IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 2, February 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002015

A Deep Learning Model for Prediction and Classification in the Diagnosis of Diabetes

Maria Navin J. R[1], Pankaja R[2], Sridevi N[3]

Sri Venkateshwara College of Engineering, Bangalore, Karnataka, India [1,2,3]

ABSTRACT: Diabetes mellitus is the most serious health challenges in many of countries. According to the International Diabetes Federation, there are 285 million diabetic people worldwide. This total is expected to rise to 380 million within 20 years. Diabetes-Mellitus refers to the metabolic disorder that happens from malfunction in insulin secretion and action. It is characterized by hyperglycemia where the persistent hyperglycemia of diabetes leads to damage, malfunction and failure of different organs such as kidneys, eyes, nerves, blood vessels and heart. In the past decades several techniques have been implemented for the detection of diabetes. The diagnosis of diabetes is very important now days using various types of techniques. Here, there are various techniques, their classification and implementation using various types of software tools and techniques. The diagnosis of diabetes can be done using Artificial Neural Network, K-fold cross validation and classification, Vector support machine, K-nearest neighbor method, Data Mining Algorithm, etc.

Recent advances in Deep Learning (DL) provide an analysis framework that can be used to automatically classify images and objects with human-level accuracy. A key advantage of Deep Learning is its ability to perform unsupervised feature extraction over massive data sets making big data part of the solution not the problem. Deep Learning is a technology which become a key tool at many of the top technology companies around the world. By using various deep learning techniques along with h2o package we are trying to predict the pima disease at early stage with atmost accuracy.

KEYWORDS: Diabetes, Deep Learning, Data Mining, Artificial Neural Networks, H2O Package, R Language, Confusion Matrix, Accuracy.

I. INTRODUCTION

Diabetes is one of the common and rapidly increasing diseases in the world. It is a major health problem in most of the countries. Condition of the diabetes is caused where your body is unable to produce the required amount of insulin needed to regulate the amount of sugar in the body [1]. This leads to various diseases including heart disease, kidney disease, blindness, nerve damage and blood vessels damage. There are two general reasons for diabetes: (1) the pancreas does not make enough insulin or the body does not produce enough insulin. Only 5-10 % of people with Type-1 diabetes have this form of the disease. In Type-2 diabetes cells do not respond to the insulin that is produced. Insulin is the hormone which regulates uptake of glucose from the blood into most. If the amount of insulin insufficient, then glucose will not have its usual effect so that glucose will not be absorbed by the body cells that require it. Detection and diagnosis of diabetes at an early stage is the need of the day.

Deep learning helps researchers analyze medical data to treat diseases. It helps in analyzing medical images which enhances doctor's ability. It's advancing the future of personalized medicine. The deep neural network has the ability to learn relevant information, thus outperforming results obtained from raw data. Deep learning methods emerge as a highly effective method because they have a structure that allows extracting attributes from data without preprocessing. Extracting attributes from data with classical methods is an extremely tiring process. However, Deep learning methods that do this automatically can produce more effective results. Deep learning techniques are trying to imitate the working mechanism of the human brain [2] [3].

II. RELATED WORK

The NB Tree model has the lowest accuracy of 70.60 percent when compared with the medical diagnosis that the error MAE is 0.3327 and RMSE is 0.454 [4]. In [5], Decision Tree, Naive Bayes and Support Vector Machine produced precision of 85%, 77% and 77.3% respectively. In [6], Pima Indians Diabetes database with 8 features, 500 non-

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002015

diabetic and 268 diabetic patients were considered. Using DL, SVM and RF an accuracy of 76.81%, 65.38% and 83.67% was obtained respectively.
In another Research paper presented by Yang Guo , GuohuaBai , Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, Thediscovery of knowledge from medical databases isimportant in order to make effective medical diagnosis. Pima Indian diabetes dataset was used. Preprocessing was done to improve the quality of data. Naïve Bayes Classifier was applied to the modified dataset to construct the model. Finally Weka tool was used to perform simulation and the accuracy obtained was 72.3%. [7]. Diabetes was closely related to BMI of a person and ANN classifier with principal component analysis produced an accuracy of 75.7% [8].
In [9], PCA and mRMR were used to reduce the dimensionality and an accuracy of 80.84% was achieved using random forest classifier.

III. METHODOLOGY

Deep learning is a branch of machine learning which is based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures [10]. The characteristics of deep learning are that it is based on ANNs where the machine learning techniques, primarily and supervised learning, are used to create new features from the input variables. Dataset is prepared by a study on Pima disease and we create a .csv file with 532 patient details where column is represented by disease attributes. The packages for performing deep learning are available in R language. In R, the machine learning is a function call, with parameters and training data being given at the same time [11].

The proposed system is implemented using the algorithm as shown.

Deep_Learning_H2O(Diabetes Data Set with 532 examples)
{
Data preparation with segregation of the data set into training and testing data sets.
Training the classification model using H2O deep learning with backpropagation error functions for output and hidden layers.
Prediction and Classification of the model on test data.
Performance analysis using Confusion Matrix.
}

1. Data preparation and uploading it to H2O

H2O is an open source predictive analytics platform with prebuilt algorithms, such as k nearest neighbor, gradient boosted machines, and deep learning. The dataset can be uploaded to the platform via Hadoop, AWS, Spark, SQL, no SQL, or your hard drive.
H2O package in R language provides the necessary methods for implementing deep learning concepts on any dataset. The prepared dataset on Pima disease have to be uploaded to H2O. To use the H2O machine-learning algorithms the data must be in the H2O cluster; all that exists on your client is a handle (a pointer) to H2O's data frame. The H2O function, h2o.uploadFile(), allows you to upload/import your file to the H2O cloud. The following functions are also available for uploads h2o.importFolder, h2o.importURL and h2o.importHDFS.

2. Create train and test datasets

We can upload our own train and test partitioned datasets to H2O. Now we can connect to H2O and use methods namely- runif(), assign() and table() to create the training and the testing dataset.

The characteristic feature variables of interest are as follows:

- npreg: This attribute indicates the number of pregnancies as a characteristic feature.
- glu: This attribute indicates the plasma glucose concentration in an oral glucose tolerance test
- bp: This attribute indicates the diastolic blood pressure (mm Hg)
- skin: This is triceps skin-fold thickness measured in mm
- bmi: This is the body mass index
- ped: This is the diabetes pedigree function
- age: This is the age in years
- type: This is a Boolean value whether diabetic or not, Yes or No

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002015

The datasets are contained in the R package, MASS. One data frame is named Pima.tr and the other is named Pima.te. Instead of using these as separate train and test sets, we will combine them and create our own in order to discover how to do such a task in R.

3. Modeling

H2O Deep Learning is a backpropagation and stochastic gradient based multi layer neural network classifier. The network created using H2O contains many hidden layer. Now we can create the predicted values using h2o.predict() by specifying the object created after calling h2o.deeplearning() function parameter. After performing the modelling we obtain the error rate and performance accuracy using a confusion matrix. The main aim of the data is to discriminatehealthy people from those with pima disease, according to "Confusion matrix" column which is set to N for healthy and Y for pima disease. The data is in ASCII CSV format. The rows of the CSV file consist of instances corresponding to one voice recording.

IV. RESULTS AND DISCUSSIONS

```
>str(pima.scale)

'data.frame':        532 obs. of  8 variables:
 $ npreg: num  0.448 1.052 0.448 -1.062 -1.062 ...
 $ glu  :num  -1.13 2.386 -1.42 1.418 -0.453 ...
 $ bp   : num  -0.285 -0.122 0.852 0.365 -0.935 ...
 $ skin :num  -0.112 0.363 1.123 1.313 -0.397 ...
 $ bmi  :num  -0.391 -1.132 0.423 2.181 -0.943 ...
 $ ped  :num  -0.403 -0.987 -1.007 -0.708 -1.074 ...
 $ age  :num  -0.708 2.173 0.315 -0.522 -0.801 ...
 $ type : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 1 1 1 2 ...

>h2o.table(train[,8])
type Count
1   No   247
2  Yes   117
[2 rows x 2 columns]

>h2o.table(test[,8])
type Count
1   No   108
2  Yes    60
[2 rows x 2 columns]

>dlmodel<- h2o.deeplearning(x=1:7, y=8, training_frame = train, validation_frame = test, seed = 123, variable_importances = TRUE)
```

The data set is segregated into training and test data set examples with 70% and 30% respectively. The parameter x represents the input variables of the data set, y is the target response variable, training_frame represents the training data set examples, validation_frame holds the test data examples, 123 represents an initial seed for the sampling. The final parameter namely variable_importance as true indicates the most important feature attribute that contributes during classification.

A confusion matrix is a N x N table that is often used to evaluate the performance of a classifier. The actual target values are compared with the predicted values generated by the machine learning model. The different values of the confusion matrix are True Positive, True Negative, False Positive (Type 1 Error) and False Negative (Type 2 Error). True Positive indicates the number instances where the actual predicted values are both positive. True Negative indicates the number instances where the actual predicted values are both negative. False Positive (Type 1 Error) indicates instances where the actual value was negative and predicted value was positive. False Negative (Type 2 Error) indicates instances where the actual value was positive and predicted value was negative.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002015

Confusion Matrix for train set (vertical: actual; across: predicted) for F1-optimal threshold:

Table 1

Confusion Matrix for Training Data Set	No	Yes	Error	Rate
No	238	9	0.036	=9/247
Yes	6	111	0.051	=6/117
Totals	244	120	0.041	=15/364

Confusion Matrix for test set (vertical: actual; across: predicted) for F1-optimal threshold:

Table 2

Confusion Matrix for Test Data Set	No	Yes	Error	Rate
No	105	3	0.027	=3/108
Yes	2	58	0.033	=2/60
Totals	107	61	0.029	=5/168

From Table 1, we observe that the deep learning model is trained using 70% of the instances from the data set. The Confusion matrix generated after the training procedure represents an accuracy of 95.8%. Similarly, from Table 2 during the testing procedure the accuracy of the prediction model is 97%. Hence, deep learning approach in machine learning play a vital role as classifier model for performing prediction with good accuracy.

V. CONCLUSION

In this paper, we cope the problem of Pima disease identification by means of deep neural network. The deep learning technique with h2o package is used to diagnose the pima disease. The deep learning method has ability to extract hidden features which considerably increases the performance of h2o. In the analysis of pima diabetes disease dataset we take pima disease dataset for analysis. We perform performance analysis on Deep Learning method using Confusion Matrix. We predict and classify the presence of the disease in the training data and test data with an accuracy of 95.8% and 97% respectively.

In regards to future works, we aim to extending the dataset with more data of individuals also our next idea is to learn a fusion schema that considers all attributes while making decisions about individuals.

REFERENCES

[1] Mujumdar A, Vaidehi V. (2019) Diabetes Prediction using Machine Learning Algorithms, ScienceDirect, Procedia Computer Science 165, 292–299.

[2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, (2017) Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal 15, 104–116.

[3] A.J. DinuGanesan R, Felix Joseph and Balaji. (2017) A study on Deep Machine Learning Algorithms for diagnosis of diseases, International Journal of Applied Engineering Research, 12, 6338-6346.

[4] S. LowanichchaiJabjone, and T. Puthasimma, Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree.

[5] P. Sonar and K. JayaMalini(2019). Diabetes Prediction using different Machine Learning Approaches", In2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)

[6] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques, In 2019 1st International Informatics and Software Engineering Conference (UBMYK), pp. 1-4.

[7] Y.Guo, G. Bai, and Y. Hu, Using Bayes Network for Prediction of Type-2 Diabetes, IEEE International Conference for Internet Technology and Secured Transactions.

[8] T.M. Alam, M.A. Iqbal, Y. Ali, A. Wahab, S.Ijaz, T. I.Baig, A. Hussain, M.A. Malik, M.M. Raza, S. Ibrar, and Z.Abbas (2019) A model for early prediction of diabetes, ScienceDirect, Informatics in Medicine Unlocked, 16, 100204.

[9] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, Predicting Diabetes Mellitus with Machine Learning Techniques, Front. Genet, https://doi.org/10.3389/fgene.2018.00515.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 2, February 2021 ||

DOI:10.15680/IJIRSET.2021.1002015

[10] L. Deng, G. Hinton, and B. Kingsbury(2013) New types of deep neural network learning for speech recognition and related applications: An overview. In ICASSP, 8599–8603.

[11] G. E. Hinton, S.Osindero, and Y. W. Teh, (2006).A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527 1554.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details