Optimization Accuracy of Diabetes Prediction using Machine Learning Algorithm

Abhinav Sharma[1], Prof. Satyarth Tiwari[2], Prof. Suresh. S. Gawande[3]
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]
Co-guide, Department of Electronics and Communication, Bhabha Engineering Research Institute, Bhopal[3]

Abstract:-
In this research, the main objective will to classify the data as diabetic or non-diabetic and improve the classification accuracy. It presents an automatic prediction system for diabetes mellitus through machine learning techniques by taking into account of several limitations of traditional classifiers and provides a great relationship between patient's symptoms with diabetes diseases and the blood sugar rate. Machine learning provides a reliable and excellent support for prediction of a DM with correct case of training and testing. Diagnosis of diabetes mellitus desires great support of machine learning classifiers to detect diabetes disease in early stage, since it cannot be cured which brings great complication to our health system. This research work will consist of three phases. The first work contributed to develop a classification algorithm for prediction of DM. The second work will contribute as diabetes classification based on Extreme Learning Machine. The third work will contribute with optimization techniques for gradient boosting to obtain best output solution with higher accuracy. Optimization technique is used for searching and classifying the good diabetic data.

Keywords: - Diabetes Mellitus (DM), Machine learning, Early Stage, FPGA Application

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.
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.

II. MACHINE LEARNING

Machine learning (ML) is one of the widespread methods includes the several domains such as computer science and reaching applications. The computational learning theory belongs to statistics branch are used to analysis the performance and computation of machine learning algorithms. ML is used to designing algorithms which allows a computer to learn. Learning is the process of finding the statistical regularities or other patterns in the data. Therefore, it resembles how human might approach a learning task. In ML, data plays a crucial role, and the learning algorithm is used to identify and learn knowledge or properties from the data. In ML, data plays an indispensable role, and the learning algorithm is used to discover and learn knowledge or properties from the data. The quality or quantity of the dataset will affect the learning and prediction performance.

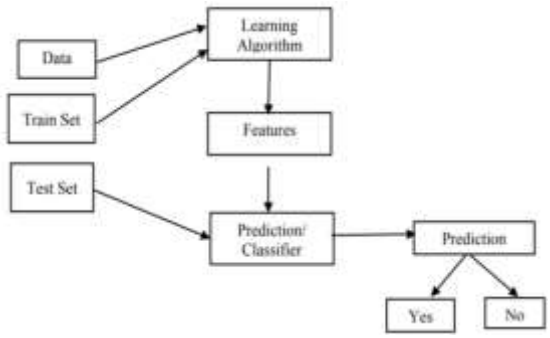

Figure 1: ML Approach

In machine learning, data plays an indispensable role, and the learning algorithm is used to discover and learn knowledge or properties from the data. The quality or quantity of the dataset will affect the learning and prediction performance.

In machine learning, an unknown universal dataset is assumed to exist, which contains all the possible data pairs as well as their probability distribution of appearance in the real world. The acquired dataset is called the training set (training data) and used to learn the properties and knowledge of the universal dataset. In general, vectors in the training set are assumed independently and identically sampled from the universal dataset. The figure 4.1 shows the process of machine learning in which the data sets are classified as train set and test set are forwarded to the learning algorithms then it predicts the presence or absence of certain features [2].

III. PROPOSED METHODOLOGY

Gradient boosting (GB) successively makes new models from a gathering of feeble models with the possibility that each new model can limit the misfortune work. This misfortune work is estimated by angle plunge technique. With the utilization of the misfortune work, each new model fits all the more precisely with the perceptions, and accordingly the general exactness is improved. In any case, boosting should be ultimately halted; something else, the model will tend to overfit. The halting measures can be an edge on the precision of forecasts or a most extreme number of models made.

Consider the arbitrary activation function f(x). The derivation of the activation function is denoted by F(x).

$$Y_{-ink} = \sum_i z_i w_{jk} \tag{1}$$

$$Z_{-inJ} = \sum_i V_{ij} X_i \tag{2}$$

$$Y_k = f(y_{-ink}) \tag{3}$$

The error to be minimized is

$$E = 0.5\sum\nolimits_k [t_k - y_k]^2 \tag{4}$$

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} 0.5\sum\nolimits_k [t_k - y_k]^2 \tag{5}$$

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial}{\partial w_{jk}} 0.5\sum\nolimits_k [t_k - f(y_{-ink})]^2 \tag{6}$$

$$\frac{\partial E}{\partial w_{jk}} = -[t_k - y_k]\frac{\partial}{\partial w_{jk}} f(y_{-ink}) \tag{7}$$

$$\frac{\partial E}{\partial w_{jk}} = -[t_k - y_k] f(y_{-ink}) \frac{\partial}{\partial w_{jk}} (y_{ink}) \tag{8}$$

$$\frac{\partial E}{\partial w_{jk}} = [t_k - y_k] f^1(y_{ink}) Z_j \tag{9}$$

Let us define

$$\delta_k = -[t_k - y_k] f^1(y_{-ink}) \tag{10}$$

Weight on connection to the hidden unit Z_j

$$\frac{\partial E}{\partial v_{ij}} = -\sum_k [t_k - y_k] f(y_{-ink}) \frac{\partial}{\partial v_{ij}} y_k \tag{11}$$

$$\frac{\partial E}{\partial v_{ij}} = \sum_k [t_k - y_k] f(y_{ink}) \frac{\partial}{\partial v_{ij}} y_{-ink} \tag{12}$$

$$\frac{\partial E}{\partial v_{ij}} = \sum_k \delta_k \frac{\partial}{\partial v_{ij}} y_{-ink} \tag{13}$$

Proposed calculation is a gradient learning method, utilized for characterization and relapse issues. It can create a viable model comprising of powerless students, for the most part choice trees. The essential thought of the proposed strategy is to assemble and sum up the group model in a stage wise design by enhancing a target subjective misfortune work. The proposed method develops its model from the past misfortune capacity of negative angle in an emphasis way. In the ML, limiting the misfortune work is a significant issue and should be enhanced. At the end of the day, the misfortune work addresses the contrast between the anticipated yield and the objective. A low worth of

misfortune work implies a high expectation or grouping result. At the point when the misfortune work diminishes successively and iteratively, the model heads continuously along a particular course, which is the Gradient of misfortune work.

IV. SIMULATION TOOLS

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 (14) and (15).

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.

Therefore, F-measure was utilized in this study as a performance metric to provide a balanced and fair measure using the formula in (16).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

Where,

TP—True Positive, FP—False Positive, FN—False Negative

V. SIMULATION RESULTS

In data set used for this study, there are categorical variables such as Cp, chest pain type which is represented as 1, 2, 3 and 4. 1, 2, 3 and 4 does not have ordinal relationship with each other therefore it gives wrong results when applied directly to machine learning algorithms. Thus, OneHotEncder is used to encode chest pain type values into binary values, this resolves the issue of ordinality. In this data set the dependent variable or the value to be predicted is multi class. It ranges from 0 to 4.

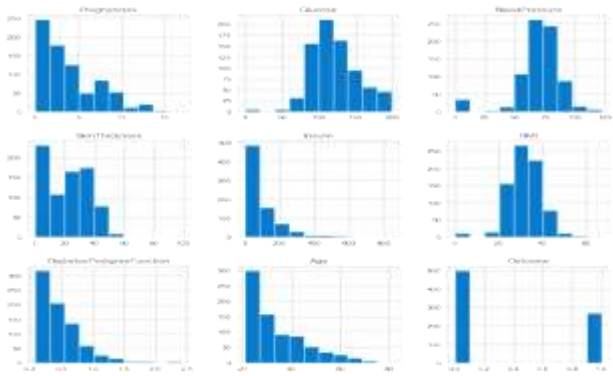

Figure 2: Attributes of Dataset

Figure 3 shows the histogram of attributes shows the range of dataset attributes and code which is used to create it.

In figure 4 are the status of diabetes health ranging from healthy to severely unhealthy. Blue bar represents male population and red bar represents female population. It can be seen that; in this data set the male population is more prone to diabetes disease.

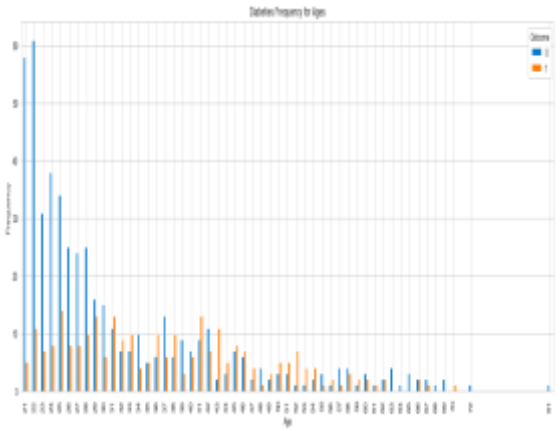

Figure 3: Bar Plot of Number of Diabetes Frequency for Ages

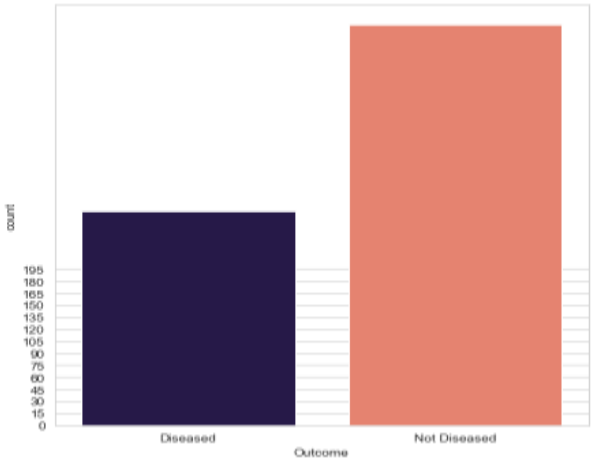

Figure 4: Outcome of Disease and Not Disease

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.

VI. CONCLUSION

In proposed algorithm we used an ensemble of gradient boosting algorithm 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. It is clear that the gradient boosting algorithm is providing best accuracy for diabetes diagnosis than previous algorithm. In future, improvement in the execution time for large size data set could be treated as a research subject.

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.