IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

ANALYSIS OF DIABETES USING MACHINE LEARNING AND NEURAL NETWORK

Hasti Pareshbhai Patel
Department of CSE
ITM(SLS) Baroda University

Vedantvaibhav Thakur
Department of CSE
ITM(SLS) Baroda University

Prof. Pradeep C.
Geetanjali Institute of Technical Studies, Udaipur

Abstract : Diabetes is a chronic condition that 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[1]. 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, Linear Regression, Random Forest, J48, IBK, ANN, Multilayer Preceptron ,Naïve Bayes ,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, Accuracy, Weka Software, Coding

1. Introduction

Diabetes is the fast growing disease among the people even among the youngsters. Symptoms of Diabetes are Frequent Urination, Increased thirst ,Tired/Sleepiness , Weight loss or increase , Blurred vision , difficulty concentrating ,frequent infections, Itching etc and These symptoms may occur suddenly. Symptoms for type 2 diabetes are generally similar to those of type 1 diabetes but are often less marked. As a result, the disease may be diagnosed several years after onset, after complications have already arisen. For this reason, it is important to be aware of risk factors [2].

This computation is done on Diabetes datasets on available repository datasets from the Sylhet Diabetes Hospital in Sylhet, Bangladesh. We have implemented different classification methods to classify the data to detect whether a patient have diabetes or not .We used different machine learning techniques such as Logistics Regression ,Linear Regression, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), J48, IBK, Multi-Layer Perception, Artificial Neural Networks and Naive Bayes (NB). The classification of Diabetes and the performance of these techniques were estimated using metrics such as accuracy, precision, recall, f-1 score, Kappa Statistics, Sensitivity, Specificity. ML algorithms evaluated based on the basis accuracy and ROC curve of Logistic Regression. Weka 3.8.4 Software[3] and Google Colab as a classification tool.

2.Methods and Material

2.1. Data Collection

In this article, we used dataset is originally from the Sylhet Diabetes Hospital in Sylhet, Bangladesh[4]. This dataset contains 520 'diabetes patients' records and 17 attribute. 61.5% samples are Positive and 38.5% samples are Negative. The size of datasets 34KB.

2.2. Machine Learning Classifier

Once the data were pre-processed and all the anomalies had been managed, they are now sent to the machine learning classifiers to train an ML model which could classify Whether a person has Diabetes or not by their Symptoms. In this section, brief information about the machine learning classifiers considered for the research is discussed[5,6,7,8,9].

2.2.1 Support Vector Machine(SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm. It can be applied to classification or regression tasks. However, it is preferred for classification problems. In the SVM algorithm, each data item is plotted as a point in x-dimensional space (where x is equal to the number of features) with the value of each feature being the value of the particular coordinate. As a solution to separate the two classes of the data points, many possible hyperplanes may be applied. Here the objective is to find a plane that has the maximum distance between data points of each class. By a maximization of a margin distance, it is provided with some reinforcement so that future data points can be classified with more confidence. The loss function that helps maximize the margin is hinge loss[9,10].

2.2.2 J48

J48 represents an open source Java implementation of the C4.5 algorithm. that is

Algorithm for generating decision trees developed by Ross Quinlan. that is

An extension of Quinlan's earlier ID3 algorithm. In this case a decision tree is generated.

C4.5 is often called statistical because it uses C4.5 for classification.

Select attributes from a set of training instances, then select initial values.

A subset of training instances. Instance attributes and subsets are now used

Build a decision tree. Remaining training instances (those not in the used subset)

for construction) is used to test the accuracy of the constructed tree. Repeated

until the trees are made. C4.5 uses information retrieval rate to select the best attributes

distinguish between instances. [11]

2.2.3. Artificial Neural Network

Artificial Neural Networks (ANN) are a type of machine learning model that are inspired by the structure and function of the human brain. They consist of interconnected nodes, or artificial neurons, that process information and make predictions based on input data. There are several key components to an ANN.

1. Hidden layers: These are the intermediate layers between the input and output layers. They consist of artificial neurons that use weighted sums of the input data to produce intermediate output values.
2. Output layer: This layer generates the final prediction based on the values computed in the hidden layers.
3. Activation functions: These are mathematical functions that determine the output of each artificial neuron. Common activation functions include the sigmoid function, the rectified linear unit function, and the hyperbolic tangent (tanh) function.

The weights and biases of the artificial neurons in an ANN can be adjusted during the training process to minimize the difference between the predicted and actual outputs. This process is known as backpropagation and is typically done using gradient descent.[12,13,14,15]

2.3. Performance Evaluation

We employed statistical analysis to evaluate the test results of several metrics in the task. Various measures, including accuracy, recall, precision, and f1 measure, were used to assess the effectiveness of the categorization techniques. Machine learning metrics include the following:

True Positive (TP): When a prediction's outcome accurately detects the presence of Diabetes in a patient.

False Positive (FP): When a patient's diagnosis of Diabetes is made erroneously as a result of a prediction.

True Negative (TN): The prediction's outcome successfully disproves the presence of Diabetes in the patient.

False Negative (FN):results occur when the prediction wrongly rules out the presence of Diabetes in the patient.

In order to evaluate the prediction performance of classification algorithms, we define and compute the classification accuracy, precision, recall and F1 score, respectively.

2.3.1 Accuracy

The classification accuracy of an ML classifier is the solution to measure how often the algorithm classifies a data point correctly. The accuracy informs about the number of correctly predicted data points out of all the data points, which is evaluated as follows:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP+TN+FP+FN)}$$

2.3.2 Sensitivity

The test sensitivity is named the true positive rate (TPR). It concerns the proportion of samples that are genuinely positive that give a positive result using the test in question. It also concerns type II errors; false negatives are the failures to reject a false null hypothesis.[16]

Sensitivity = TP/FN + TP

2.3.3 Specificity

The test specificity is named the true negative rate (TNR). It concerns the proportion of samples that test negative using the test in question that are genuinely negative. Additionally, it is referred to as type I errors, false positives are the rejection of a true null hypothesis. It is evaluated as follows: [16]

Specificity = TN/FP + TN

2.3.4 Precision

Precision measures the percentage of instances or samples that are accurately classified among those that are classed as positives. As a result, the precision calculation formula is as follows:[17]

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

2.3.5 Recall

The recall is determined as the proportion of Positive samples that were correctly identified as Positive to all Positive samples.

The recall gauges how well the model can identify positive samples[17].

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

2.3.6 F1-Score

F1 score is define as the average of precision and recall. [18]

$$F1 = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

2.3.10 Epoch

Epoch Means one cycle through the full training dataset. Usually, training a neural network takes more than a few epochs. In other words, if we feed a neural network the training data for more than one epoch in different patterns, we hope for a better generalization when given a new "unseen" input (test data). An epoch is often mixed up with an iteration. Iterations is the number of batches or steps through partitioned packets of the training data, needed to complete one epoch. [19]

3.Analysis of Result:

Table 1 Accuracy With Preprocessing using Coding

Model Name	Accuracy
KNN	85%
SVM(Kernel = Linear)	92%
Artificial Neural Network(ANN)	93.27%
Logistic Regression	92%
Naïve Bayes	91%
Linear Regression	66%
Decision Tree	96%

Table 2 Kappa Statistics, Sensitivity and Specificity With Preprocessing using Coding

Model Name	Kappa Statistics	Sensitivity	Specificity
KNN	68%	83%	90%
SVM(Kernel = Linear)	82%	91%	93%
Logistic Regression	81%	95%	84%
Naïve Bayes	79%	94%	84%
Decision Tree	91%	94%	100%

Table 3 Accuracy, Kappa Statistics, Sensitivity and Specificity of SVM Kernels With Preprocessing using Coding

SVM Kernels	Accuracy	Kappa Statistics	Sensitivity	Specificity
Linear	92%	82%	91%	93%
Polynomial	90%	78%	91%	87%
Radial basis function(rbf)	68.26%	0%	100%	0%
Naïve Bayes	68%	0%	100%	0%

Table 4 Accuracy and Kappa Value Using with Preprocessing Weka Software

Model Name	Accuracy	Kappa Statistics
48	5.96%	1%
BK	5%	9%
Multi Layer Preceptron	8%	3%
Naïve Bayes	7%	3%
Logistic Regression	2%	3%
Random Forest	6%	1.92%

Table 5 Precision, Recall, F1-Score With pre-processing Using Coding

MODEL NAME		Prec-ision	Recall	F1-SCORE
KNN	Positive	5%	3%	9%
	Negative	1%	91%	0%
Logistic Regression(LT)	Positive	3%	6%	4%
	Negative	0%	5%	8%
Naïve Bayes	Positive	3%	4%	4%
	Negative	8%	5%	6%
Decision Tree	Positive	00%	4%	7%
	Negative	9%	00%	4%
SVM Kernel="Linear"	Positive	7%	2%	4%
	Negative	4%	4%	9%
SVM Kernel="Poly"	Positive	4%	2%	3%
	Negative	3%	8%	5%
SVM Kernel="rbf"	Positive	8%	00%	1%
	Negative	%	%	%
SVM Kernel="Sigmoid"	Positive	8%	00%	1%
	Negative	%	%	%

Table 6 Precision, Recall, F1-Score with Preprocessing Using Weka Software

MODEL NAME		Prec-ision	Recall	F1-SCORE	MCC	ROC Area	PRC Area
J48	Positive	98%	95%	96%	91%	96%	97%
	Negative	92%	97%	94%	91%	96%	91%
IBK	Positive	96%	95%	96%	89%	97%	98%
	Negative	92%	94%	93%	89%	97%	93%
Multi Layer Preceptron	Positive	86%	97%	91%	75%	98%	99%
	Negative	93%	73%	82%	75%	98%	97%
Naive Bayes	Positive	92%	85%	89%	73%	94%	96%
	Negative	79%	89%	84%	73%	94%	90%
Logistic Regression	Positive	94%	93%	93%	83%	96%	98%
	Negative	89%	90%	90%	83%	96%	93%
Random Forest	Positive	97%	95%	96%	91%	96%	96%
	Negative	93%	96%	95%	91%	96%	91%

Table 7 Time Taken To Build a Model with Pre-Processing Using Weka Software

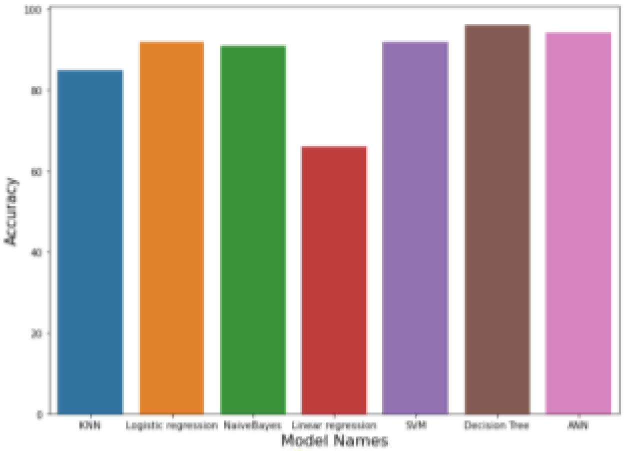
Model Name	Time Taken
48	.01 sec
BK	sec
Multi Layer Preceptron	.69 sec
Naïve Bayes	sec
Logistic Regression	.05 sec
Random Forest	.01 sec

Table 8 Epochs, Accuracy and Loss In Artificial Neural Networks(ANN)

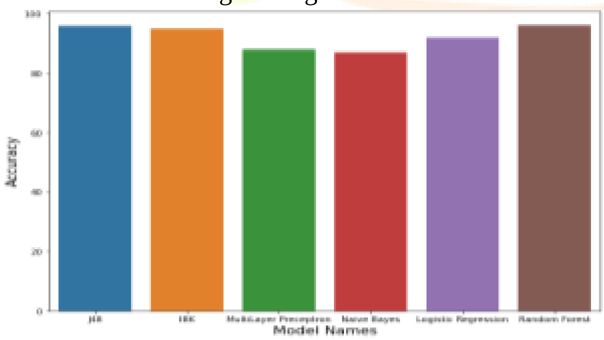
Epoch n.o	Accuracy	Loss

0	2.21%	9.89%
0	0.87%	8.61%
0	0.62%	5%
0	1.59%	2.63%
0	3.75%	2.53%
0	4.71%	8.23%
0	5.19%	8.59%
00	5.43%	5.64%

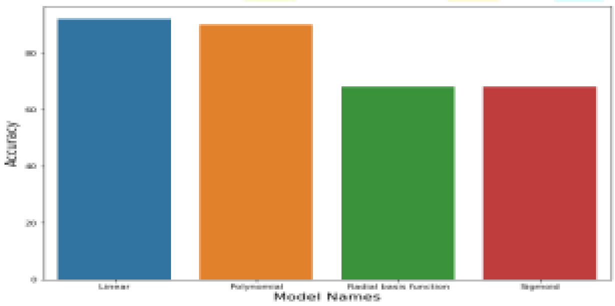
3.Analysis of Result:



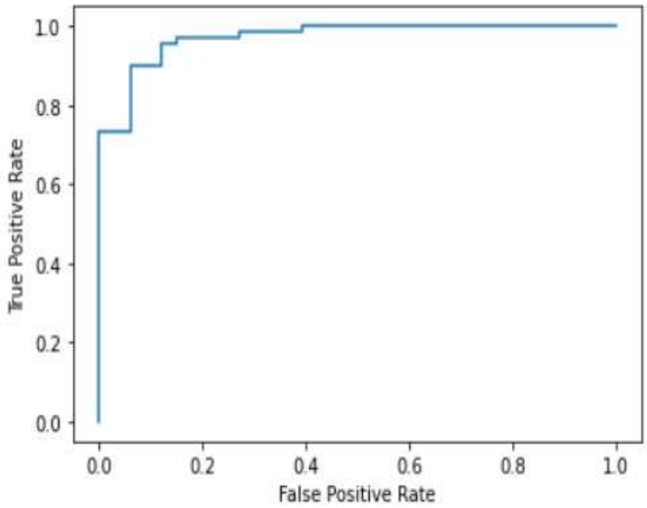
The above figure (a) Shows The Accuracy Comparision Between Model using Coding



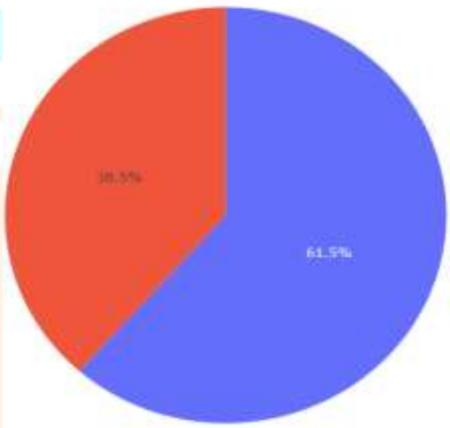
The above Figure (b) Shows Accuracy Comparision Between Models Using Weka Software



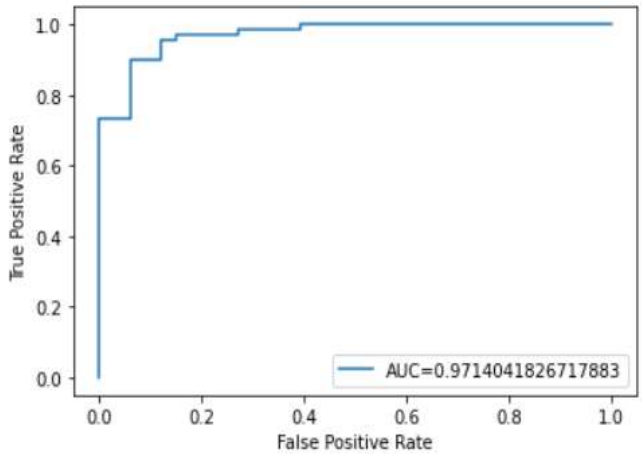
The above Figure (c) Shows Accuracy Comparision Between SVM Model Kernels



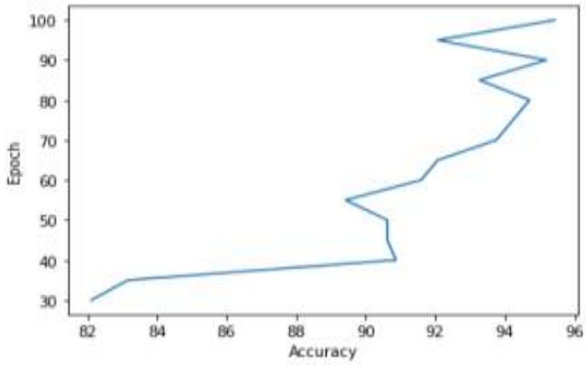
The above figure (d) is Shows Logistic Regression ROC Curve



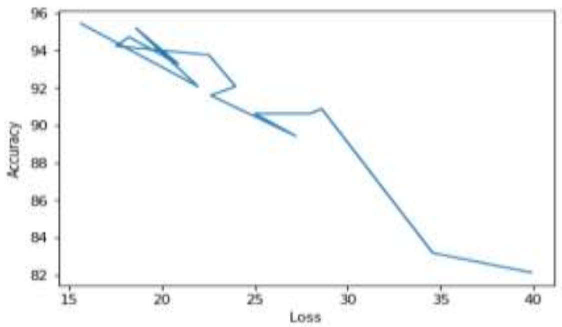
The above figure (e) is Shows Comparison Between Positive cases = 1 and Negative cases = 0.



The above figure (f) Shows Logistic Regression AUC Curve and Value of AUC is 0.97140



The above figure (g) Shows Epoch Vs Accuracy In Artificial Neural Networks



The above figure (h) Shows Loss Vs Accuracy In Artificial Neural Networks

Conclusion

The paper points out a Hybrid Supervised Machine Learning Classifier System for diabetes detection using symptoms. The performance of the classifiers has been tested on all attributes and selected features separately to obtain and compare the achieved accuracy .we had used popular machine learning classifiers such as Support Vector Machine, ANN, J48(using Weka Software), Multilayer-Perceptron ,Logistic Regression etc. Based on the experimental results, it is evident that Using Coding Decision Tree achieves an accuracy of 96%,a sensitivity of 94%, a specificity of 100%, and Kappa statistics of 91% And Artificial Neural Network(ANN) achieves an accuracy of 93.27%.Using Weka Software, J48 Algorithm yields an accuracy of 95.96% according to the testing data Additionally, the accuracy of the IBK is 95%.It is also observed that the Logistic Regression achieves an accuracy of 92%, a AUC of 0.97, a sensitivity of 95%, a specificity of 84% and Kappa statistics of 81%.Given the above, it is relevant that the Decision Tree , J48, IBK ,Artificial Neural Network Classifiers are the most appropriate machine learning-based classifiers for Diabetes Detection Using Symptoms.

References

[1] International Diabetes Federation - Home (idf.org)

[2] Diabetes (who.int).

[3] Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer

[4] Diabetes UCI Dataset | Kaggle

[5] Electronics | A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches (mdpi.com)

[6] P. Sonar and K. JayaMalini, ''Diabetes prediction using different machine learning approaches,'' in Proc. 3rd Int. Conf. Comput. Methodologies Commun. (ICCMC), Mar. 2019, pp. 367–371, doi: 10.1109/ICCMC.2019.8819841.

[7] Prediction of Diabetes Empowered With Fused Machine Learning.

[8] Diabetes Prediction Using Machine Learning KM Jyoti Rani

[9]Diabetes Prediction based on Supervised and Unsupervised Learning Techniques - A Review

[10] N. B. Padmavathi, ''Comparative study of kernel SVM and ANN classifiers for brain neoplasm classification,'' in Proc. Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICICT), Jul. 2017, pp. 469–473, doi: 10.1109/ICICICT1.2017.8342608

[11] .Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584- 1589). IEEE

[12] Artificial Neural Network Ivan Nunes da Silva ,Danilo Hernane Spatti , Rogerio Andrade Flauzino , Luisa Helena Bartocci Liboni ,Silas Franco dos Reis Alves

[13] Lakatos,G., Carson, E. R., and Benyo, Z., "Artificial neural network approach to diabetic management", Proceedings of the Annual International Conference of the IEEE, EMBS, pp.1010–1011, 1992

[14] Impact of Preprocessing for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks

[15] Diabetes Diagnosis using Artificial Neural Network Santosh Kumar*1, Dr.A. Kumaravel2

[16] Sensitivity and specificity - Wikipedia

[17] Precision and Recall in Machine Learning - Javatpoint

[18] https://en.wikipedia.org/wiki/F-score

[19] https://deepai.org/machine-learning-glossary-and-terms/epoch

[20] Matthews's correlation coefficient: Definition, Formula and advantages - Voxco

[21]A. D.-N. C. and Applications and undefined 2016, "Performance evaluation of different ma-chine learning techniques for prediction of heart disease," Springer

[22] Machine Learning and AI for Healthcare Big Data for Improved Health Outcomes

[23] Artificial Neural Network Ivan Nunes da Silva ,Danilo Hernane Spatti , Rogerio Andrade Flauzino , Luisa Helena Bartocci Liboni ,Silas Franco dos Reis Alves

[24] Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. Informatica 2007, 31, 249–268.+

[25]Cohen's Kappa Statistic: Definition & Example - Statology

[26]www.researchgate.net/publication/267965137_Prediction_of_Diabetes_by_using_Artificial_Neural_ Network

[27] Attenuated Total Reflection FTIR dataset for identification of type 2 diabetes using saliva

[28] Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd Edition by Wes McKinney

[29] Diabetes: Symptoms, Causes, Treatment, Prevention, and More (healthline.com).