DIABETIC DISEASE PREDICATION USING MACHINE LEARNING TO ACCOMPLISH PRECISE ACCURACY

Krishna kumari Matta[1] V.N.S.S.R.K.Sai Somayajulu Meduri[2]

Indra Devi Bellapukonda[3]

[1]Department of Computer Science & Engineering, VFSTR(Deemed to be university), Guntur, India mkk_cse@vignan.ac.in

[2]Department of Information Technology,Gudlavalleru Engineering College Gudlavalleru, India m.somayajulu12@gmail.com

[3]Department of Computer Science & Engineering, Sri Vasavi Institute of Engineering and technology, Nandamuru, India indiradevi329@gmail.com

Abstract

Now a day's the working environments has turned out to be stressful due to numerous reasons. These things make a drastic raise to various chronically diseases irrespective of age, gender and other relevant conditions. No one is aware or can give assurance that they will be safe and healthy for the next moment. One such kind of disease we have considered to discuss as a case study is diabetic due to which the death rates are increasing drastically throughout the globe. To be safe and to carry out a good life it is advised to trace out the probability of coming across through such kind of diseases. With drastic new cutting edge technologies surrounding us it is possible to predict the possibility or probability of occurrence of such kind of diseases with certain parametric symptoms which are associated with these diseases. Among many such categories of diseases we have chosen the prediction of diabetes at an early stage based on the symptoms. In this paper we have used ML techniques for our early prediction with accurate accuracy.

Keywords: *Decision Tree, Gradient Boosting, Gaussian Naive Bayes, Linear Regression, K-Nearest Neighbours, Machine Learning, Random Forest, Support Vector Clustering.*

1. Introduction

ML is an evolved field of computer science from artificial intelligence computational theory and pattern recognition study. Here numerous algorithm can be learned and built from a given data sets which is helpful for some kind of predictive work. In diverse situations during making decisions in activities for daily life both abstract notion and physical objects can be dealt by human beings. For example while discriminating a fruit and flower with the help of a machine cannot be carried out with the help of physical objects. While storing the abstracts of these objects with respect to the machine we can carry out the task [11]. For example with the help of various distinct features which are associated with a fruit and flower such as colour, size, shape, weight, cost play an important role for separation [10]. As these objects are represented with their features i.e. patterns so we use patterns in machine learning for predicting.

II. Classification of Machine Learning

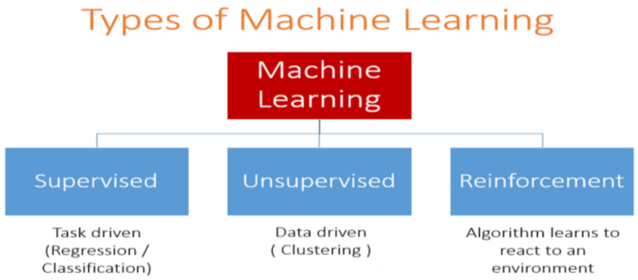

Types of Machine Learning
Machine Learning
Supervised
Unsupervised
Reinforcement
Task driven (Regression / Classification)
Data driven (Clustering)
Algorithm learns to react to an environment

Fig 1: Showing the basic classification in machine learning

Classification of Machine Learning			
Criteria	**Supervised**	**Unsupervised**	**Reinforcement**
Definition	Machine learns new things under supervision.	Machine learns new things without any supervision.	For a given environment software responds in order to take necessary action during which either rewards or errors may come across.
Dealing Problems	Regression, Classification	Association Clustering	Award/Reward
Kind of data	Labelled	unlabeled	Data not predefined
Training environment	Under supervision	No Supervision	Without Supervision
Type	Regression, classification	Clustering	Positive, Negative
Example	Detecting Fraud Card Transaction , Tumours,	Detecting abnormal accessing website, identifying topics in blog	Chess Game

Table 1: Showing basic classification of a machine learning

III. Machine Learning Classifier

In order to predict the probability of a particular event which can be carried out by training and testing data set where classification are of the following type

- ***K-NN Classifier***: If falls under supervised learning, for both regression and classification it is used. It is a kind of method with non-parametric nature [12]. To have class membership output we prefer to use classification, for a given object to have a property value we prefer to use regression. Here we use a Euclidian metrics which is show in the below equations.

$$d(x, x') = \sqrt{\left(x_1 - x_1'\right)^2 + \ldots + \left(x_n - x_n'\right)^2}$$

$$P(y = j | X = x) = \frac{1}{K} \sum_{i \in \mathcal{A}} I(y^{(i)} = j)$$

- ***Support Vector Clusters (SVC):*** it also falls in supervised learning when the considered data set is small and precise then this kind of classifier is used. They can be further classified into linear and non linear. SVC is defined as

$$\left[\frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i(w \cdot x_i - b)\right)\right] + \lambda \|w\|^2.$$

Where $\vec{w} \cdot \vec{x} - b$ defines the hyper plane. Risk which is observed in SVC is represented as

$$\varepsilon(f) = \mathbb{E}\left[\ell(y_{n+1}, f(X_{n+1}))\right]$$

it can be reduced

$$\hat{\varepsilon}(f) = \frac{1}{n} \sum_{k=1}^{n} \ell(y_k, f(X_k)).$$

- ***Linear Regression (LR)*** : It specifies relation between explanatory variables and their scalar response. This method is used for predictive model fitting. Matrix representation is given as y=xβ+ε

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

- ***Decision Tree (DT):*** By following an iterative manner weak learners are combined together to form a strong learner existing as a single entity [8]. Here the mean square error is calculated as

$$\frac{1}{n}\sum_{i}(\hat{y}_i - y_i)^2,$$

Where "n" number of samples in y

$\hat{y}_i$ is the predicted value of F(x)

y_i is real value

- ***Random Forest (RF):*** Here a number of decision trees are constructed; it is also used for regression and classification. It works on the basic principle that weak learners combine together to form strong learners, another advantage of this rule is that it does not over fit [9]. Here regression and standard deviations are given as

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \qquad \sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

- ***Gaussian NB (GNB):*** They are applied based on Baye's theorem. They exhibit high scalability, having linear parameters for a given learning problem. For a given vector **x** = (x1,x2,x3,....xn) which represents n independent variables whose probabilities are given as $p(C_k \mid x_1, \ldots, x_n)$

On applying bayes theorem we get

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \qquad \text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

- ***Gradient Boosting (GB):*** it's a kind of prediction model used for prediction models which are not strong; they use the technique of both classification and regression. For a given model F having M stages can be expressed using gradient boosting as

$$F_{m+1}(x) = F_m(x) + h_m(x) = y$$

IV.Experimental Results

Deficiency of insulin content in blood is considered as the major cause of diabetes beside which there are many factors which are affecting the same. Increased hunger, thirsty feeling and increased frequency of urination are the warning bell for high sugar content in blood this can be referred as a prior diabetes. Which needs to be addressed at a proper time if not then it may lead to death also. Malfunction of exocrine gland for not producing sufficient content of insulin is the cause. Though there exists numerous decision support systems which helps in assisting health specialist but the major one is given importance with high accuracy. Diabetes may be classified into three types Type 1 includes the case where hypoglycaemic agent fails to produce in enough content by pancreas, it can also be referred as juvenile diabetes. Type 2 it arises due to the fail in response by the cells to hypoglycaemic agent in an effective manner, by the coming years this type of diabetes will increase in large. Type 3 is related to gestational which mainly occurs in women during pregnancy, during pregnancy for elderly women the risk with this kind of diabetes is very high.

The data which is used from the data set [1] is first normalized and then standardized with the following requirements. The data is normalized using

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

In order to get better solution of coefficients, features are scaled using normalization. After normalization of data then it is standardized using the equation

$$z = \frac{x_i - \mu}{\sigma}$$

Where z is the data which is rescaled for standardized with **σ** = 1 and **μ** = 0

Among many parametric symptoms which are helpful for tracing the diabetes are Age, Body Mass Index, Blood Pressure, Insulin Percentage in body, Diabetes Pedigree Function, Skin Thickness and Pregnancies are important among them.

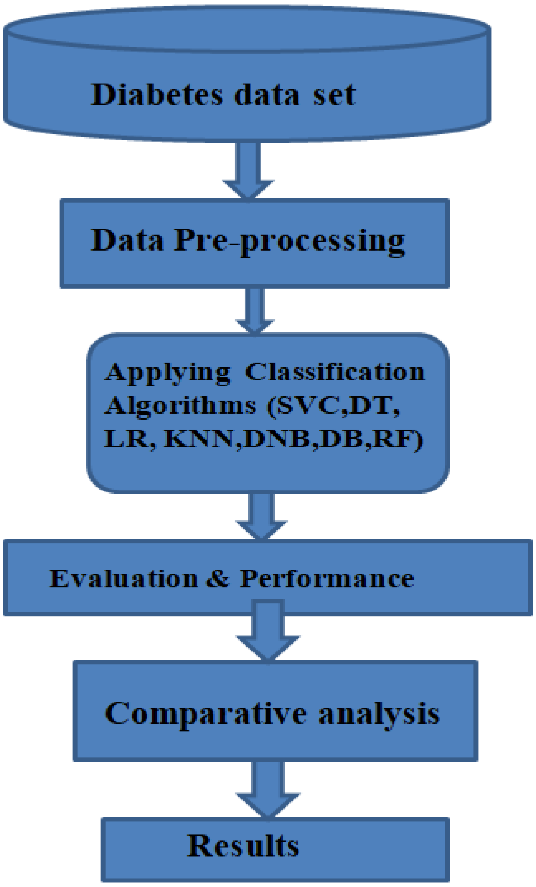
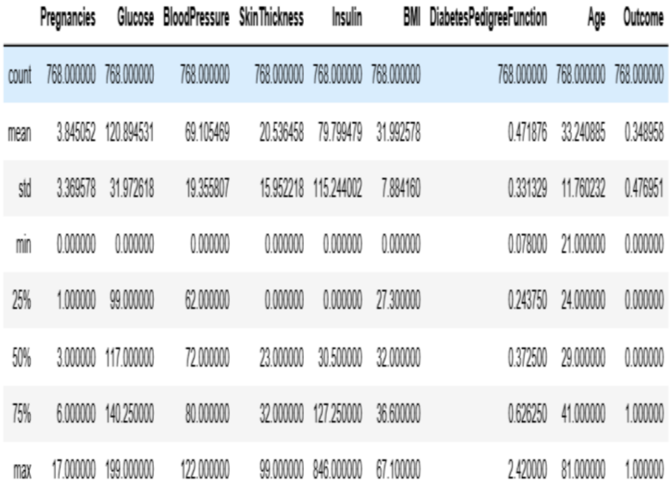


Fig. 2: Proposed Architectural Model

The above architecture shows the way in which the experiment is being carried out here first we have considered a relevant dataset for disease prediction, then the collected data is normalized and then standardized, then various classification methods are being applied in order to know the exact accuracy and prediction given by them, then the obtained results are analyzed and final result is discussed.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Fig.3: A primary view of the dataset used for prediction

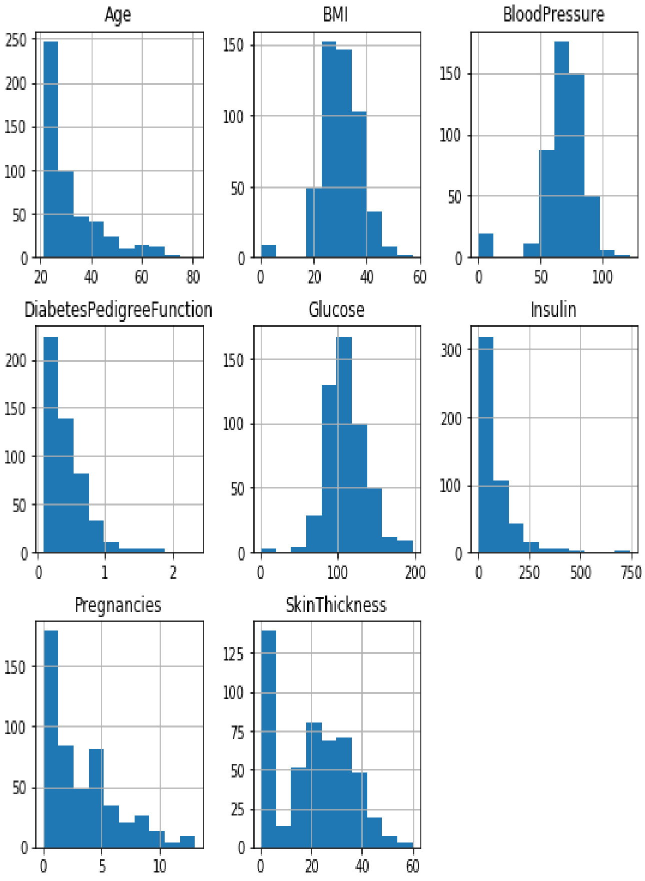


Fig.4: Showing the scaling factors affecting diabetes

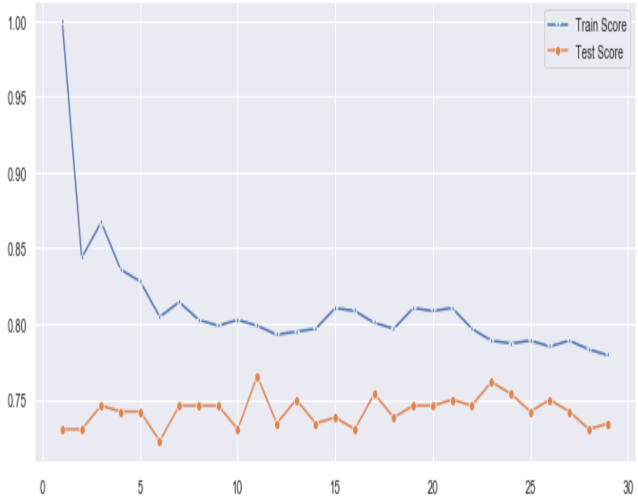


Fig.5: Results obtained from KNN Method with a nearest

The above fig shows the nearest value of k=11 for which the testing and training data give the correct results.[2][3][4] showed the prediction of diabetes using various methods. After carrying out necessary operation on the considered dataset with respect to linear regression and KNN visualization of data is clearly observed in the above diagram which clearly shows the factors which are much effective causing the high rate of diabetes.

Classifier	KNN	SVC	LR	DT	GNB	RF	GB
Accuracy	0.7365	0.6610	0.77051	0.70019	0.75525	0.73853	0.7702
Precision	0.7831	0.7027	0.8192	0.7444	0.8029	0.7852	0.8188
Recall	0.6253	0.5518	0.6432	0.5845	0.6305	0.6166	0.6430
F1 Score	0.8302	0.7450	0.8684	0.7891	0.8512	0.8324	0.8681

Table 2: showing the obtained results for various classifiers

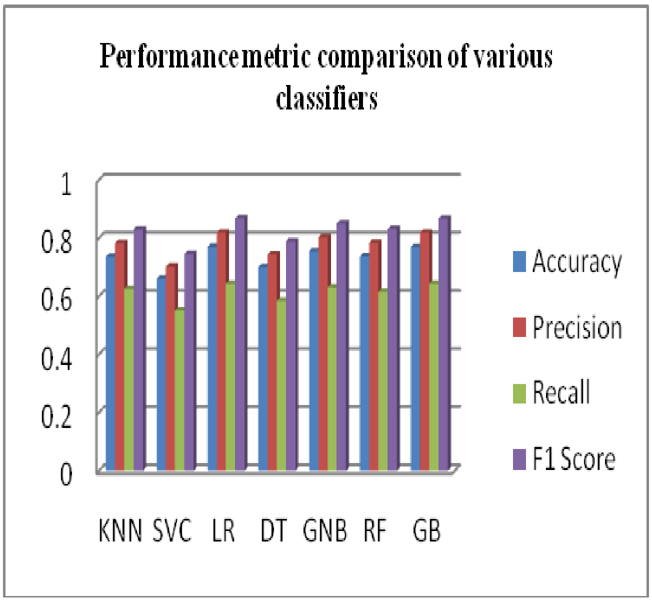


Fig.6: Showing obtained values of Accuracy, Precision, Recall and F1 Score of various classifiers

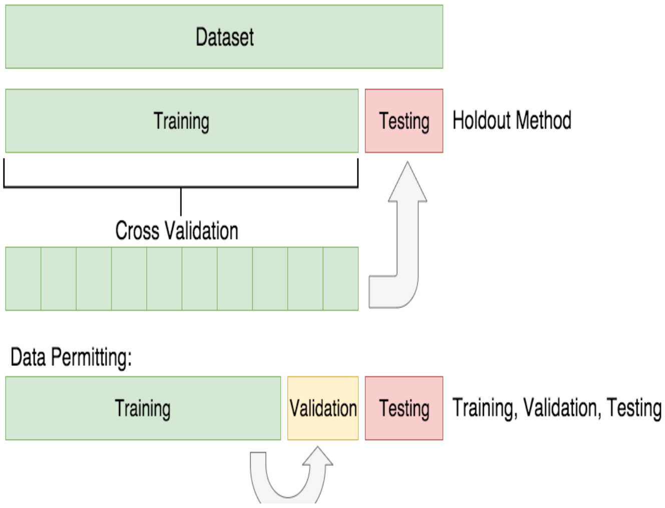


Fig.7: Showing how training and testing of a given data is carried out in the machine model

The above fig. Shows that how the collected or considered data is used effectively for training and testing the effect of diabetes with the considered parameter. The same is used for analyzing whether our methods are giving an apt results or not [5][6][7] shown effectively the usage of ML techniques used for effective predictions.

V. Performance Analysis

For a proposed model there are methods to trace out its performance which include the following

- ***Confusion Matrix:*** Used for performance summarization of an proposed classification algorithm with the outputs in certain binary formats.
- ***ROC (Receiver Operating Characteristic) :*** Used for describing how well a model is successfully distinguishing the required things, if it does it is labelled as better model if it not it is referred as a poor model.

Precision, Recall and F1-Score play an important role. Precision is defined as number of TP are the cases where model has defined them as false, FP are the case where model have defined as positive. Recall is defined as the ratio between TP and FN. F1 Score is needed when there is a balance between Precession and Recall. Where all the three are defined as

Precision =TP / (TP+FP).

Recall = TP / (TP+FN).

F1 score = (Recall * Precision)/ (Recall + Precision)

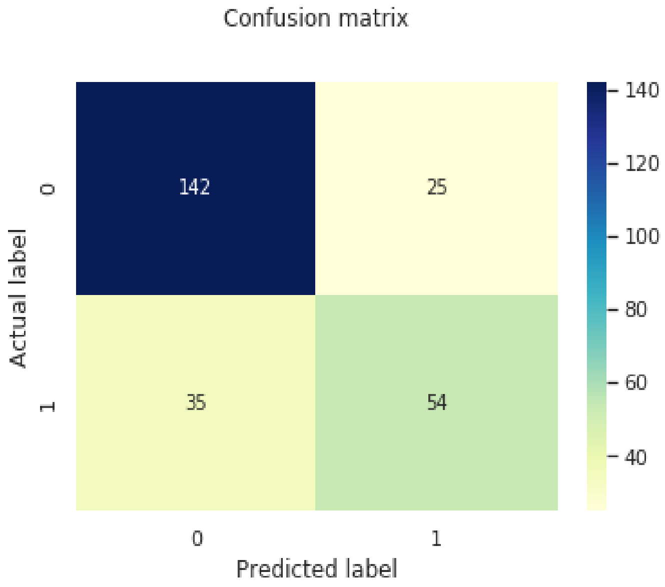


Fig.8: A sample confusion matrix generated for KNN classifier

The Decision tree which is generated for a given dataset with the help of which we make an early prediction of diabetes. Importance of tree or graph in detection is well explained by [8].

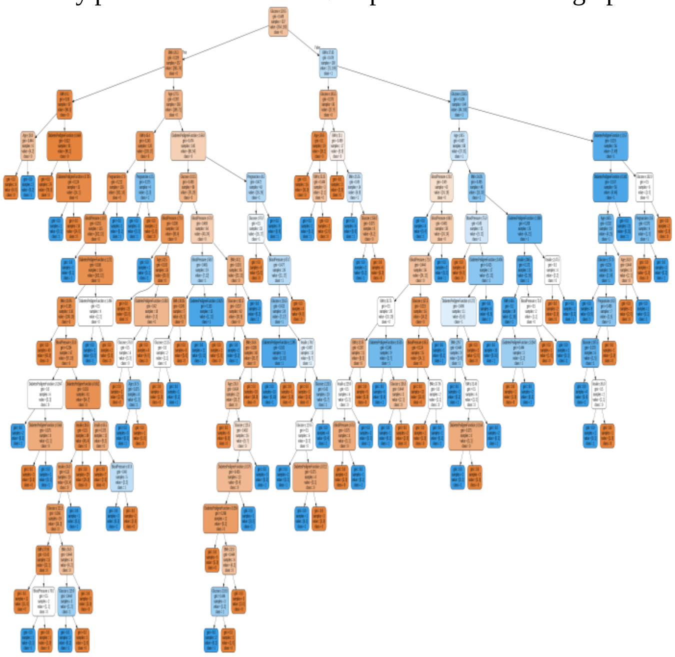


Fig.9 : Showing a Decision Tree generation for the dataset

VI. Conclusion and Future Work

Every time we may not get the same results with all different kinds of datasets for different kind of predicted outputs. So the choice of the method, precision values and available number of parametric values plays a vital role in the choice of the method. For the above mentioned ML methods if we apply AI techniques then further accuracy can be obtained which may be helpful for further correct prediction and to prepare further improvement measures to overcome the effect, also proper choice of validation and performance method are also important out of many available methods as discussed in the paper.

References

[1]. https://www.kaggle.com/uciml/pima-indians-diabetes -database.

[2]. Z. Punthakee, R. Goldenberg, and P. Katz, “Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome,” Can. J. Diabetes, vol. 42, pp. S10–S15, 2018.

[3]. M. N. Piero, “Diabetes mellitus – a devastating metabolic disorder,” Asian J. Biomed. Pharm. Sci., vol. 4, no. 40, pp. 1–7, 2015.

[4]. G. Swapna, R. Vinayakumar, and K. P. Soman, “Diabetes detection using deep learning algorithms,” ICT Express, vol. 4, no. 4, pp. 243–246, 2018.

[5]. Aakansha Rathore and Simran Chauhan, “Detecting and Predicting Diabetes Using Supervised Learning”. International Journal of Advanced Research in Computer Science, Volume: 08, MayJune 2017.

[6]. K. Kannadasan, D. R. Edla, and V. Kuppili, “Type 2 diabetes data classification using stacked autoencoders in deep neural networks,” Clin. Epidemiol. Glob. Heal., no. December, pp. 2–7, 2018.

[7]. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques,” Front. Genet., vol. 9, no. November, pp. 1–10, 2018.

[8]. P. Anusha, Aala Ravikiran,V. Lakshman Narayana,(2020), “Energy Priority With Link Aware Mechanism For On-Demand Multipath Routing In Manets”, International Journal of Advanced Science and Technology, Vol. 29, No. 03, (2020), pp. 8979 - 8991.

[9]. Venkata Rao Maddumala, K.Maha Lakshmi, P.Anusha, V.Lakshman Narayana,(2020), “Enhanced Morphological Operations for Improving the Pixel Intensity Level”, International Journal of Advanced Science and Technology, Vol. 29, No. 03, (2020), pp. 9191 - 9201.

[10]. B. Tarakeswara Rao, V. Lakshman Narayana,(2020), “Use of Blockchain in Malicious Activity Detection for Improving Security”, International Journal of Advanced Science and Technology, Vol. 29, No. 03, (2020), pp. 9135 - 9146.

[11]. C.R.Bharathi, V.Lakshman Narayana,(2020), “Unlimited Bandwidth for RF Applications Using Design and Examination of CMOS LNA”, International Journal of Advanced Science and Technology, Vol. 29, No. 03, (2020), pp. 9056 - 9062.

[12]. Lakshman Narayana Vejendla and Bharathi C R ,(2018),“Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs”, Smart Intelligent Computing and Applications, Vo1.1, pp.649-658. DOI: 10.1007/978-981-13-1921-1_63

[13]. Mohiddin, Shaik Khaja, and Y. Suresh Babu. "UNIQUE METHODOLOGY TO MITIGATE ANTI-FORENSICS IN CLOUD USING ATTACK-GRAPHS."