International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

ENHANCED PREDICTION OF DIABETES USING MACHINE LEARNING TECHNIQUES

Prachi Patel, Prof. Manoj Yadav
M. Tech. Scholar, Dept. of Computer Science & Engineering, School of Engineering, SSSUTMS, Sehore
Assistant Professor, Dept. of Computer Science & Engineering, School of Engineering, SSSUTMS, Sehore

***Abstract*:-** Due to its intricate dependencies on numerous factors, diabetes diagnosis is a very difficult task in the early stages. In order to assist medical experts in the demonstration method, it is necessary to establish restorative symptomatic emotionally supportive networks. Neural system functions have been successfully linked to the diagnosis of many medical conditions. Gradient boosting machine learning is used in this thesis to train the diabetes diagnosis and classify diabetic patients into two groups based on their class values. To attain an accuracy of 81.95% in the suggested strategy, we employed an ensemble of gradient boosting techniques.For diabetic disease dataset, the majority vote-based model, which includes Naïve Bayes, Decision Tree, and Support Vector Machine classifiers, achieved an accuracy of 76.56%, sensitivity of 79.16%, and specificity of 77.476%.
.

Keywords*: *Diabetes, Machine, Learning, Prediction, Dataset, Ensemble

I. INTRODUCTION

Diabetes is a grave health issue that is spreading rapidly around the world and is particularly deadly in industrialized and developing nations. This persistent illness may lead to long-term problems and even death. It can increase the risk of heart disease, neurological system damage, blindness, and kidney failure. In this illness, the body is unable to produce new cells or use insulin properly—a hormone that allows glucose to enter the body and fuels it [1].

Despite the bloodstream's abundance of glucose, cells that lack insulin starve themselves of its energy. Diabetes complications are associated with blood vessel illnesses and are often categorized as vascular vessel diseases, such as diabetic retinopathy, which affects the eyes. Patients who have had diabetes for at least five years experience protein discharge in the small blood vessels in their retinas and back of their eyes. A blood vessel disease can also result in the development of tiny aneurysms and new, brittle blood vessels that can induce retinal detachments and scarring, which can impair vision [2].

Diabetic nephropathy is the term for kidney damage caused by diabetes. Urine protein leakage is first caused by diseased renal blood vessels. Kidneys eventually lose their ability to filter and distribute blood. Dialysis is necessary because blood contains hazardous waste materials.Diabetic neuropathy is the term for the damage diabetes causes to nerves; small vein disease is another contributing factor. Restrictions on the blood supply to the nerves cause them to lose blood flow, which can lead to injury or death. Neural injury manifests as burning, painful feet and lower extremities [3]. Diabetes comes in two main forms: type I and type II. Type I diabetes, also referred to as juvenile diabetes, is primarily diagnosed in youngsters, while type II diabetes is the most prevalent kind of the disease.

Although up to 20% of patients with type II diabetes receive insulin treatment to regulate blood glucose levels, insulin is not necessary for the patient's survival [4]. Diabetes mellitus is a degenerative infection characterized by either an inability to produce insulin or resistance to the hormone, which is essential for the breakdown of glucose. The pancreas in a healthy person produces insulin to aid in the breakdown of blood sugar and maintain blood glucose (sugar) levels within normal limits. Diabetics are unable to release glucose from the bloodstream because they are either insulin-insensitive or unable to administer insulin. Blood glucose levels rise and result in serious health problems regardless of the presence of insulin or insulin protection.

II. LITERATURE REVIEW

This research [2] introduced the Random Forest method for the prediction of diabetes in order to create a machine learning system that can more accurately predict diabetes in patients at an early stage. The suggested model provides the best results for diabetic prediction, and the outcome demonstrated that the prediction system can accurately, quickly, and most significantly forecast the onset of diabetes.

Diabetes Prediction Using Machine Learning Techniques, a work [4] presented, uses three different supervised machine learning methods—SVM, Logistic Regression, and ANN—to predict diabetes. An efficient method for early diabetic illness diagnosis is proposed by this effort.

This research [5] described a method for predicting the beginning of diabetes using ensemble supervised learning. Five popular classifiers were utilized for the ensembles, and the results were aggregated using a meta-classifier. The outcomes are displayed and contrasted with those of related research in the literature that made use of the same dataset. It is demonstrated that diabetes onset prediction can be performed more accurately by employing the suggested strategy.

This research [6] presents an Intelligent diabetic Disease Prediction System assembled by data mining that uses a database of diabetic patients to analyze diabetes cases. They suggest using algorithms like Bayesian and KNN (K-Nearest Neighbor) in this system to apply to a database of diabetes

patients and analyze them by taking into account several diabetes-related factors for the prediction of diabetes disease.

Six distinct machine learning algorithms were used in this work [7] to predict diabetes using machine learning in the healthcare industry. There is a discussion and comparison of the applicable algorithms' accuracy and performance. An analysis of the various machine learning approaches included in this research identifies the most appropriate algorithm for diabetes prediction. Researchers are increasingly interested in diabetes prediction in order to train programs to determine whether or not patients have diabetes by using the appropriate classifier on the dataset. Previous research has shown that there hasn't been much improvement in the classifying process. Therefore, a system is needed to address the problems found based on earlier research, as diabetes prediction is a crucial topic in computers.

According to this research [9], diabetes is one of the deadliest and most chronic diseases that raises blood sugar levels. If diabetes is left undiagnosed and untreated, several entanglements arise. A patient is seen by a symptomatic focus and counseling professional as a result of the tedious distinguishing measure. But the advancement of AI techniques addresses this fundamental problem. The goal of this study is to develop a model that can most accurately represent the likelihood that a patient would develop diabetes. In order to detect diabetes at an early stage, this research uses three AI characterisation computations, namely Decision Tree, SVM, and Naive Bayes.

Several AI techniques are applied in this study [10] to conduct predictive analysis on massive amounts of data across several domains. Although it can be a challenging task, vision screening in medical services can assist professionals in making highly informed decisions on the treatment and well-being of their patients. This study discusses the future research in healthcare; six different AI computations are used in this analysis. Six unique AI computations are applied to a dataset consisting of patient clinical records in order to analyze the data.

This research [11] views the extraction of complicated head and hand developments coupled with their continuously changing shapes as a problematic problem in PC vision for gesture-based communication acknowledgment. Three unique example estimates are used for CNN preparation; each estimate has a different subject and survey point layout. The prepared CNN is tested using the remaining two instances. To improve recognition accuracy, unique CNN models were developed and tested using communication data derived from our selfie gestures. Comparing our 92.88% recognition rate to other classifier models published on the same dataset, we achieved a higher accuracy.

In this paper [12], the remarkable progress in biotechnology and health sciences has led to the generation of a critical mass of data, including clinical data and high throughput genetic data generated from massive electronic health records (EHRs). With the main class having all the makings of the most well-known, the purpose of the current examination is to conduct a methodical survey of the uses of artificial intelligence (AI), information mining strategies, and instruments in the field of diabetes research regarding a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management. A wide range of AI computations were used. Typically, 15% of the ones used were by lone learners, while affiliation rules were used to illustrate 85% of directed learning draws.

III. PROPOSED METHODOLOGY

The suggested methodology is an ensemble learning method for regression and classification issues. It can generate a useful model using weak learners, typically decision trees. The fundamental idea behind the suggested approach is to optimize an objective arbitrary loss function in order to gradually construct and generalize the ensemble model. The suggested method builds its model iteratively using the prior negative gradient loss function. Minimizing the loss function is a crucial problem in machine learning that requires optimization. Stated differently, the difference between the target and the expected output is represented by the loss function.

To prevent bias in training and testing, the dataset was split into two datasets (70%/30%, training/testing). Thirty percent of the data were utilized to evaluate the suggested activity classification system's performance, with the remaining seventy percent being used to train the machine learning model. Equations (1) and (2) offer the formulas to compute precision and recall.

Precision gives you an idea of how well your model predicts the real positives from all of the positives your system is supposed to anticipate. By categorizing these as genuine positives, recall indicates how many real positives were actually caught by our model. In cases where data is uneven, the F-measure is recommended over accuracy because it can offer a compromise between precision and recall.
In order to offer a fair and balanced measure, the F-measure was used in this study as a performance statistic using the formula

Precision = TP TP + FP × 100 (1)

The formula for recall is TP TP + FN × 100 (2)

Measure F is equal to 2 ∗ Precision ∗ Recall Precision + Recall × 100 (3)
where FP stands for false positive, FN for false negative, and TP for true positive

The research is based on the Pima Indians diabetes dataset, which has 768 information occurrences and 9 attributes. The dataset is obtained from the UCI machine learning store [13]. All of the patients in this dataset are Pima Indian women, identified by the codes "0" and "1," where "0" indicates a negative diabetes test and "1" indicates a positive test. The patients are all at least 21 years old and reside in the Phoenix, Arizona, area.
.

Table 1: Attributes of Diabetes Data Set

Attribute No.	Attribute	Description	Missing Value
1.	Pregnant	A record of the number of times the woman pregnant	110
2.	Plasma glucose	Plasma glucose concentration measured using two hours oral glucose tolerance test (mm Hg)	5
3.	Diastolic BP	Diastolic blood pressure	35
4.	Triceps SFT	Triceps skin fold thickness (mm)	227
5.	Serum-Insulin	Two hours serum insulin (muU/ml)	374
6.	BMI	Body mass index (weight Kg/height in (mm) ²)	11
7.	DPI	Diabetes pedigree function	0
8.	Age	Age of patient(year)	0
9.	Class	Diabetes on set within five year	0

Gradient Boosting

Using an ensemble of weak models, gradient boosting (GB) iteratively builds new models with the goal of minimizing the loss function in each new model. The gradient descent method is used to measure this loss function. The accuracy is increased overall because each new model matches the observations more closely when the loss function is used. Nevertheless, boosting must finally come to a halt, otherwise the model will begin to overfit. A cap on the number of models produced or a threshold on prediction accuracy might be used as the terminating criteria.

Steps in the algorithm:

Data entered: D= {($\boldsymbol{x}$1,), ($\boldsymbol{x}$2,),…, ($\boldsymbol{xN}$,$\boldsymbol{yN}$)}, $\boldsymbol{L}(\boldsymbol{y},\boldsymbol{O}(\boldsymbol{x}))$

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

Start

($\boldsymbol{x}$) =argmin/w ∑_(i=1)^ initializen ⟦L(yi,w) ⟦

$\boldsymbol{form}$=1:$\boldsymbol{M}$

(∂L(yi,O(xi)))/(∂O(xi)) = r_im

Utilizing training data, train the weak learner C_m (x) by computing w:w_m=arg⟦0⟧min⟦0⟧∑_(i=1)^yi,O_(m-1) (xi)+wC_m (xi)) = N⟦

Revision:⟦O_m (x)= O⟧_(m-1) (x)+wC_m (x)

End for

End

Output: O_m (x)

IV. EXPERIMENTAL RESULTS

The Scikit-learn machine learning library contains the LabelEncoder and OneHotEncoder classes. Essentially, LabelEncoder converts the categorical data into ordinal numbers. The data set utilized in this study includes categorical variables such Cp and the type of chest pain, which are denoted by the numbers 1, 2, 3, and 4. Since 1, 2, 3, and 4 have no ordinal relationship with one another, using them straight to machine learning methods yields incorrect results. Consequently, the ordinality problem is resolved by encoding chest pain type values into binary values using OneHotEncoder.

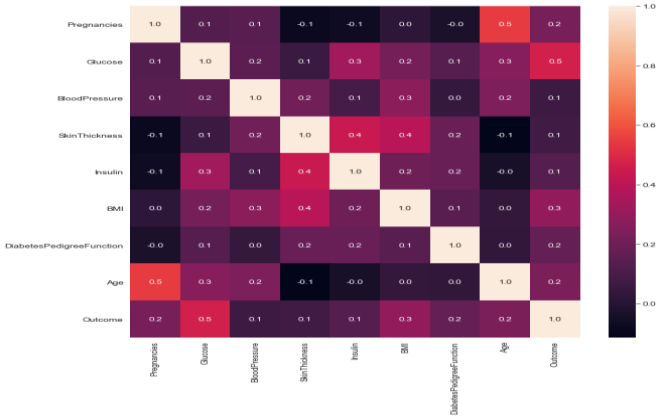

Figure 1: attribute representation through heatmap

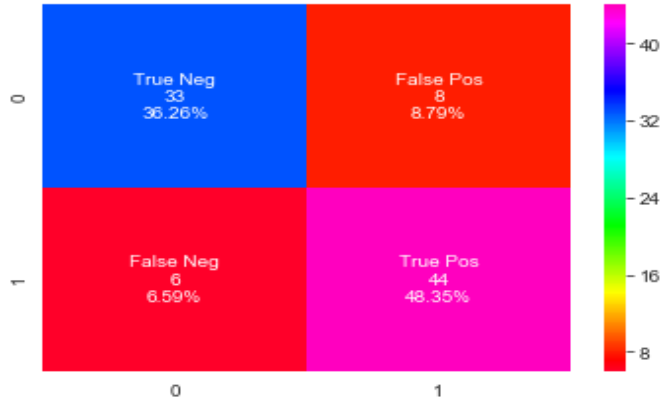

Figure 2: Confusion Matrix For Proposed Algorithm

We employed an ensemble of SVM, KNN, and ANN in the suggested technique to attain an accuracy of 81.95%. For the diabetic illness dataset, the majority vote-based model, which uses Naïve Bayes, Decision Tree, and Support Vector Machine classifiers, produced an accuracy of 76.56%, a sensitivity of 79.16%, and a specificity of 77.476%.

We discover that the accuracy of the KNN is significantly more efficient than that of other algorithms after using the machine learning approach for testing and training. The confusion matrix of each algorithm, as illustrated in Figure 5.12, should be used to calculate accuracy. Here, the number of counts for TP, TN, FP, and FN are given. Value is calculated using the accuracy equation, and it is concluded

that the proposed algorithm has the highest accuracy of all of them, at 81.95%.

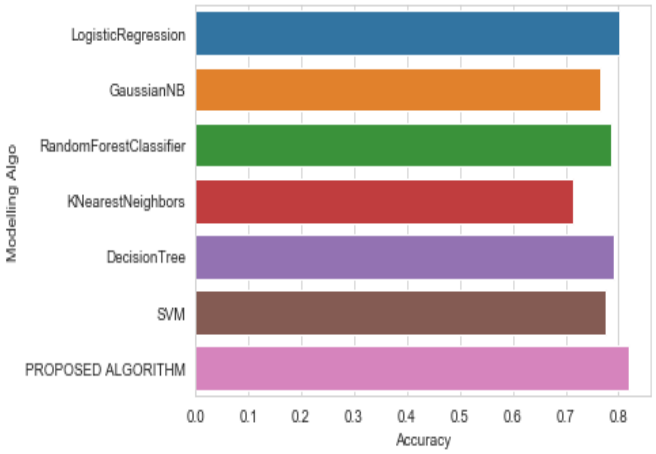

Figure 3: Overall accuracy comparison

V. CONCLUSION

This study work's primary goal was to develop and apply diabetes prediction using machine learning techniques and performance analysis of those techniques, and it was successful in doing so. We employed an ensemble of SVM, KNN, and ANN in the suggested technique to attain an accuracy of 81.95%. For the diabetic illness dataset, the majority vote-based model, which uses Naïve Bayes, Decision Tree, and Support Vector Machine classifiers, produced an accuracy of 76.56%, a sensitivity of 79.16%, and a specificity of 77.476%. It is evident that when compared to earlier algorithms, the gradient boosting approach offers the best accuracy for diagnosing diabetes. Enhancing the execution time for large data sets may be studied as a research topic in the future.

REFERENCES

1. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
2. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Compu- tation Automation and Networking, 2019.
3. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Perfor- mance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 Feb- ruary, 2019.
4. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineer- ing Research and Application, Vol. 8, Issue 1, (Part -II) Janu- ary 2018, pp.-09-13
5. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
6. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabe- tes Disease Prediction Using Data Mining ".International Con- ference on Innovations in Information, Embedded and Com- munication Systems (ICIIECS), 2017.
7. Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.
8. A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.
9. 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.
10. 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.
11. 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.
12. Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang' "Predicting Diabetes Mellitus With Machine Learning Techniques", Springer, 2018.
13. A. Asuncion, D. Newman. UCI machine learning repository. 2007.