International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com
doi : https://ijsrcseit.com/CSEIT2390442

An Effective Machine Learning Approach for Diabetes Prediction

Appana Naga Lakshmi[1], Durga Thokala[2], Soma Ganesh Kumar[3], Bollapally Rakesh Reddy[4], Dr. G Vinoda Reddy[5]

[1]Assistant Professor, Department of Artificial Intelligence, Madanapalle institute of Technology & Science, Andhra Pradesh, India

[2]Assistant Professor, Department of Artificial Intelligence & Machine Learning, Pragati Engineering College, Andhra Pradesh, India

[3]Department of CSE(AIML), Sri Indu College of Engineering and Technology, Hyderabad, India

[4]Department of CSE(AI&ML), Sri Indu college of Engineering and Technology, Hyderabad, India

[5]Professor, CSE (AI & ML) Department, CMR Technical Campus, Hyderabad, India

ARTICLEINFO

Article History:

Accepted: 01 Aug 2023
Published: 10 Aug 2023

Publication Issue
Volume 9, Issue 4
July-August-2023

Page Number
325-335

ABSTRACT

Diabetes is a chronic condition that could lead to a global health care disaster. 382 million people worldwide have diabetes, according to the International Diabetes Federation. This will double to 592 million by 2035. Diabetes is a condition brought on by elevated blood glucose levels. The symptoms of this elevated blood sugar level include frequent urination, increased thirst, and increased hunger. One of the main causes of stroke, kidney failure, heart failure, amputations, blindness, and kidney failure is diabetes. Our bodies convert food into sugars, such as glucose, when we eat. Our pancreas is then expected to release insulin. Insulin acts as a key to unlock our cells, allowing glucose to enter and be used by us as fuel. However, this mechanism does not function in diabetes. The most prevalent forms of the disease are type 1 and type 2, but there are other varieties as well, including gestational diabetes, which develops during pregnancy. Data science has an emerging topic called machine learning that studies how machines learn from experience. The goal of this study is to create a system that, by fusing the findings of several machine learning approaches, can more accurately conduct early diabetes prediction for a patient. K closest neighbour, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree are some of the techniques employed. Each algorithm's accuracy is calculated along with the model's accuracy. The model for predicting diabetes is then chosen from those with good accuracy.

Keywords: Machine Learning, Diabetes, Decision tree, K nearest neighbor, Logistic Regression, Support vector Machine, Accuracy.

I. INTRODUCTION

Even among children, diabetes is a condition that is spreading quickly. We must comprehend what occurs in the body without diabetes if we are to comprehend diabetes and how it arises. Sugar (glucose) is derived from the meals we eat, notably those high in carbohydrates. Everybody requires carbs, including those with diabetes, as they are the body's primary source of energy. Bread, cereal, pasta, rice, fruit, dairy products, and vegetables (particularly starchy vegetables) are all examples of foods that include carbohydrates. These foods are converted into glucose by the body when we eat them. In the bloodstream, glucose circulates throughout the body. For us to think effectively and function, some of the glucose is transported to our brain. The remaining glucose is transferred to our body's cells for usage as fuel, and it is also stored as energy in our liver for later use by the body. Insulin is necessary for the body to use glucose as fuel. The beta cells in the pancreas create the hormone insulin. Insulin functions as a door's key. In order to allow glucose to enter the cell from the blood stream, insulin attaches to the cell's doors, opening them. Glucose builds up in the bloodstream (hyperglycemia) and diabetes occurs if the pancreas is unable to generate enough insulin (insulin deficit) or if the body is unable to utilise the insulin it produces (insulin resistance). Diabetes Mellitus is characterised by elevated glucose (sugar) levels in the blood and urine.

There are millions of people affected by diabetes worldwide, which is one of the deadliest chronic diseases that can adversely affect every system of the body. Diabetes affects 422 million people worldwide, or 8.5% of the world's population, according to WHO (who.int, 2019).

Diabetes worsens kidney and nerve damage, heart and blood vessel disease, sluggish wound healing, hearing loss, and a number of skin problems. 1.6 million fatalities from diabetes were reported in 2016. Before the age of 70, over half of all deaths took place. Diabetes was the seventh most common cause of death in 2016, according to the WHO.

According to research by Katulanda and colleagues, South East Asia will have the greatest concentration of diabetes patients worldwide by the year 2025 (Katulanda et al., 2009). It's crucial to detect diabetes early if you want to live a healthy life. Type 2 diabetes can be prevented or delayed by following a balanced diet, engaging in regular exercise, maintaining a healthy weight, and giving up alcohol and smoking. Diet, exercise, medication, and therapy for complications can all help control diabetes and delay or prevent its effects.

Machine learning is an application of artificial intelligence that gives systems the capacity to learn from their own experience and advance without explicit programming. The Support Vector Machine (SVM), Decision Tree, Random Forest (RF), Naive Bayes, and Neural Network classification algorithms were employed in earlier studies to predict diabetes. The "Pima Indian Diabetes Dataset" (PIDD) was the dataset used in nearly all previous investigations. The accuracy and precision of those distinct Machine Learning models' performance have been examined.

The machine learning model was constructed using a fresh dataset, and the associated risk factors were first identified.

Gender, age, BMI, waist measurement, frequency of exercise and consumption of fruits and vegetables, high blood pressure, and risk score have all been identified as risk factors. This study intends to advance health care in Sri Lanka by identifying diabetes early and offering suggestions for maintaining a healthy lifestyle.

II. RELATED WORK

In order to determine if a person has diabetes or not, Yasodhaet al.[1] classify many sorts of datasets. The hospital's data warehouse, which has 200 instances with nine attributes, was used to create the data set for the diabetic patient. Both blood tests and urine tests are mentioned in these instances of the dataset. Due to its excellent performance on tiny datasets, WEKA may be used in this study's implementation to categorise the data. The data is then evaluated using the 10-fold cross validation approach, and the results are compared. We employ the naive Bayes, J48, REP Tree, and Random Tree. The results showed that among the others, J48 works best, with an accuracy of 60.2%.

By researching and examining the patterns that emerge in the data via classification analysis utilising Decision Tree and Nave Bayes algorithms, Aiswaryaet al.'s [2] goal is to find ways to diagnose diabetes. The goal of the study is to suggest a quicker and more effective technique of diagnosing the illness, which will aid in the timely treatment of the patients. The study found that the J48 method provides an accuracy rate of 74.8% while the naive Bayes provides an accuracy of 79.5% by adopting a 70:30 split, using the PIMA dataset and cross validation approach.

Gupta et al.'s [3] study compares the performance of the same classifiers when implemented on some other tools, including Rapidminer and Matlab, using the same parameters (i.e. accuracy, sensitivity, and specificity). The study aims to find and calculate the accuracy, sensitivity, and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA. The JRIP, Jgraft, and BayesNet algorithms were used. According to the results, Jgraft has the highest accuracy (81.3%), sensitivity (59.7%), and specificity (81.4%). Additionally, it was found that WEKA performs better than Matlab and Rapidminner.

Following the application of the resample filter to the data, Lee et al.'s [4] main focus is on using the decision tree algorithm CART on the diabetes dataset. The author places a strong emphasis on the issue of class imbalance and the need to address it before implementing any method in order to increase accuracy rates. The class imbalance is most frequently found in datasets with dichotomous values, which implies that the class variable has two alternative outcomes and may be readily managed if discovered earlier in the data preprocessing stage. This will also assist to increase the predictive model's accuracy.

The World Health Organisation reports that diabetes is now the biggest cause of death globally (Who.int, 2019). The majority of diabetic patients live in low- and middle-income nations. Numerous studies have been conducted explicitly on the application of machine learning and neural networks in the diagnosis of diabetes mellitus, and several approaches are addressed along with their objectives, tools and techniques employed, outcomes, and conclusions.

In their research work, Kaur and Kumari (Kaur and Kumari, 2018) presented "Predictive modelling and Analytics for Diabetes using Machine Learning." The main goal of that study was to determine which of five predictive models—the "Linear Kernel" and "Radial Basis Function" (RBF), "Multifactor Dimensionality Reduction" (MDR), "k-Nearest Neighbour" (kNN), "Kernel Support Vector Machine" (SVM), and "Artificial Neural Network" (ANN)—was the most accurate at predicting diabetes mellitus. The Pima Indian Diabetes Dataset, which was originally owned by India's "National institute of diabetes and digestive and kidney diseases," has been examined using the R data processing tool. There are 768 cases in this dataset that are divided into diabetes and nondiabetic groups. There are also eight other risk factors. They used 70% training data to train their model, and the remaining 30% was used for testing. These five alternative models were created using the supervised learning techniques discussed above, and they have all been tested in the R programming environment.

(Zou et al., 2018) Zou and others "Prediction of Diabetes Mellitus with Machine Learning Techniques" was explored in their academic paper." The primary goal the purpose of this investigation was to determine which is the precise machine learning method to diabetes mellitus is predicted. Scientists have got two distinct datasets with the names Pima dataset and the Luzhou dataset. Luzhou hospital physical acquired the dataset data from an examination in Luzhou, China. And this dataset consists of two sections: healthy individuals, diabetic individuals, and it includes 14 distinct examination indexes. Furthermore, the Pima dataset and the earlier described studies.

III. METHODOLOGY

Types of Diabetes

Diabetes Subtypes When a person has type 1 diabetes, their immune system is weakened and their cells are unable to make enough insulin. There are no convincing studies that demonstrate the causes of type 1 diabetes, and there are also no effective preventative measures at this time. Type 2 diabetes is characterised by either insufficient insulin production by the cells or improper insulin utilisation by the body. 90% of people with diabetes have this kind of diabetes, making it the most prevalent type. Both genetic and lifestyle factors contribute to its occurrence.

Symptoms of Diabetes

- Frequent Urination
- Increased thirst
- Tired/Sleepiness
- Weight loss
- Blurred vision
- Mood swings
- Confusion and difficulty concentrating
- frequent infections

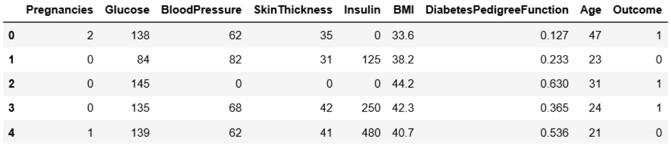
The primary cause of diabetes is genetics. It is brought on by at least two defective genes on chromosome 6, a chromosome that influences how the body reacts to different antigens. The development of type 1 and type 2 diabetes may also be influenced by viral infection. According to studies, having viruses such hepatitis B, CMV, mumps, rubella, and coxsackievirus increases the risk of acquiring diabetes.

We will learn about the different classifiers used in machine learning to predict diabetes in this part. We will also describe the technique we have suggested in order to increase accuracy. In this paper, five alternative methodologies were used. The various techniques are described below. The machine learning models' accuracy measurements are the output. The model can then be applied to make predictions.

Dataset Description

The diabetes data set was originated from https://www.kaggle.com/johndasilva/diabetes.

Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

Table 1. Dataset Description

The dataset contains a total of 8 attributes (features) and 1 binary target variable (outcome). The features are:

Pregnancies: Number of times pregnant.

Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

Blood Pressure: Diastolic blood pressure (mm Hg).

Skin Thickness: Triceps skinfold thickness (mm).

Insulin: 2-Hour serum insulin (mu U/ml).

BMI: Body mass index (weight in kg/(height in m)^2).

Diabetes Pedigree Function: Diabetes pedigree function, which provides information about the genetic influence of diabetes among relatives.

Age: Age in years.

The majority of the data that was gathered could be skewed by carelessness. In addition, data quality is critical because it significantly influences prediction accuracy and results [4]. So that sampling may be done effectively for improved prediction results, datasets need to be correctly balanced and divided between testing and training data at a specific ratio. Sampling is the process of choosing a representative subset of data in order to systematically extract traits and parameters from big datasets; as a result, it can improve the machine learning training model. We need to apply several sampling strategies (linear sampling, shuffled sampling, stratified sampling, and automatic sampling) on the dataset in order to retain that consistency. These sampling approaches test the prediction model by randomly dividing the dataset into subsets. These various sampling methods combine and permute a representative collection of data from the gathered data in ways that aren't exactly like what is depicted here.

The dataset is divided into divisions using a linear sampling technique to represent the dataset. Additionally, it keeps the order of the tuples and fields in the subsets unchanged.

Shuffled sampling: This sampling technique divides the applicable dataset into subsets at random. data assembled from randomly chosen subsets.

Stratified sampling: This sampling method creates subsets by randomly dividing the dataset. However, the method also demonstrates that the class distribution over the dataset should be constant. For instance, stratified sampling approach constructs produce arbitrary subgroups in such a manner that each subset has about the same proportions of the two values of class labels if the employed dataset used binominal classification.

Automatic: Depending on the characteristics of the dataset, the automated sampling method defaults to stratified sampling. If the method doesn't fit the type of data, a suitable one is used instead.

IV. PROPOSED WORK

The major goal of this study is to present the most promising characteristics that are required to early detect diabetes in a patient. The intrusive automated discovery of diabetes has been the subject of a significant amount of research. To acquire the best results, it therefore depends entirely on the characteristics that were extracted and the sort of classifier that was used. Therefore, it has been determined through analysis of diversity of learning that these dataset features can be used to categorise various risk factors for prophesy [6]. The comprehensive research done on the PIMA dataset is presented in this report. A well-organized format for the discovery of diabetes will be maintained by the association of several classification algorithms that have been tested on various strata, as well as management of their risk factors and treatment approaches for medical professionals.

The two key components of the suggested methodology are model validation and how accuracy is attained utilising various categorization models. Different machine learning approaches are useful for analysing hidden patterns and determining risk factors for diseases like diabetes. Furthermore, it has been noted that a high data dimension prevents the presentation of conventional approaches from reaching the level of acceptance in voice and object recognition [4]. Machine learning algorithms' shortcomings fueled DL research, which now outperforms other algorithms in terms of accuracy and tends to deliver more reliable results. By utilising DL in anomaly detection, extensive study has been conducted in the field of healthcare. On the PIMA dataset, our suggested model, which is related to diabetes prediction, attained the best accuracy to date, or 98.07%.

On the PIMA dataset, four data mining algorithms—DT, NB, ANN, and DL—are used to assess efficiency, which is closely correlated with reliable results. Based on a task connected to the diabetes illness forecast, our suggested approach was selected. Rapid Miner offers a user-friendly and interactive Graphical user interface for quickly and accurately putting together prediction models and pre-processing data. Because of this, Rapid Miner Studio 9.2.000 has been employed in our suggested methodology. It contains a variety of functions, including drag and drop, wisdom of crowds, and many more, for in-person suggestions during the workflow. For many data mining features that Weka does not offer, Rapid Miner offers 400 extra operators. These 400 extra operators include a variety of categorization approaches, pre-processing methods, validation strategies, and visualisation techniques not found in Weka. All of the work has been done on the rapid miner tool since it has a very convenient user interface and is quicker for researchers to use than programming languages [7]. Rapid miner provides additional benefits that are further highlighted, aside from the interface.

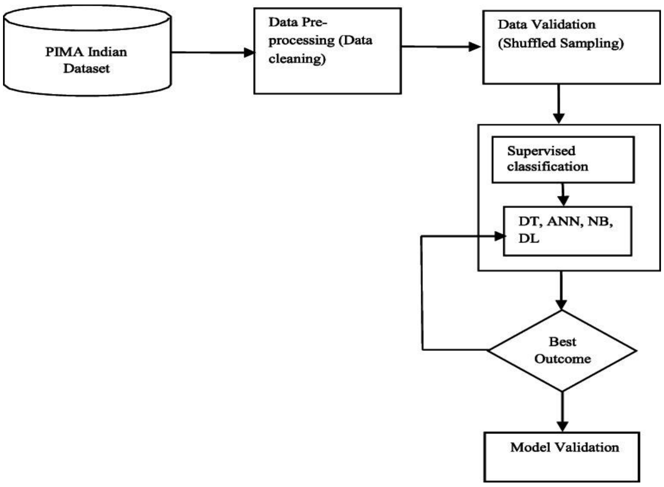

PIMA Indian Dataset
Data Pre-processing (Data cleaning)
Data Validation (Shuffled Sampling)
Supervised classification
DT, ANN, NB, DL
Best Outcome
Model Validation

Fig 1. Architecture of the work.

Because quick miners employ the same dataflow as tree-based structures, usability can be seen as one of their primary benefits. For large-scale data mining, which is a bit complex and challenging in a graph-based structure, it ensures the automatic validations and optimisation. The quick miner's efficiency is its second strength, since users have noted that it is more effective than Weka at handling larger datasets while using less memory [2]. The proposed model's flowchart is depicted in Figure 1.

V. DEEP LEARNING ARCHITECTURE

Machine learning is a broad artificial intelligence technique that analyses associations in data without being explicitly coded or specifying the historical link between the data parts [4, 5]. DL is a type of machine learning that differs from conventional approaches in that it learns from a variety of raw data representations. It enables various computational models built on ANN that have several processing layers to handle and represent data at different degrees of abstraction [8].

DL is a multilayer feed-forward perceptron-based model trained with stochastic gradient descent via back-propagation, which also facilitates the characteristics of ANN. The network is made up of four layers that act as nodes and neurons and are connected in a single direction (uni-direction). Each node has two hidden layers and a single way link to the next node. Through the use of its local data, each node trains a replica of the global model parameters. Additionally, it processes the model using many threads and applies averaging to let the model be accessible across the entire network. The learning model employs backpropagation and hidden layer neurons along with stochastic gradient descent training to enable more sophisticated characteristics including tanh, rectifier, and maxout activation, learning rate, and rate annealing. Maxout produces the most noticeable benefits out of all the activation techniques. As shown in Fig. 5, our suggested model used a DL neural network with one input layer for data entry, one output layer for prediction results, and two hidden layers for iteratively processing the information.

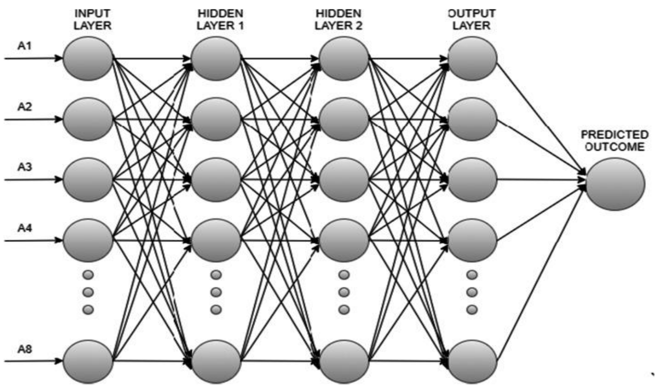
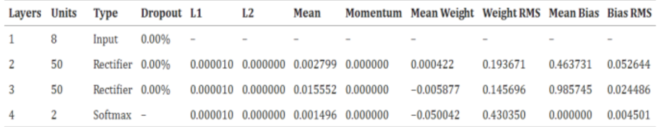


Figure 2. Multilayer DL neural network used as a prediction model

One of the most difficult problems in the machine learning model implementation is model optimisation or parameter setting. Model optimisation often refers to tweaking the code to reduce testing error, whereas deep learning optimises its model by adjusting external factors that have a significant impact on its behaviour and categorization. Advanced features like adaptive learning rate, mean bias, momentum training, dropout, and L1 or L2 regularisation have been taken into consideration for reducing the testing error because the criterion for parameter sets are variable and hidden [23]. In Table Table22, some of the important characteristics are discussed.



Layers	Units	Type	Dropout	L1	L2	Mean	Momentum	Mean Weight	Weight RMS	Mean Bias	Bias RMS
1	8	Input	0.00%	–	–	–	–	–	–	–	–
2	50	Rectifier	0.00%	0.000010	0.000000	0.002799	0.000000	0.000422	0.193671	0.463731	0.052644
3	50	Rectifier	0.00%	0.000010	0.000000	0.015552	0.000000	−0.005877	0.145696	0.985745	0.024486
4	2	Softmax	–	0.000010	0.000000	0.001496	0.000000	−0.050042	0.430350	0.000000	0.004501

Table 2. Key parameters used in the DL model optimization.

The speed of learning progress in a model is measured by the learning rate, which is referred to as the mother of all hyperparameters. The next deep learning method parameter is the number of hidden units. This is a fundamental parameter in deep learning algorithms since it controls the model's capacity for representation. L1 & L2 Regularisation is a further parameter that is essential for avoiding model overfitting. In order to reduce the complexity of the model and address the overfitting, feature selection, several regularisation techniques are used. A model is referred to as Lasso Regression if it employs the L1 regularisation approach, and as Ridge Regression if it employs the L2 regularisation technique [9].

The "absolute value of magnitude" of the coefficient is added as a regularisation component to the loss function in Lasso Regression (Least Absolute Shrinkage and Selection Operator) or L1 regularisation to prevent underfitting. It may be estimated using (1), and a further benefit of Lasso Regression is that it trains the model with the most significant parameters by shrinking the less significant parameters to zero.

Decision Trees

The DT graph, which shows the results as a splitting rule for each particular attribute, is used in decision analysis. It is a branching graph that can be used to expressly and aesthetically represent the results of decisions. Every attribute is viewed as a branching node that builds a rule at the tip of the branch to separate values into several classes. As its name implies, it has a tree-like structure and ends with a decision, which is referred to as the tree's leaf. The root is the attribute that has the greatest potential for use in predicting how a rule will turn out. Along with these benefits, a DT is straightforward and quick to use, and it also makes better predictions of the outcomes [3]. Iteratively building new nodes continues until a base condition is not satisfied. The largest value of the rule, which results in the leaf node during the DT analysis, is the sole basis for the class label attribute [2]. Decision trees are vulnerable to overfitting since they are built upside down with their roots at the top. Pre-pruning is the process of removing the leaves that are not significant and meaningful for tree construction since overfitting is an issue when a tree gets over skillful with data and its leaf displays minimum impurity. For the creation of a DT model, pre-pruning stipulates that the base criteria should be greater than the depth of the tree. Pre-pruning also contributes to improved prediction accuracy. Information gain, which is suitable to be more accurate for the prediction of outcomes from the other criteria, is another crucial DT split criterion. Entropy is determined for each property in this manner, and the attribute with the lowest entropy is chosen for the split.

Naive Bayes

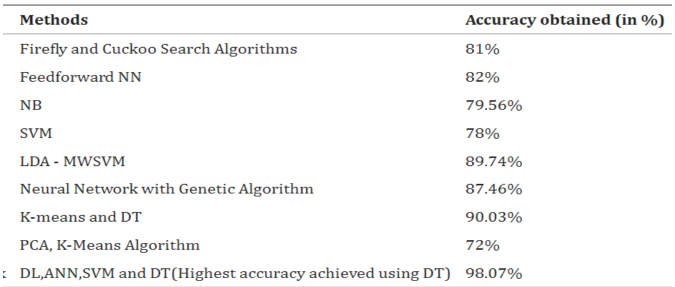
NB is a DT based supervised classification algorithm [35] which only differs in the representation of its outcome. Where the DT provides the rules at the end, NB defines the probability. Both algorithms are used for prediction purposes. Moreover, NB provides a conditional probability. The major advantage of the NB is that it can deal with a small dataset and its high passes the low variance classifier which works using Bayes theorem and finds the feasibility of the attribute associated with an object by using the important information. Along with this, it is easy to implement and computationally low-priced. In NB all the attribute values are independent of each other, therefore, it is inexpensive in computation and separately simplifies the assumption and calculation using (5). In Naive Bayes classifier parameter tuning and optimization is limited [6].

Artificial neural network

ANN is another technique for classification which is a machine learning algorithm and provides more accurate results in comparison with the existing algorithms. It is a mathematical model that is inspired by the functioning and structure of biological neurons. A neural network is a connection of multiple neurons connected as the human brain is a connection of 86 billion biological neurons. The functional connectivity in artificial neurons is mesh connectivity and each neuron has equal weight [7]. The interconnectivity of neurons works on the principle of the connectionist approach (the principle of connectionist follows that the mental phenomena described by the simple and uniform connectivity of neurons). Along with this ANN consists of one or more hidden layers that process the information through neurons and each node works as an activation node; it classifies the outcome of artificial neurons for a better outcome. The major finding of an ANN is that it finds the complex relationships between data and draws useful patterns [8].

VI. RESULT AND DISCUSSION

In this research work, outcomes were achieved by applying four classification algorithms (DL, ANN, NB, and DL) to display maximize accuracy in diabetes prediction. From these four classifiers, DL and DT provide promising accuracy (98.07%) which can be proven as a prominent tool for the prediction of diabetes at an early stage. In our proposed system we use the PIMA dataset and apply it on a DL approach. Further, it can help the healthcare practitioner and can be the second estimation for the betterment of decisions depending on extracted features [9]. Many researchers have been previously worked on the PIMA dataset with a diverse algorithm to predict diabetes. Thus some of the researcher's work has been represented with their applied methods and achieved accuracy. Table Table33 shows all the promising work done on Pima dataset till time and our proposed method achieved the highest accuracy i.e. 98.7 on PIMA Indian dataset.



Methods	Accuracy obtained (in %)
Firefly and Cuckoo Search Algorithms	81%
Feedforward NN	82%
NB	79.56%
SVM	78%
LDA - MWSVM	89.74%
Neural Network with Genetic Algorithm	87.46%
K-means and DT	90.03%
PCA, K-Means Algorithm	72%
DL,ANN,SVM and DT(Highest accuracy achieved using DT)	98.07%

Table 3. Comparative study of related research works for diabetes detection with Pima Indian dataset

Thus these additional measures are Class recall, class precision, and F-measure. Class recall can be described as the number of attributes, which were classified correctly. It can be explained in another way, that it is the number of total positive predictions divided by the number of total positive class values, It can also be called Sensitivity or the True Positive Rate as represented below.

Recall=TruePositives/TruePositives+FalseNegatives

The second measure for performance evaluation is Class Precision, It can be defined as the sum of true positive and true negative. In another way, it is the number of True Positives predictions divided by the number of True Positives and False Positives as shown using below.

Precision=TruePositives/TruePositives+FalsePositives

Another measure for performance evaluation is F-measure or-F-score, it conveys the balance between the recall and prediction. The formula for representing the F-score is given as.

F–Score=2∗precision∗recall/precision+recall

The accuracy obtained through diverse classifiers is shown below by the confusion matrix which consists of class precision, diabetes prediction yes, diabetes prediction no, class recall. Another performance measure could be Specificity, which is the proportion of values without the disease who test negative. In the form of probability, notation Sensitivity could be calculated using.

PT–D–=TN/TN+FP

Actual Values Predicted Values	True No	True Yes	Class Precisior
Predicted No	137	4	97.16%
Predicted yes	3	63	95.45%
Class recall	97.86%	94.03%	

Table 4. Decision Tree (Accuracy: 96.62%)

Actual Values Predicted Values	True No	True Yes	Class Precision
Predicted No	118	27	81.38%
Predicted yes	22	40	64.52%
Class recall	84.29%	59.70%	

Table 5. shows the outcomes of the Naive Bayes Classifier having an accuracy of 76.33%.

Actual Values Predicted Values	True No	True Yes	Class Precisi
Predicted No	128	8	94.12%
Predicted yes	12	59	83.10%
Class recall	91.43%	88.06%	

Table 6. shows the outcomes of Neural Network having an accuracy of 90.34%.

Actual Values Predicted Values	True No	True Yes	Class Precision
Predicted No	139	3	97.89%
Predicted yes	1	64	98.46%
Class recall	99.29%	95.52%	

Table 7. shows the accuracy of deep learning architecture at 98.07%.

Table Table88 represents the four performance measures (Accuracy, Precision, Recall, F-measure) for all classification algorithms that are applied to PIMA dataset for diabetes prediction. This knows that DL outperforms in all the performance parameters and provides the best results for diabetes onset with an accuracy of 98.07%. Figures 6 and and77 shows the comparison between the performance matrices of diabetes prediction technique.

Measures	Methods			
	DL	DT	ANN	NB
Accuracy (%)	98.07	96.62	90.34	76.33
Precision (%)	95.22	94.02	88.05	59.07
Recall (%)	98.46	95.45	83.09	64.51
F-Measure (%)	96.81	94.72	85.98	61.67
Specificity (%)	99.29	97.86	91.43	84.29
Sensitivity (%)	95.52	94.03	88.06	59.70

Table 8. Performance Evaluation of Diabetes Prediction techniques

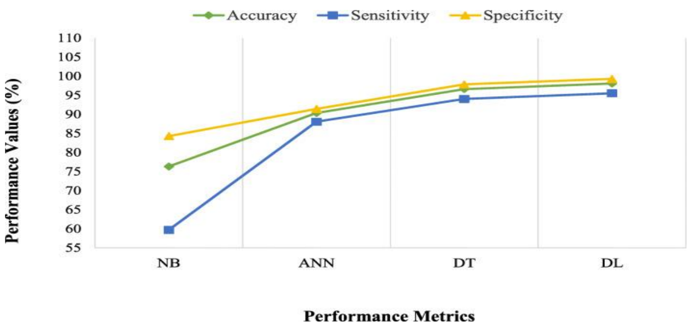


Fig. 6. Comparison of Accuracy, Sensitivity, and Specificity for Various Classification Methods

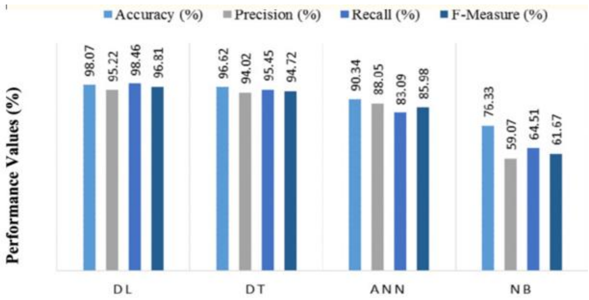


Figure 3. Comparison of Accuracy, Precision, Recall and F-score for Various Classification Methods.

As shown in the above Figs. 6 and and77 DL provides the highest accuracy among all algorithms and proven to be the best classifier algorithm for diabetes prediction. The accuracy of 98.07% has been achieved on the PIMA dataset which is the highest accuracy obtained to date. The maximum accuracy can be obtained by consequential and significant data collection. Those attributes that don't contribute to the classification outcome should be prune. In this study, we have some facts about the classification algorithm that information gain gives better results in DT classifier and activation should be max out at the time of functioning in DL for a better outcome.

VII. CONCLUSION AND FUTURE WORK

The goal of this study was to put into practise a prediction model for diabetes risk assessment. As was already mentioned, diabetes affects a sizable portion of the human population. If left unchecked, it poses a serious risk to the entire planet. Therefore In our suggested research, we have demonstrated that data mining and machine learning algorithms may reduce risk factors and improve the outcome in terms of efficiency and accuracy by putting several classifiers into practise on the PIMA dataset. The outcome on the PIMA Indian dataset utilising the data mining algorithms presented in Table Table1.1 is better than other recommended methodologies on the same dataset. The four classifiers' (DT, ANN, NB, and DL) accuracy ranges from 90 to 98 percent, which is significantly higher than that of other techniques. With an accuracy percentage of 98.07%, DL is regarded as the most effective and promising classifier for analysing diabetes among the four that have been offered. The proposed DL algorithm can be used in the future to assist healthcare professionals in the early detection of diabetes. This system will likely take the shape of an app or website.

VIII. REFERENCES

[1]. "Global Report on Diabetes, 2016". Available at: https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=2BC28035503CFAFF295E70CFB4A0E1DF?Sequence=1.

[2]. "Diabetes: Asia's 'silent killer'", November 14, 2013". Available at: ww.bbc.com/news/world-asia-24740288.

[3]. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. 2015;3(11). 10.1371/journal.pmed.0030442.

[4]. Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4):243–246. doi: 10.1016/j.icte.2018.10.005.

[5]. Wu H, et al. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 2018;10:100–107. doi: 10.1016/j.imu.2017.12.006.

[6]. Ravindra Changala ,Development of Predictive Model for Medical Domains to Predict Chronic Diseases (Diabetes) Using Machine Learning Algorithms And Classification Techniques, ARPN Journal of Engineering and Applied Sciences, Volume 14, Issue 6, 2019.

[7]. Ravindra Changala, "Evaluation and Analysis of Discovered Patterns Using Pattern Classification Methods in Text Mining" in ARPN Journal of Engineering and Applied Sciences, Volume 13, Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018.

[8]. Ravindra Changala "A Survey on Development of Pattern Evolving Model for Discovery of Patterns in Text Mining Using Data Mining Techniques" in Journal of Theoretical and Applied Information Technology, August 2017. Vol.95. No.16, ISSN: 1817-3195, pp.3974-3987.

[9]. Emerging T, Factors R. Diabetes mellitus , fasting blood glucose concentration , and risk of vascular disease : a collaborative meta-analysis of 102 prospective studies. The Lancet. 2010;375(9733):2215–2222. doi: 10.1016/S0140-6736(10)60484-9.

[10]. Zhang LM. Genetic deep neural networks using different activation functions for financial data mining. In: Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015; 2015. p. 2849–51. 10.1109/BigData.2015.7364099.

[11]. Grundy SM. Obesity, Metabolic Syndrome , and Cardiovascular Disease. 2004;89(6):2595–600. 10.1210/jc.2004-0372.

[12]. Palaniappan S. Intelligent heart disease prediction system using data mining techniques, (march 2008). 2017. 10.1109/AICCSA.2008.4493524.

[13]. Craven MW, Shavlik JW. Using neural networks for data mining. Futur Gener Comput Syst. 1997;13(2–3):211–229. doi: 10.1016/s0167-739x(97)00022-8.

[14]. Radhimeenakshi S. 2016 International Conference on Computing for Sustainable Global Development (INDIACom) 2016. Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural networks; pp. 3107–3111.

[15]. Ravindra Changala, "A Generalized Association Rule Mining Framework for Pattern Discovery" published in Ravindra Changala et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Impact Factor 2.93 in Vol. 5 (4) , 2014, pp:5659-5662, ISSN: 0975-9646,August 2014.

[16]. El-Jerjawi NS, Abu-Naser SS. Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology. 2018;121:55–64. doi: 10.14257/ijast.2018.121.05.

Cite this article as :

Appana Naga Lakshmi, Durga Thokala, Soma Ganesh Kumar, Bollapally Rakesh Reddy, Dr. G Vinoda Reddy, "An Effective Machine Learning Approach for Diabetes Prediction", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 4, pp.325-335, July-August-2023. Available at doi : https://doi.org/10.32628/CSEIT2390442
Journal URL : https://ijsrcseit.com/CSEIT2390442