed in children which is usually known as Juvenile diabetes and type II is most common form of diabetes. A patient with type II diabetes do not require insulin cure to remain alive, although up to 20% are treated with insulin to control blood glucose level [4]. Diabetes mellitus is a degenerative infection described by either absence of insulin or a protection from insulin, a hormone which is pivotal for digestion of glucose. In a sound individual, the pancreas produces insulin to help process sugar in the blood and keep up blood glucose (sugar) levels inside their ordinary range. Diabetics can't deliver insulin or are impervious to insulin, and subsequently can't expel glucose from the circulatory system. Regardless of whether there is insufficient insulin or insulin protection, glucose levels in the blood increment and cause extreme medical issues. There are two noteworthy sorts of diabetes. Sort 1, or insulin-subordinate adolescent diabetics, is hereditary in cause and is described by the body's powerlessness to create insulin, and the subsequent development of glucose in the blood. It normally happens amid pubescence or immaturity however can happen amid adulthood. Side effects incorporate outrageous craving and thirst, visit or over the top pee, and weight reduction. Definitive impacts of diabetes incorporate coronary illness, kidney malady, malignancy, hypertension, gangrene, diseases, visual deficiency, strokes, and demise.

Type II, or non-insulin-subordinate grown-up beginning diabetes, is the more typical frame and is portrayed by the body's protection from insulin. Around 90% of diabetes are type II and 80% of them are overweight when analyzed, for the most part amid middle age. Most stout diabetics have lifted insulin levels, however it doesn't control their glucose on account of the deficiency in insulin receptor cells and insulin protection. Stoutness and overabundance calories make a protection from insulin - that is, the pancreas keeps on creating insulin in light of blood glucose, yet the body's cells

oppose the activity of insulin. The mix of heftiness and high glucose prompt a lessening in the quantity of insulin receptors, destinations to which insulin connects to start transformation of glucose to glycogen or fat for capacity. Weight reduction and diminished caloric admission cause an expansion in the quantity of receptor cells prompting more proficient insulin digestion [5].

Indications of Type II diabetes are the same as side effects of Type I, however weight reduction is once in a while experienced in a Type II diabetic without an adjustment in count calories. Dissimilar to Type I diabetes, which for the most part requires normal insulin infusions, Type II diabetes can ordinarily be controlled by regular strategies, including diet, weight control, hormonal adjusting, compound treatment and home grown supplements. By and large, doctors consider a fasting plasma glucose level over 140 mg/gl as exorbitant.

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.

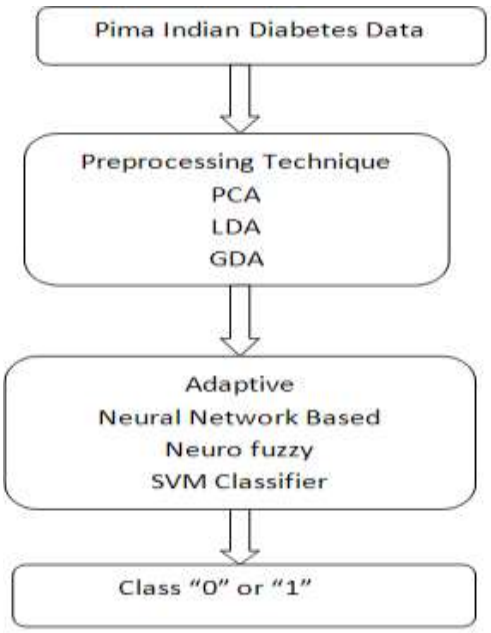

Fig. 1: Data pre-processing of Pima Indian data set

On the off chance that a preparation informational collection contains immaterial qualities characterization examination may deliver less exact outcomes. Information preprocessing is required to enhance the prescient exactness. The issue of missing information postures trouble in the examination and basic leadership forms and the missing information is supplanted before relegate it to NN demonstrate. Without this pre-handling, preparing the neural systems would have been moderate. It can be helpful to scale the information in a similar scope of qualities for each info highlight to limit predisposition inside the neural system for one element to another.

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 FRAMEWORK

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.

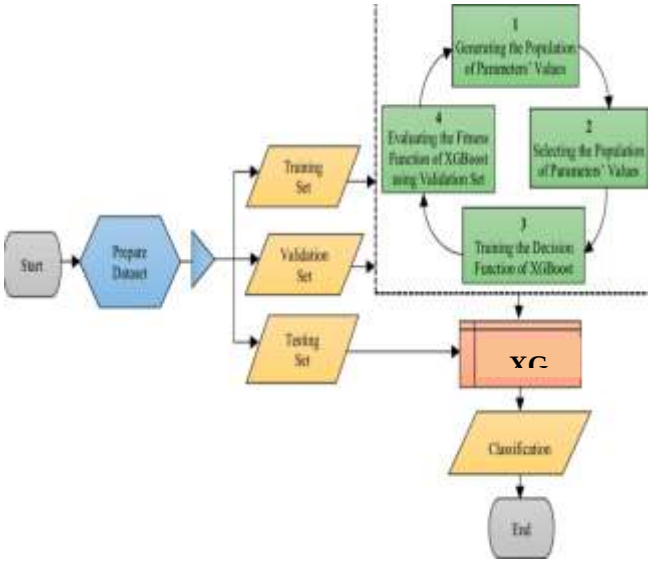

Fig. 2: Flow chart of Proposed Algorithm

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.

Algorithm steps:

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

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

Begin

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

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

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

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

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

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

End for

End

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

IV. EXPERIMENTAL 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.

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

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

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%

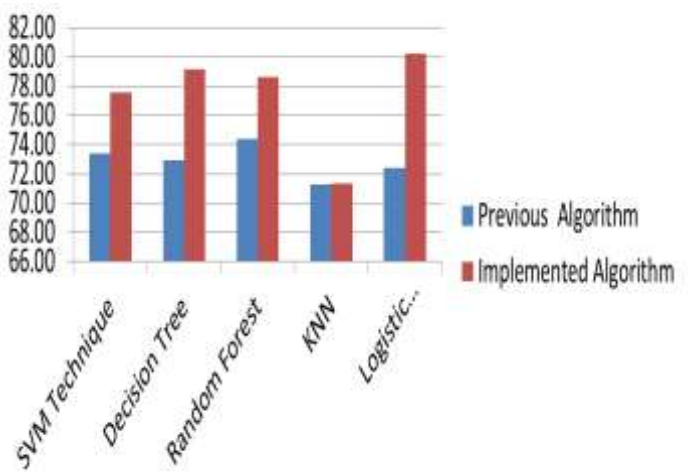

Fig. 3: Bar Graph of the Previous and Implemented Algorithm for Accuracy

V. CONCLUSION

Diabetes is metabolic disease that arises due to high blood glucose level in body. Insulin is not sufficient enough in body of diabetic patients to regulate the sugar level. Further several other diseases also arise from diabetes which is hazardous to health. It is necessary to detect such a serious health issue as early as possible Diabetes is cause of various diseases in human body. To make diabetes diagnosis easier for Physicians, there have been several methods employed. The diabetic data set is tested with selected classification algorithm. 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] 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.

[2] 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 2018.

[3] 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.

[4] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] K. Saxena, Z. Khan, S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbour Algorithm", International Journal of Computer Science Trends and Technology (IJCST), 2014.

[13] L. Hermawanti, S.G. Rabiha, "Combining of Backward Elimination and K-Nearest Neighbor Algorithms To

Diagnose Heart Disease", Prosiding SNST Ke-5 Fakultas Teknik Universitas Wahid Hasyim, pp. 1-5, 2014.

[14] R.A. Vinarti, W. Anggraeni, "Identification of Prediction Factor Diagnosis of Breast Cancer Rates with Stepwise Binary Logistic Regression Method", Jurnal Informatik, vol. 12, no. 2, pp. 70-76, November 2014.

[15] Muhammad Waqar Aslam, Zhechen Zhu and Asoke Kumar Nandi, “Feature generation programming with comparative partner selection for diabetes classification”, “Expert Systems with Applications”, 5402-5412, IEEE 2013.