International Journal on Emerging Technologies **11**(3): 867-872(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Identification of Diabetes with Recursive Partitioning Algorithm using Machine Learning

L. Lakshmi[1], M. Purushotham Reddy[2], A. Praveen[3] and K.V.N. Suniha[4]
[1]Professor, Department of Computer Science & Engineering,
BVRIT HYDERABAD College of Engineering for Women, Hyderabad,500090, INDIA.
[2]Associate Professor, Department of Information Technology,
Institute of Aeronautical Engineering, Hyderabad,500043, INDIA
[3]Assistant Professor, Department of Information Technology,
Institute of Aeronautical Engineering, Hyderabad,500043, INDIA
[4]Professor & Principal, Department of Computer Science & Engineering,
BVRIT HYDERABAD College of Engineering for Women, Hyderabad,500090, INDIA.

(Corresponding author: L. Lakshmi)

(Received 26 March 2020, Revised 20 May 2020, Accepted 22 May 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Diabetes is also called Diabetes Mellitus, which is a word for several conditions involving how your body turns food into energy. It is due to metabolic disorders and the reduction of blood glucose levels in the body. There are stages in diabetes based on the severity. So, it is an emergency task to identify diabetes at an early stage to decrease the severity of the problem. By considering all the complications many research studies are through to solve the problem in an effective way. As most of the existing methods used the Pima Indians Diabetes data set, we also used the same data set with different machine algorithm called recursive partitioning algorithm to improve the accuracy in predicting diabetes at the early stage of human life. As a first step we have performed pre-processing of data to improve the quality of training data. Next we have performed feature subset selection to select only important features from given data set. Then we have divided our data into training and test data. Next we applied an efficient nondependent recursive partitioning algorithm to perform prediction in the neural network model. Our investigational results revealed that the recursive partitioning approach provides a substantial performance improvement over prevailing approaches.

Keywords: Diabetes, Machine Learning, Prediction, Feature subset selection, Recursive partition algorithm,

Abbreviations: *SVM*, support vector machines; PCA, principal component analysis; ML, Machine Learning; TP, true positives; TN, true negatives; FP, false positives; FN false negatives

I. INTRODUCTION

Diabetes mellitus is a typical infection that influences a huge majority share of the individuals in numerous pieces of the world. Diabetes influences individuals generally after the age of 20. Diabetes commonness has been expanding more in low and middle revenue nations [1]. It turns into a reason for different sicknesses likewise like visual deficiency, kidney failure, heart maladies, and cholesterol. Forecast of diabetes at a beginning time would assist the patients with maintaining the sugar level under control.

Diabetes is a constant illness or gathering of metabolic ailment where an individual experiences an all-encompassing degree of blood glucose in the body, which is either the insulin creation is lacking, or because the body's cells don't react appropriately to insulin. The steady hyperglycemia of diabetes is identified with long haul damage, brokenness, and failure of different organs, especially the kidneys, eyes, nerves, veins, and heart.

The main objective of this study is to make use of noteworthy features, propose a good machine learning prediction algorithm, and locate the ideal classifier to give the nearest result contrasting with clinical results [2] [14]. The proposed technique expects to concentrate on choosing the properties that mainly create problems in the early location of Diabetes Miletus utilizing predictive analysis.

The existing algorithm's results show that the Naive Bayes, PCA, and decision tree algorithm has the most noteworthy explicitness of 94.50, 95.20% and 95.00% separately holds best for the examination of diabetic information [3]. Naïve Bayesian algorithm outcome states the finest accurateness of 82.30%. The exploration also takes a broad view of the assortment of ideal features from the dataset to increase the classification precision.

Some of the research methods use deep learning architectures to classify normal and diabetic HRV signals. They used convolutional neural network (CNN), and their combinations to extract complex temporal dynamic characteristics from HRV input data. These characteristics are transmitted to the classification vector machine (SVM) for classification. They improved the accuracy by 0.03% using CNN architecture compared to previous work without using SVM. The existing classification system can help clinicians diagnose diabetes using ECG signals with a very good accuracy of 92.7% [13].

For the experimental analysis, Pima Indians Diabetes [4] data set is selected and divided into train and test dataset. Feature selection is an important step in data preprocessing. This will choose the subset of features from the entire list of capabilities depending on the factual score and will expel repetitive features that don't contribute to experimental performance. After the collection of features, the algorithm for classification is applied to build a good classification training model. Then the model is tested with test data set for foreseeing the diabetes risk. The performance metrics accuracy, sensitivity, and specificity are measured and evaluated.

II. LITERATURE REVIEW

As diabetes is a chronic condition in the body it is essential to create and then a model is to be predicted so that we can easily identify the condition of the body related to diabetes, keeping this in mind Pima Indians set is being proposed to conduct further implementations. Diabetes is mainly of two types, namely type1 and type 2. This is precisely based on the complexity of the problem. It depends on the glucose levels and metabolism of the person. Type 1 diabetes can be defined as the scenario in which the cells of the pancreas that produce insulin are damaged. It is an autoimmune condition; this may be due to genes sometimes. Type 2 diabetes is due to insulin is produced by the pancreas but the problem is, it is not used by the body appropriately. This type of diabetes can be seen in teens and adults. According to a survey, almost 90% of people are suffering from this type.

According to a survey by Neha Sharma (2018), early recognition and screening assume a fundamental job in inadequate anticipation of diabetes. The method of learning begins with review or information, like examples, models, precise events, or orders, to look for patterns in the information and settle on acceptable decisions [5].

Duyguç, and Esin D (2011) proposed “An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier.” In some journals, tongue images are used to detect diabetes. For this, MATLAB is used for diabetic identification [6]. In this scenario, both the diabetic and non-diabetic patients are diagnosed and then outputs are predicted. Patients with diabetes type 2 are influenced significantly with higher covering range of yellow fur, thick fur, and also bluish tongue ($P < .001$) compared to a group of people under control. Likewise, an expressively higher percentage of patients having long term diabetics and those with yellow fur than the short term were observed.

Sudajai *et al*. (2011) proposed “Knowledge based DSS for an Analysis Diabetes Of Elder using Decision Tree” and discussed diabetes examination in seniors. The outcome indicated that the Random forest model has the most elevated precision in the classification is 94.50 percent when contrasted and the clinical finding that the error with RMSE, 0.0447, and with MAE, 0.004 [7]. The NBTree model has the most minimal accurateness in classification is 70.60 percent when contrasted and the clinical conclusion that the error with MAE, 0.3327, and with RMSE, 0.454.

Alghamdi *et al.* (2017) proposed “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project”, the detection of information from clinical databases is significant to make a compelling clinical conclusion [8]. They used the Pima Indian diabetes data set for training and testing the model. Preprocessing was utilized to improve the nature of the information. The classifier was applied to the changed dataset to improve the Naïve Bayes model accuracy. At last, the WEKA tool was utilized to do a reenactment, and the precision of the subsequent model was 72.3%.

Jack W. Smith *et al.* (1988) proposed “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus”, and applied the classification techniques of data mining to classify diabetes clinical information and anticipate whether the patient was influenced with diabetes or not [9]. They introduced a framework that gave training information on that information include pertinence examination is done then the correlation of classification algorithm, choosing classifier at that point improved classification calculation is applied and afterward discovered the assessment that contrasted and training information. They used the C4.5 algorithm, which gave a classification accuracy of 91%.

K.C. Tan *et al.* (2009) proposed “A hybrid evolutionary algorithm for attribute selection in data mining”, and discussed a short filtering technique that expels bothersome features before classification starts while the wrapper strategy applies classification technique to choose ideal features [10]. Wrapper strategy gives higher classification exactness. The main drawback of the wrapper methodology would be a more extended runtime because the ML algorithm needs to run repetitively in the quest for quality subsets.

Swapna *et al.* (2018) proposed "Diabetes detection using deep learning algorithms" mentioned that diabetes can also be identified by giving heart rate variability as input. Through this, they estimated 97% accuracy. This method gave more accuracy as it is based on blood vessels and its area and perimeter resulted in motivational outputs [11]. Some of the researchers proved that even the tongue samples can be used to examine the presence of diabetes by using some mechanisms. They are various methods for this kind of detection too. According to certain studies, it is said that even eyes can lead to diabetes. This can be explained as if the retina's blood vessels are damaged it causes diabetes which is usually called Diabetic Retinopathy. This may lead to blindness.

Here, if we just go back and could recollect all the solutions identified for the identification of diabetes we probably could draw the major corners that involve in detection. For example, most of the articles include the Pima Indian dataset, Heart rate variability, ECG, CNN, Deep Learning algorithms [12], data classification. Logistic regression is used to predict the probability of the target variable. It is a supervised learning classification algorithm. The targeted nature of the variable dependency is considered as dichotomous, it states that there might be simply two probable classes. Here, logistic regression can be two more categories based on the target variables which can be expected by it. Based on those categories, this Logistic regression can be majorly divided into the following types – binary, nominal, ordinal [13,14].

Deep Learning Techniques are broadly used to solve problems more efficiently. Now-a-days vast number of deep learning and machine learning algorithms are used in diabetes analysis and detection to get the best predicted results of the available solutions. Increasing neural network architecture size will result in improve the accuracy of results. In diabetes detection, we can also use machine learning algorithms like Support Vector Machine (SVM), Naïve Bayes, Random Forest algorithm, Recursive Partitioning Algorithm, and many others. In this vast technology filled with various algorithms, it is proposed to implement the Recursive Partitioning algorithm to improve accuracy in diabetes prediction.

III. THE ARCHITECTURE OF THE PROPOSED SYSTEM

The architecture of the proposed system is shown in Figure 1 uses Pima's Indian diabetes database with nine features in which eight features are used as input and one feature is used for output and it consists of both non-diabetic and diabetic records. As a first step data preprocessing is not only involved the data cleaning but also handling of missing values, dimensionality reduction, inconsistent data handling, feature selection. Feature selection is performed using the gain feature selection technique to select highly correlated features to achieve classification goals. After feature selection, the data set is divided into 70 percentage training data and 30 percent of testing data.

In this paper, we have proposed a classification model by utilizing a recursive partitioning algorithm to foresee the diabetes hazard in the example in the training data set. Supervised learning is the most widely used approach in the prediction and classification of data. Supervised learning is used to the precise classification of data with known class labels and maps the given input to output. The recursive partitioning algorithm will construct a classification or regression model and the outcome is acquired in for binary tree representations. The data is to be modeled properly so that there is no redundancy in the data and have no missing values to get accurate results for the esteemed problem.

The classification is model is built using a supervised recursive partitioning algorithm as it classifies data samples into different groups. The Random Forest classifier is a classification algorithm that contains various Decision trees. This is a statistical method. It is used for multivariable analysis. It creates decision trees by recursively partitioning based on several independent variables. Its features include bagging and also randomness while constructing individual trees to create uncorrelated forest possessing trees. By doing this it is the fact that it gives more accuracy than any of the individual trees.

After building the model, the model is tested with 30 percentage test data for accuracy, specificity, sensitivity, and also the precision of the Neural Networks. After testing the model for accuracy, then it predicts diabetes accuracy.

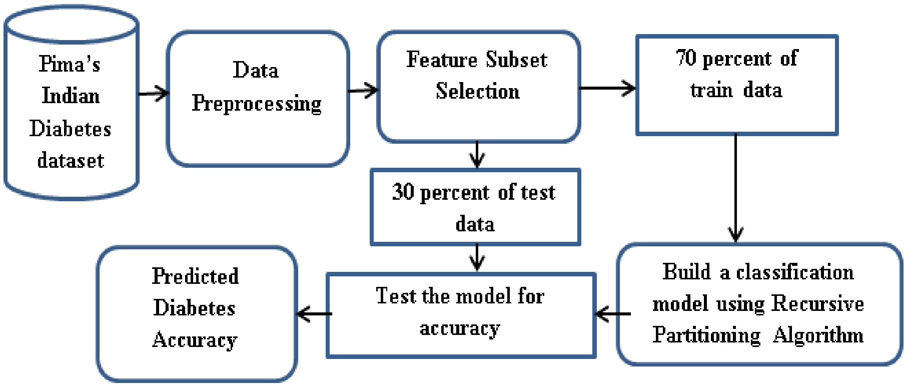
Pima's Indian Diabetes dataset
Data Preprocessing
Feature Subset Selection
70 percent of train data
30 percent of test data
Predicted Diabetes Accuracy
Test the model for accuracy
Build a classification model using Recursive Partitioning Algorithm

Fig. 1. The architecture of the proposed system.

The performance of the model is evaluated using the three performance metrics namely accuracy, specificity, and sensitivity. These measurements are determined from the given confusion matrix. The confusion matrix mainly used for performance prediction of a specified classification model on the given training data set. This matrix summarizes the trained results of the recursive partitioning classifier and it mainly consists of True negatives, true positives, false negatives, and false positives.

This paper measures the performance of the algorithm using three performance metrics namely, accuracy, sensitivity, and specificity. These metrics are calculated from the given confusion matrix. The confusion matrix is the table that is used to predict the performance of a classification model on a sample set of data. It is used for summarizing the results of a classifier. It is a matrix that shows all the number of True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN). The representation of the confusion matrix is shown in Table 1.

Table 1: Confusion matrix representation.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

The formulae used to compute the performance measures are presented in the given equations. For better identification of a binary classification test or which may exclude the conditions correctly can calculate by accuracy. To calculate the true positive rate or recall we used sensitivity. It can be used to measurement of the portion of the true positives that can be accurately recognized as the true positives. The third performance metric is the true-negative rate also known as precision or specificity which measures a portion of true-negatives correctly identified as true-negatives.

$$Total = TP + TN + FP + FN \tag{1}$$

$$True_{Total} = TP + TN \tag{2}$$

$$Accuracy = \frac{True_{Total}}{Total} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

The 'Recursive Partition Algorithm' is a greedy and top-down algorithm that produces decision trees. It uses the repetitive process for the selection of best split at a given node. An example binary decision tree model produced by the algorithm is shown in Figure 2.

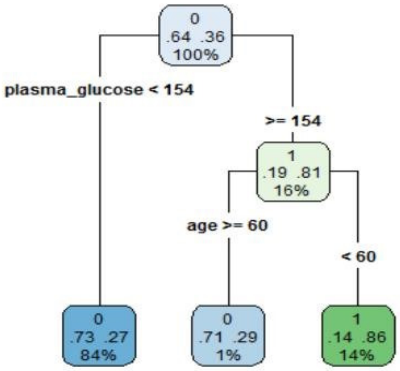

Fig. 2. Representation of binary decision tree model.

IV. PERFORMANCE ANALYSIS

We have evaluated the performance of the random partitioning algorithm on the Pima Indian data set; the preliminary analysis discloses the subsequent insights of data. The given data set comprises of female patient data and their ages ranging from 21-81. The feature set used in the dataset includes the features A1-A8. The features and their descriptions are represented in the Table 2.

Table 2: Feature set used in Dataset.

Feature_ID	Feature Name	Feature Description
A1	Pregnant_Times	How many number of times pregnancy occurred
A2	Plasma_Glucose	Concentration of Plasma Glucose Level
A3	Diastolic_BP	The blood pressure in arteries
A4	Skin_Thickness	Skin fold thickness used to measure body fat
A5	Serum_Insulin	Insulin levels in the blood
A6	Body_Mass_Index	BMI used to measure body fatness
A7	Pedigree_Function	Used to know ancestors history function
A8	Age_Num	Age should be represented in numbers
A9	Prediction variable	True or False

The following figure represents plasma glucose levels that can be measured as a factor in diabetes risk.

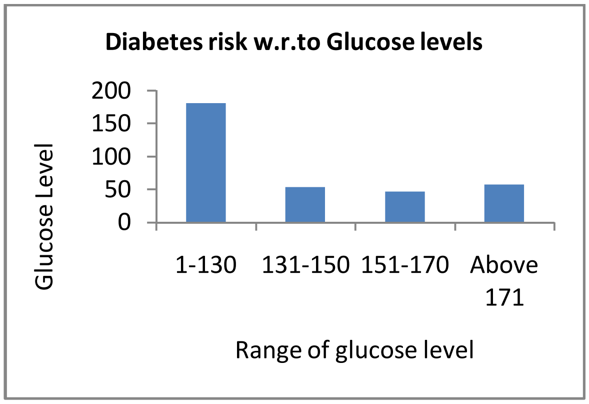

Fig. 3. Diabetic risk with respect to glucose levels.

As shown in the Figure 3 it is evident that glucose level plays a little impact on the analysis of diabetes risk. Similarly, we can evaluate the performance of the diabetic risk against the remaining attributes in the given data set.

Now the data set is classified into 70 percentages of training data and 30 percentages of test data. Here we are representing the results feature subset selection to improve the accuracy of the results. The main purpose of a random selection algorithm is dividing the feature set into sub-feature sets to improve classification accuracy.

Table 3: Accuracy, Sensitivity, and Specificity evaluation with respect to attribute set.

Attribute List	Accuracy	Sensitivity	Specificity
A1, A2, A3, A4	81.05	91.25	75.56
A1, A2, A3, A4, A6	81.06	91.50	75.63
A1, A2, A3, A4, A5, A6	81.56	91.34	74.53
A2, A4, A5, A6, A8	81.24	90.24	73.24
A1, A2, A3, A7	78.56	86.53	69.58
A1, A2, A3, A4, A5, A6, A7, A8	78.58	85.68	68.53

The Table 3 represents the selection of a combination of features for the attainment of good accuracy. From the Table 3, it is observed that when we include some features like attribute A7 the accuracy of the model is decreasing meaning those features less impact on the model. To correctly evaluate the system we are evaluating the model one to eight features and we are estimating the accuracy of the model. The classification accuracy of the model concerning different features is represented in the following Table 4.

Table 4: Classification accuracies with respect to different attribute sets.

Attribute Set	Accuracy	Sensitivity	Specificity
Full attribute set	78.58	85.68	68.53
A8	69.54	70.31	70.12
A7,A8	81.26	91.56	75.56
A6,A7,A8	78.45	86.53	65.82
A5,A6,A7,A8	76.85	82.31	59.56
A4,A5,A6,A7,A8	75.63	82.46	55.45
A3,A4,A5,A6,A7,A8	75.23	94.23	46.52
A2,A3,A4,A5,A6,A7,A8	71.25	85.64	42.52

The results represented in the Table 4 have shown the highest accuracy when the attributes named Diabetes Pedigree Function and the other one named age are both removed from the attribute set. The algorithm provides good accuracy when compared to Naïve Bayes, linear regression algorithms. The training model is improved for all most all attribute sets when training and testing data set is increases. The variation of testing accuracy, sensitivity, and specificity are represented in the following Figure 4.

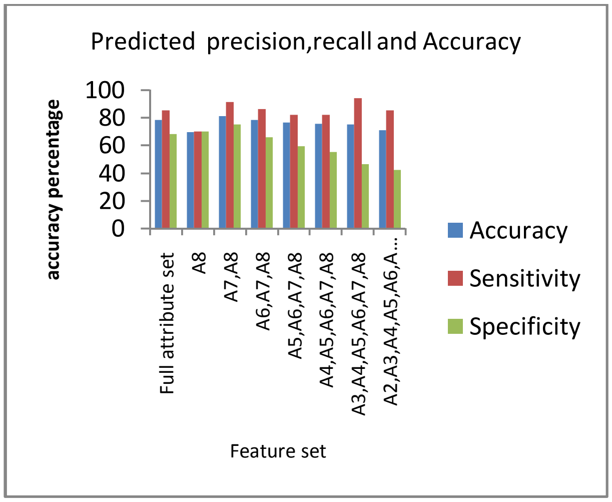

Fig. 4. Accuracy percentage vs. feature set.

The experimental results represented in above Figure. 4 reveals that if we remove attribute Diabetes Pedigree Function, then the model achieved the highest accuracy. The above graph also represents the relationship between precision, recall, and F1 score and also the good balance between the measures.

IV. CONCLUSION

Early-stage diabetes detection plays an important role in human life. This paper provides detailed insight into various diabetes detection algorithms in machine learning and also presents experimental evaluation with one of the machine algorithm called recursive partitioning algorithm to improve the accuracy in predicting diabetes at the early stage of human life. The performance of the model is calculated by making use of performance measures like accuracy, specificity, and sensitivity. Our investigational results revealed that recursive partitioning approach provides a substantial performance improvement over prevailing approaches.

V. FUTURE SCOPE

The performance of model using random partition algorithm is good. To improve the accuracy of model we can use large data with more training data so that model will be well trained with all features of diabetes.

Conflict of Interest: The research work submitted by me is original and executed. Even through many researchers are working on the same problem, we have worked Indian diabetes data set and results produced by us are original.

REFERENCES

[1] World Health Organization, Geneva, (1985), "Report of a study group: Diabetes Mellitus," World Health Organization Technical Report Series 727.

[2] Kemal Polat, Salih Gunes, and Ahmet Arslanc, (2008), "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," *Expert Systems with Applications,* vol. *34*. 1, 482-487.

[3] Kayaer K and Yildirim T, (2003) "Medical diagnosis on Pima Indian diabetes using general regression neural networks," Proceedings of the international conference on artificial neural networks and neural information processing, 181-184.

[4] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Neha Sharma (2018). "Diabetes Detection and Prediction Using Machine Learning/IoT", IEEE.

[6] Duyguç.,Esin D, "An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier." Expert Syst. Appl. *38*, 2011, 8311–8315.

[7] Sudajai Lowanichchaia, Saisunee Jabjone b, TidanutPuthasimmac, (2011) " Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree", in International Journal on Soft Computing, vol. 2. 2, 15-23.

[8] Alghamdi M., Al-Mallah M., Keteyian S., Brawner C., Ehrman J., Sakr S, (2017). "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project." PLoS One 12:e0179805. 10.1371/journal.pone.0179805.

[9] Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, (1988), "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Proc. Annu. Symp. Comput. Appl. Med. Care, 9, 261-265.

[10] K.C. Tan, E.J. Teoh, Q. Yua, K.C. Goh, (2009).A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, Volume 36, Issue 4, 8616-8630.

[11] Swapna G., Kp S., Vinayakumar R,(2018), "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals", Procedia Comput. Sci., 132,1253-1262.

[12] Goutham, Swapna& R, Vinayakumar & Kp, Soman. (2018). Diabetes detection using deep learning algorithms.ICT Express.4. 10.1016/j.icte.2018.10.005.

[13] Ratna Patil, Sharavari and Tamane, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes", *International Journal of Electrical and Computer Engineering* (IJECE), 2018, *8*(5), 3966-3975.

[14] L. Alekya, L. Lakshmi, G. Susmitha, and S. Hemanth, (2020), "A Survey On Fake News Detection In Social Media Using Deep Neural Networks", in *International Journal Of Scientific & Technology Research*, *9*(3), 5063-5066.

How to cite this article**:** Lakshmi, L., Purushotham Reddy, M., Praveen, A. and Sunitha, K.V.N. (2020). Identification of Diabetes with Recursive Partitioning Algorithm using Machine Learning. *International Journal of Emerging Technologies*, 11(3): 867–872.