Diabetes Predictor System using Machine Learning Algorithms over Live Cloud Infrastructure

Deeksha K [1], Amrutha M [2], Harshitha [3], Ashwini Gotyal [4] Dr.M. Kusuma [5]
[1][2][3][4] BE Students, Department of Information Science and Engineering
[5] Professor, Department of Information Science and Engineering
[1][2][3][4] Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India.

Abstract:- Diabetes the silent killer which kills part by part of human life. Diabetes can knock anyone at any point of time. It is not only an infirmity but also an establisher of different kinds of diseases. Early prediction in such diseases plays a vital role in the developing economy. With the rising technologies such as deep learning within machine learning will help in finding a remedy to this outgrowth. Machine learning techniques increase medical diagnosis accuracy and reduce medical costs.The classification model in the machine learning approach helps medical specialists to improve their forecasting, diagnosis or treatment planning procedures. Whereas Deep learning within machine learning requires relatively keen knowledge about the data being used, the algorithms to be applied and analysis being made.

Big data analytics provides an excellent basis for handling the vast quantity of data. Methods of Data analysis and large scale parallel processing have been exploited in the Healthcare sector over the cloud services, thus making immense progress in developing solutions to health-related problems.

***Keywords*:-** *Diabetes, Deep Learning, Accuracy, Machine learning algorithms.*

I. INTRODUCTION

Diabetes Mellitus has become one of the greatest threat to the rising population and as it is a chronic disease there exists no cure to the problem[1].It is found that in the last few decades the cases of people living with diabetes jumped to almost 350 million worldwide.It is caused due to malfunctioning of the pancreases,destroyed cells present in the pancreas that makes insulin within the immune system and also due to obesity and lack of exercise[1]. Pancreases being one of the most consequential organs in our body releases a hormone called insulin [2]. When this organ fails to deliver enough insulin it leads to hyperglycemia(high blood sugar)or hypoglycemia(low blood sugar)[1].

Through the studies, it has been found that glycemic control is the imperative criterion to prevent organ damage and other complications.

There are two essential divisions in diabetes namely Type1 and Type2 diabetes. Both types of diabetes are chronic diseases that affect the way the body conducts blood sugar or glucose. Glucose acts as the fuel that feeds the body cells, and to enter the cells it needs a key,Insulin acts as that key. People with type 1 diabetes don’t produce insulin. People with type 2 diabetes don’t respond to insulin as well as they should and in the later stages do not make enough insulin[3].

Some of the major symptoms of diabetes patients are increasing in thirst, urination and blurred vision, hunger fatigue, coldness in the feet or hands, cuts that do not amend early and inexplicable weight loss [2].

As the cases of diabetes are increasing tremendously the need to analyze the cause and remedy becomes indispensable, for which a new field of science called Data Science is being exploited worldwide. It is a field that applies scientific methods along with algorithms for training and processing data which can be either structured or unstructured in essence.The structured is the formatted data and unstructured is the unformatted data. Data science is a concept that combines statistics, data analysis, machine learning, and their correlated approaches in order to know the spectacles held within the data [2].

Now moving towards what is Machine Learning,It is a tool to convert information into knowledge. It’s an element of data science that provides the system the aptitude to study and progress from practice without being explicitly programmed.It has many forms such as supervised, unsupervised, semi-supervised and reinforcement learning. Supervised requires a data scientist to provide both input and desired output, which requires them to arbitrate what should be the input variables and output variables. Whereas Unsupervised need not be trained with the expected output as they undergo extensive training using which they learn by themselves and predict the output[2].

Deep learning is one portion of a broader family of machine learning and requires relatively keen knowledge about the data being used, the algorithms to be applied and analysis being made.

It has networks of learning unsupervised from data that is unstructured or unlabeled.

As deep learning works extensively with large amounts of data, requires exhaustive computing making scalability of machine learning models inoperable.Tensorflow can be used to overcome this challenge as it provides an architecture that builds processing on both CPU and GPU.thus developed models are cloud computing equipped[7].

Cloud Infrastructure, The present trend in the internet ecosphere is to link all the devices to the internet with a mission of enhancing the quality of everyday lives. Cloud computing is an incredible platform to handle the huge data which is produced from the IoT atmosphere, it is also used to handle the real-time processing of information, called big data. Cloud processing networks act as a monitor or guidance platform to guide edge to edge processing of available resources. More importantly, in this cloud framework, cloud processing can be utilized network-wide globally. By doing this essential utilization of available resources can be achieved[4].

II. LITERATURE SURVEY

[5] presents a comparative study conducted on the PIDD dataset. The technique gave encouraging conclusions for both the datasets. Outlier detection was employed as a pre-processing step to detect and remove outliers in diabetes datasets.

The proposed framework used an ensemble of four MLP's to achieve greater accuracy.

The AutoMLP gave higher accuracy by achieving an accuracy of 88.7%.

In [9], the authors used an LSTM network with one LSTM layer, one bi-directional LSTM layer along with several fully connected layers were used to predict blood glucose concentration.

The LSTM network was pre-trained with both silico data and real patient data to generate a “global model”.

The authors in [11] flourished a disease prediction model (DPM) for type 2 diabetes and hypertension by integrating iForest, SMOTETomek, and ensemble learning.

The forest was used to detect and waive the outlier data from the dataset while SMOTETomek was utilized to balance the imbalanced dataset. The proposed model enhanced the accuracy to 96.4% as it is an ensemble model it performs better than the individual.

III. METHODOLOGY

A. Data collection and processing:

There are various sources available for collecting required data around diabetes patients. Some of the many are UCI Machine learning Repository, Kaggle and PREST database.

The dataset from the UCI Repository was acquired from two sources namely, an automatic electronic recording device and paper records[8]. Kaggle has the dataset named by PIMA Indians Diabetes Database[9]**.**

Whereas PREST is a database of ratification embarked in 2010,it combines information from different sources to obtain clinical variables[7].

The PIDD dataset from UCI has 768 samples, it contains 268 diabetic and 500 non-diabetic samples.It consists of 8 numeric valued features and a class label containing 2 values,1 and 0.

No.	variable	New Variable	Value
1	Pregnancies	comp1	Integer
2	Glucose	comp2	Integer
3	Blood pressure	comp3	Integer
4	Skin thickness	comp4	Integer
5	Insulin	comp5	Integer
6	BMI	comp6	Numeric
7	Diabetes function	comp7	Numeric
8	Age	comp8	Integer
9	Class	class	Integer

Table 1:- The attributes of PIDD

Where 0 refers to negative to diabetes and 1 refers to positive to diabetes[7].

The data is not balanced meaning there exists more negative cases than positive cases. Hence we focus more on data pre-processing so that when this data is passed to the prediction model it produces accurate results.

Naive bayes is one of a classification model that is considered for training the samplings and resolves for the missing values in the dataset.In [3],Quian Wang suggests ADASYN method,an adaptive method that can synthesize the samples through K-nearest neighbours and reduce class imbalance of a dataset.

Extracted features of diabetes data are projected to a new space using principal component analysis, then it is modeled by applying classification methods on these newly formed attributes. the principal component analysis must be applied to comprehend the features that contribute the most in finding accuracy [6]. Principal component analysis (PCA) is an important technique to understand in the fields of data science. An accuracy of 82.1% was obtained by using this method in paper [6] for predicting diabetes which has reformed over existing classification methods.

B. *Analysis of data:*

Before feeding the data to prediction model , it must be first split into training and testing sets, i.e. a machine needs to be "trained" by explicitly feeding it data that has the correct answers affixed.This training data will help the machine to connect the articulation in the data to the right answer[1].

Once trained in this way, a machine can now be given test data that has no answers for it to analyse and produce results on its own, i.e. without human intervention.

All classical methods are trained following a 80-20% training and testing articulation fused with a K-fold cross validation to warble parameters[2].

The performance analysis can be done by splitting processing into two sets where one set makes use of a supervised algorithm and another makes use of the deep learning algorithm and parallelly processes both the algorithms.

The ability to run models in parallel results in a significant boost in terms of performance compared to a sequential approach.

The two algorithms being discussed here for solving the problem are decision tree and RNN algorithm.

Decision tree - It is a supervised algorithm, which uses a tree-like structure model in which each internal node is depicted as “test” on a feature, each branch represents the outgrowth of the test, and each leaf node represents a class label. The paths from root to leaf are characterized as classification regulations.[2].

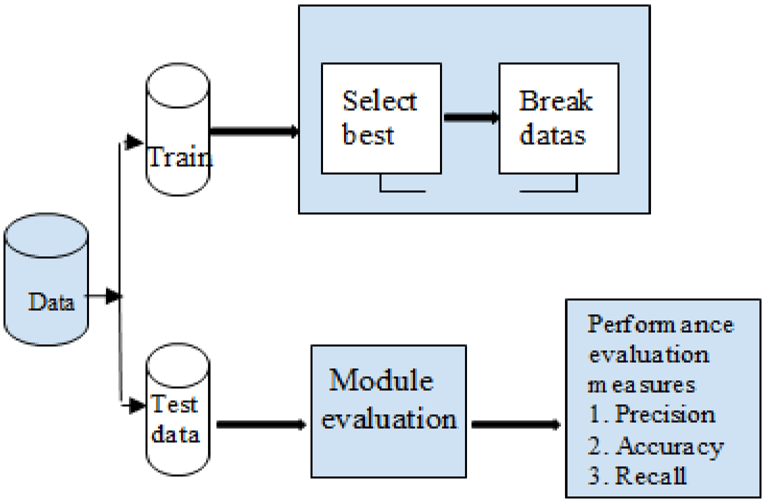


Fig 1:- Decision Tree architecture schema

In particular, due to the consecutive nature of medical information and the underlying dependence from the patient’s condition related to the involved tests recited, implementation of a Recurrent Neural Network (RNN) for the patient’s classification is most operative[7].

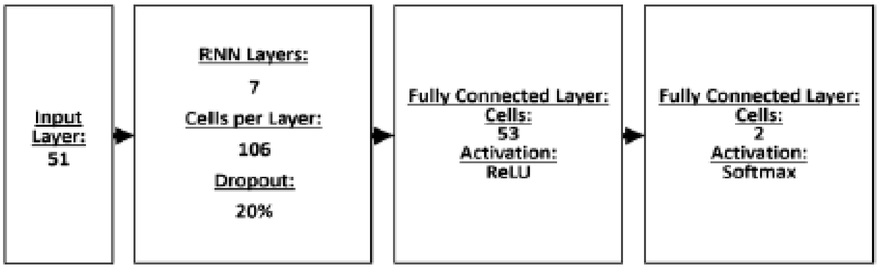

Input Layer: 51
RNN Layers: 7
Cells per Layer: 106
Dropout: 20%
Fully Connected Layer: Cells: 53 Activation: ReLU
Fully Connected Layer: Cells: 2 Activation: Softmax

Fig 2:- RNN architecture schema

The Recurrent Neural Network class of artificial neural network has proven effective results in terms of accuracy was considered in [7].The RNN model of this paper hosted 7 layers with 106 cells and was trained using 5000 epochs using a passel of 250 records per class and implemented Adam Optimizer for optimization function[7].

The testing was done by randomly selecting 280 records per class and hence were isolated completely from the training stage.

Data Parallelism was applied while working with a huge amount of available training data where two approaches were used to train the data,one is by cropping data and other is by using a bootstrap technique[7].

The motive of conducting this whole program of Data Parallelism was to make sure time does not become the reason for the decrease in efficiency.

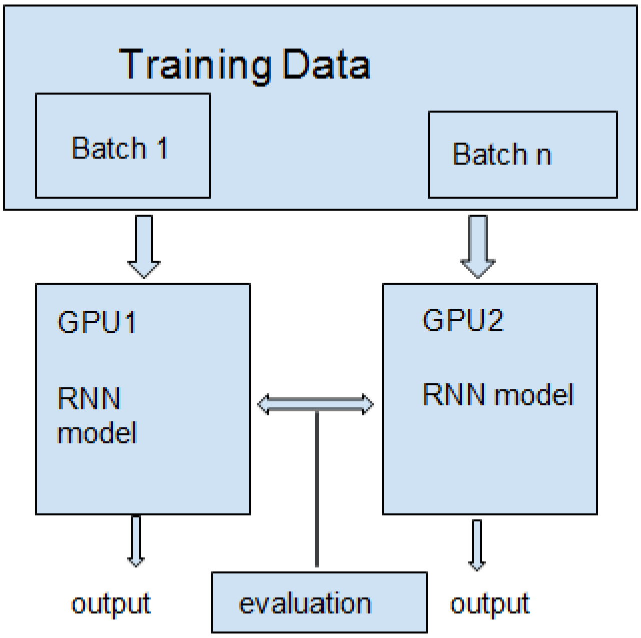


Fig 3:- Data parallelism schema

C. *Metrics*

- *Evaluation Criteria:*

Root mean square error (RMSE, [mg/dl]), correlation coefficient (CC), time lag (TL, [min]) and fit must be used to estimate the prediction performance. The evaluation criteria indicate the overall predictive ability of the models by comparing the test datasets and the predicted value [9].

RMSE indicates the difference between the target data and the predicted data.

$$\boldsymbol{RMSE} = \sqrt{E\left(\left(G - \hat{G}\right)^{2}\right)} = \sqrt{\frac{1}{N}\sum \left(G - \hat{G}\right)^{2}}$$

Where G and G’ are actual and predicted glucose values.

Fit is calculated on the basis of the fraction of RMSE and root mean square difference between target and its mean value.

$$Fit = \left(1 - \frac{\sqrt{\frac{1}{N}\sum(G-\hat{G})^2}}{\sqrt{\frac{1}{N}\sum(G-G_{mean})^2}}\right) * 100\%$$

➢ *Performance Metrics:*

The relation between the actual output and the corresponding estimated value using the clustering technique based on the Model predictions can have four different potential outcomes: true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). TP and TN outcomes are correctly classified, whereas FP outcomes classified as positive when they are actually negative, and FN outcomes classified as negative when they are actually positive [8].

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

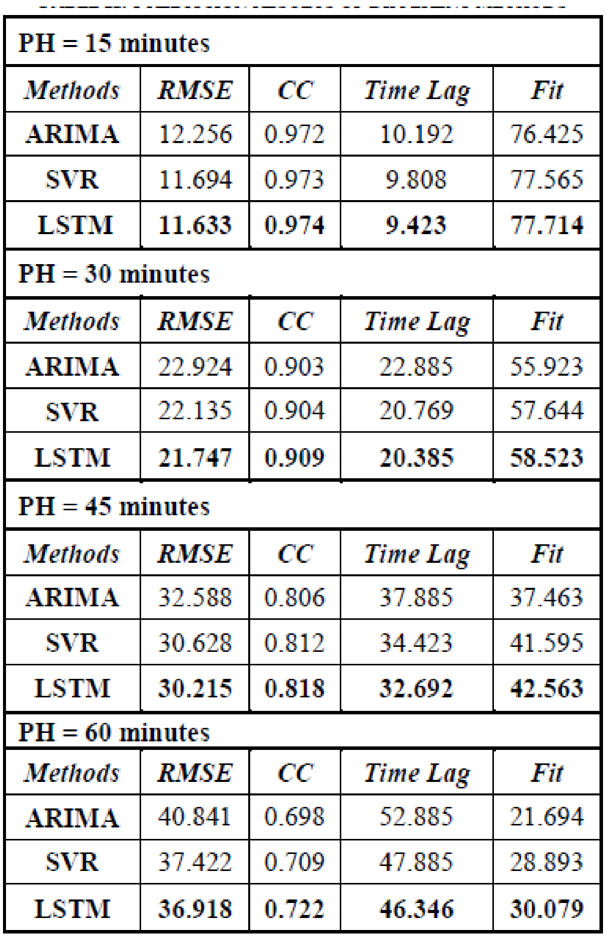
$$Precision = \frac{TP}{TP+FP}$$

The accuracy measures the proximity between the predicted values and works well only for balanced and the real valued data whereas precision denotes the proportion of the true positives that identify relevant instances ie it points out how much of the model saying right is actually right[8].

IV. RESULT COMPARISON

The accuracy measures the contiguity between the predicted values and the real values, the precision denotes the proportion of the true positive cases among the positive cases. Prediction juxtaposition of different models and hence drawing the conclusion from different models will help in grasping the pros and cons of each model.

The comparison is done for different models as well as for different features.



PH = 15 minutes				
Methods	*RMSE*	*CC*	*Time Lag*	*Fit*
ARIMA	12.256	0.972	10.192	76.425
SVR	11.694	0.973	9.808	77.565
LSTM	**11.633**	**0.974**	**9.423**	**77.714**
PH = 30 minutes				
Methods	*RMSE*	*CC*	*Time Lag*	*Fit*
ARIMA	22.924	0.903	22.885	55.923
SVR	22.135	0.904	20.769	57.644
LSTM	**21.747**	**0.909**	**20.385**	**58.523**
PH = 45 minutes				
Methods	*RMSE*	*CC*	*Time Lag*	*Fit*
ARIMA	32.588	0.806	37.885	37.463
SVR	30.628	0.812	34.423	41.595
LSTM	**30.215**	**0.818**	**32.692**	**42.563**
PH = 60 minutes				
Methods	*RMSE*	*CC*	*Time Lag*	*Fit*
ARIMA	40.841	0.698	52.885	21.694
SVR	37.422	0.709	47.885	28.893
LSTM	**36.918**	**0.722**	**46.346**	**30.079**

Fig 4:- Result comparison of different models

V. CONCLUSION

All ML Classification algorithms are beneficial for predicting measures of diabetes. The prediction analysis is the technique in which the model predicts the future on the basis of current situations.The model should be able to handle any amount of data , any type of data and correspondingly produce accurate results,for which the model must be trained extensively.The training of the data results in better accuracy as we take care of class distribution, class imbalance and missing values, hence results in inaccurate prediction.It is also observed that learning with more sample data set can improve the accuracy with reducing error rate.

The most appropriate attribute for diabetes prediction in PIDD was surveyed to be: plasma glucose concentration, diastolic blood pressure and number of times pregnant.Using these details it becomes easy for us to realise the dependency between feature and output hence move in that direction to obtain the best accuracy. From studying different papers we can conclude that LDA and SVM in supervised models perform the best. Whereas the RNN in an unsupervised model is considered the best. The performance of the prediction is related to the way the model is constructed, for this reason, we ought to obtain better results by applying different different techniques all together rather than individual models.

REFERENCES

[1]. Meng, G., &amp; Saddeh, H. (2019). Performance Analysis of Different Classifiers for Diabetes Diagnosis International Journal of Computer Application & amp: Information Technology, 11(2), 265-270.

[2]. Thomas, J., Joseph, A., Johnson, I., & Thomas, J. (2019). Machine Learning Approach For Diabetes Prediction. *International Journal of Information*, *8*(2).

[3]. Wang, Q., Cao, W., Guo, J., Ren, J., Cheng, Y., & Davis, D. N. (2019). DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. IEEE Access, 7, 102232-102238.

[4]. Sharma, S. K., &amp; Wang, X. (2017). Live data analytics with collaborative edge and cloud processing in wireless IoT networks. IEEE Access, 5, 4621-4635

[5]. Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., &amp; Nawaz, R. (2017, September). An expert system for diabetes prediction using autotuned multi-layer perceptron. In 2017 Intelligent Systems Conference (IntelliSys) (pp. 722-728). IEEE.

[6]. Chen, M., Yang, J., Zhou, J., Hao, Y., Zhang, J., & Youn, C. H. (2018). 5G-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds. *IEEE Communications Magazine*, *56*(4), 16-23.

[7]. Sierra-Sosa, D., Garcia-Zapirain, B., Castillo, C., Oleagordia, I., Nuno-Solinis, R., Urtaran-Laresgoiti, M., & Elmaghraby, A. (2019). Scalable Healthcare Assessment for Diabetic Patients using Deep Learning on Multiple GPUs. *IEEE Transactions on Industrial Informatics*

[8]. Ateeq, K., & Ganapathy, D.G. (2017). The novel hybrid Modified Particle Swarm Optimization – Neural Network (MPSO-NN) Algorithm for classifying Diabetes.

[9]. Sun, Q., Jankovic, M. V., Bally, L., & Mougiakakou, S. G. (2018, November). Predicting Blood Glucose with an LSTM and Bi-LSTM Based Deep Neural Network. In *2018 14th Symposium on Neural Networks and Applications (NEUREL)* (pp. 1-5). IEEE.

[10]. H. Roopa and T. Asha, "A Linear Model Based on Principal Component Analysis for Disease Prediction," in IEEE Access, vol. 7, pp. 105314-105318, 2019.

[11]. Fitriyani, N. L., Syafrudin, M., Alfian, G., &amp; Rhee, J. (2019). Development of Disease Prediction Diabetes and Hypertension.IEEE Access, 7, 144777-144789.