Hybrid Machine Learning and Implementation in Diabetes Classification

M Vamsi Krishna ECE,
SSN College of Engineering Kalavakkam,
Chennai

Akkala Thanmayi
IT,SSN College of Engineering, Kalavakkam,
Chennai

***Abstract*— This paper manages to predict the Diabetes of patient by applying an examination of three different hybrid Machine learning techniques. Further, by fusing all the current gamble elements of the dataset, we have noticed a steady exactness subsequent to characterizing and perform Cross-valuation. We have figured out a method to accomplish a steady and most noteworthy precision of 100% with triple hybrid algorithm which has a combination of XG Boost, Ada Boost and Random Forest. The main aim of this paper is to track diabetes at initial level and determine the accuracy rate using ML techniques and it has been successfully achieved in the results.**

Index Terms—Diabetes , insulin level, machine Learning algorithms, accuracy, k-nearest,Random forest, etc

I. INTRODUCTION

Diabetes has become one of the most common disease in today's generation. Diabetes is caused due to increase level of insulin in blood sugar level . In many cases it is difficult to predict diabetes at an initial stage. In order to overcome this drawback we have implemented machine learning techniques to predict diabetes at initial stage.

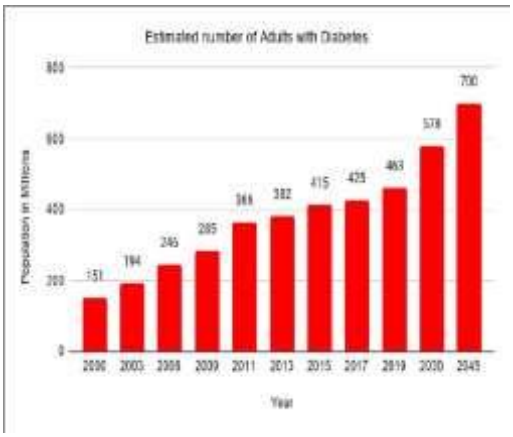

Fig. 1. Number of patients diagnosed by diabetes

In 2020 , approx. 468 million growing population predicted the patient from the age of 20-78 years had been diagnosed , with 72% of the population are women.

Diabetes is expanding gradually on the planet on account of ecological, hereditary elements. The numbers are rising quickly because of a few variables which incorporates undesirable food sources, actual inertia and some more. Diabetes usually occur in pregnant women due to their hormonal change and does affect the blood sugar level. Deep yearning, thirst and successive pee are a portion of the recognizable attributes. Certain gamble factors like age, BMI, Glucose Levels, Pulse, and so forth, assume a significant part to the commitment of the infection.

In the Fig. 1 gives an crystal clear view of increase in the number of cases for every year. This scenario is an crucial thing that diabetes has become as fastest growing disease in past few years and affect the health of an individual around the world.

Due to emergency of large amount of data Machine Learning has become very popular these days as it is used almost in all field starting from automation to healthcare industry where huge amount of data has to be processed.

Machine learning is one of the major and significant method that is at present being utilized in the business for performing information investigation and acquiring understanding of information. Machine learning mining utilizes various data mining methods. In this review, ML method is utilized for predicting disease. AI gives a pool of devices and strategies, utilizing these instruments and procedures raw information can be changed over into some noteworthy, significant data by PCs. There are four sorts of AI calculations that are at present being utilized.

Unsupervised learning is used o predict the pattern and used in modeling the process, where unsupervised learning is used to solve regression and clustering problem, semi supervised learning is an mix of both learning techniques. In this paper we have used this technique to resolve issues in medical field.

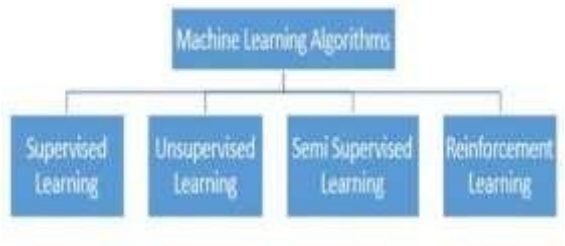

Fig 2- Types of ML algorithm

With the rise of ML and its relative techniques, it has become visible that the critical issues and blocks in its identification confronted before, can now be facilitated with much effortlessness, yet, giving a point by point and precise result. As of the present day, it is grasped that AI has become much more compelling and accommodating as a team with the space of Medication. Early assurance of an illness can be made conceivable through Ml by concentrating on the qualities of a person. Thus in this paper we help people in predicting diabetes in patient at early stage using machine learning algorithms.

1.2 Types of Diabetes

Type one diabetes – affects the pancreas ,it is well known as one polygenic disease or adolescent disease .This is commonly found in young people below age of 19.These patient use insulin on daily basis. Usually an diabetic patients should follow healthy diet and good exercise.

Type 2- thus may cause to damage of cell and when we neglect this process it will affect our hypoglycemic operation .The main reason for such type 2 diabetes is due to overweight. The WHO predicted the people affected by this type type 2 will be doubled at the end of 2026.

3) Type 3 Gestational diabetes is caused in women ho are pregnant , due to hormonal changes of the women there is increase in blood sugar level ,studies say that around 19% of people tend to have diabetes during pregnancy

The corpulence is one of the principle explanations behind type-2 diabetes. The type 2 diabetes are taken into immense care and we have to take proper exercise to overcome this diabetes. Studies show that around 30 milion people are affected by type 2 diabetes in India.

II. Literature Review

Over 70% of Indians experience chronic disease. The prediction of human disease like diabetes, malignant growth, asthma, Hypertension is finished utilizing different Machine learning approaches[7]. The prediction of type 1&2 are done by recognizing the symptoms like weight gain in rare cases, weightless, obscured eyesight. Veena has talked about, the diabetes disease is caused due to increase in sugar level in the plasma

Enormous informational collection assembled from various research centers, facilities, EHR and PHR handled in Hadoop, eventual outcomes then, at that point, conveyed over various servers as indicated by the topographical areas. Jiang in [13] introduced an exhaustive overview of writing over enormous information and examination. The central view of the author is to apply ML on modern power frameworks and to reduce the deficiencies when power is is increased.

In [14], a medical services expectation framework in view of Naive bayes is introduced. Proposed framework finds and concentrates stowed away information connected with various sicknesses from illness data set. This framework permits clients to share their wellbeing related issues and afterward utilizing Guileless Bayes anticipate the right ailment. For better expectation of heart illnesses in regular constant infection flare-up networks, creators smooth out the ML Technique in [15].

P. Suresh has introduced the calculations like decision Tree, for recognizing diabetes utilizing data mining methods [2]. Ruban has inspected the delayed consequence of estimations that are executed to propel the assurance trustworthiness [7].Laftaa developed an framework that helps the patients to predict cardiovascular breakdowns in [19]. The author explained in detail about coronary illness model, as indicated by the outcomes the framework likewise gives proposals to the patients that about the need of stepping through a few exam and visiting a specialist. The principle part of proposal framework depends on time Series information examination calculation. For advancement of the purposed framework genuine information was utilized. Writers led a pilot study on gathering of cardiovascular breakdown patients and accumulated information utilizing day to day clinical readings.

The [5] the author have utilized various of ML techniques to predict sickness and handle data by directing them to investigate data with classifiers. The main working of any processed data is to extract the data and utilize the information for handling data to produce the output. This in turn have reduced the computational time and increased the efficiency. Alongside the precision measure, we have used k-nearest algorithm for better results.In [10], the author has used images to examine ML to forecast the Hydrocephalus. The support vector machine has been implemented to ventricular components of 29 children[12]. Simitra et al. examined the meaning of early disclosure of female desolateness in [14].

The classifying algorithms are used in [13] for the model expectation and execution measurements: responsiveness, particularity and precision were dissected. Highlight choice (Head Part Investigation) was done and the outcomes are confirmed utilizing all elements, single component like Glucose . In [14] the author in his paper of different ML Technique and predicted accuracy and precision value The informational collection was pre-handled and

characterized and the precision was examined. With the outcomes RF was demonstrated to be more productive with an exactness of 75%. Augmenting region under the Recipient working trademark (ROC) bend utilizing the ideal hyper parameters. A 10-overlap cress-approval is done to break down the pattern in the forecast of concealed dataset. Arbitrary Backwoods ends up being critical.

(SVM) classifier is used to as a multi parameter detector since it detect blood pressure level, sugar level, heart beat rate [16]. This classifier helps in monitoring health where SVM has been modified and used in various software [17].

III.PROPOSED SYSTEM

3.1WORKING

In this paper we predict the diabetes using hybrid machine learning techniques . In this process the first step is dataset collection .we have collected data based on the different parameters like glucose level, Pulse rate , age limit, and people who are pregnant. After checking data we pre- process the data by splitting data into train data and test data with different classifier and random forest method so the data which has obtained is used to optimize the result and provide accuracy.

IV MODULES DESCRIPTION

In this paper we will be using combination of three hybrid algorithms for increasing the accuracy. First hybrid algorithm being KNN and random forest. The second hybrid algorithm is ADA Boost and Random forest. A random forest is used to solve regression. Third hybrid algorithm is combination of XG Boost, Adaboost, Random forest. When an input data is given for the disease prediction process, the features will be compared with the features of the model file and it will predict the presence of disease. In the project, we can easily determine the presence of diabetes with higher accuracy.

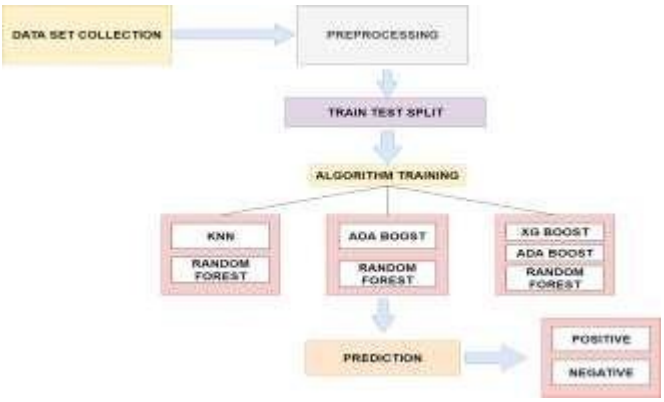

Fig 3 Proposed System Architecture

Dataset Collection
Dataset Pre-processing
Training with Hybrid Machine Learning Algorithms
Prediction

4.1 Dataset

The initial stage is to collect the dataset. The highlights of dataset are:

- Records rate of pregnancy
- Sugar level in blood
- Pulse rate.
- Weight Record (BMI)
- Skin crease thickness in mm
- Insulin esteem in 2 hour
- Genetic variable Family work
- Time of patient in years

We will collect the dataset for battery duration and the data collected in this model is fed into ML Algorithm. As dataset increase we can also increase the accuracy rate of the process.

Classification-Classification is used to answer yes or no inquiries and to make a multi-class arrangement .

Regression-For a calculation to yield some numeric worth. For instance, assuming that we invest an excess of energy concocting the right cost for our item. Since it relies upon many variables, regression calculations can support assessing this worth.

Ranking. Positioning is effectively used to suggest motion pictures in video web-based features or show the items that a client could buy with a high likelihood in view of their past hunt and buy exercises.

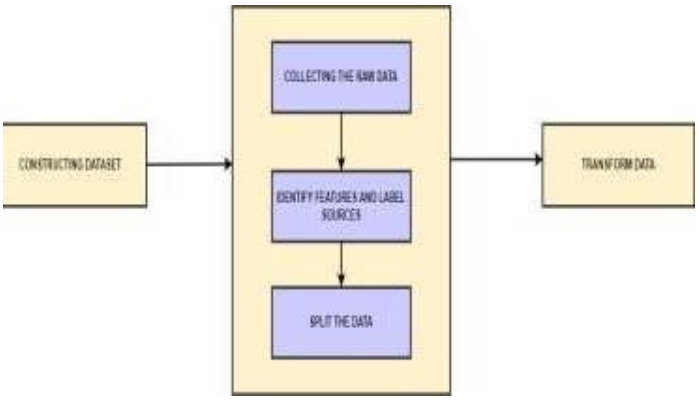

Fig 4 Dataset collection

The training of data set plays an major role after processing data into different category, where these data has to be split into train data and test data since there are huge amount of data .After splitting data we can categories the accuracy by implementing different classifier algorithm

After preprocessing the dataset, it can be directly applied to the AI for calculation. Experimentation technique is utilized for setting the rate for preparing and testing information so the characterization should be possible. The underneath calculations can measure up for expanding the presentation of the framework to foresee the presence of illness.

4.2 Pre-processing

Pre-processing technique such as Information Preprocessing is utilized to change the raw data into machine understandable data. At the end of the day, the data collected from various sources in raw format isn't feasible for examination.

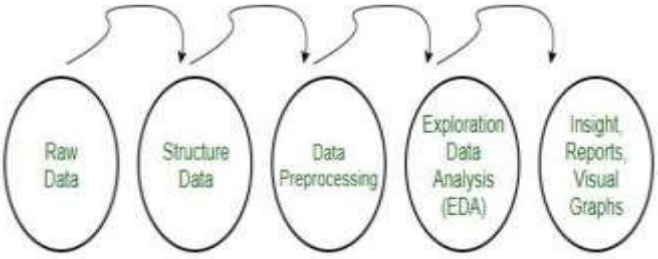

Fig 5 Process flow for pre-processing data

4.3 Training with Hybrid Machine learning Algorithms

In this paper, first the KNN and random forest are used as hybrid combination to further develop precision. This algorithm is used to tackle regression and complex issues. Secondly a combination of Ada Boost and Random forest is used and thirdly XG Boost, Ada Boost and Random forest is used for training the model.

The XG Boost algorithm is used to determine the data and compute them at high efficiency and with less computational time.The main objective is to create a model and access some key methods to implement sparse matrix to automatically store the missing data.

Table 1 -Data Statistic

Attribute	Count	Mean	STD	Minima	Maxima
Num_pregnant	800	3.92	3.26	0.00	18.00
Glucose_con	800	121.89	32.6	0.00	198.00
Bp_num	800	79.10	19.3	0	123.00
Thickness	800	23.51	15.92	0	98.00
Insulin	800	89.9	112.2	0	821.00
BMI	800	32.6	0.33	0	68.20
Age	800	0.57	12.7	0	3.42

4.4 Prediction

The main objective is to predict the diabetes of the patient that would be beneficial for the determining the health condition. This paper deals with prediction model i.e. the best Machine Learning algorithms which will accurately predict the presence of diabetes. In this project we have used combinations of KNN, Random forest, Ada boost and XG Boost algorithms for prediction. The better result is produced by algorithm combining XG Boost, ADA Boost and Random forest .

V. RESULT ANALYSIS

The main aim of this paper is to predict the diabetes using machine learning algorithm and use classifier algorithm to pre-process the data. To split the data test and train method has been implemented. The dataset collected for this project is shown in the below figure.

These datasets are then preprocessed to convert the data into vector format so that it can be trained .

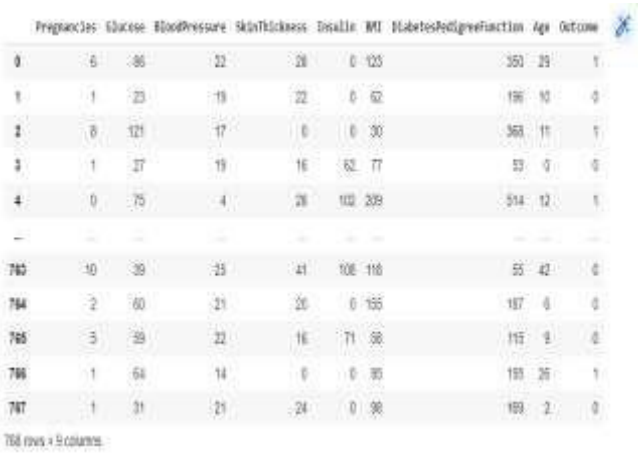

Fig 6 Dataset Preprocessing

After pre-processing the dataset, dataset is split into train and test so that the evaluation metrics can be calculated.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

print("X_train", X_train.shape)
print("y_train", y_train.shape)
print("X_test", X_test.shape)
print("y_test", y_test.shape)
```

```
X_train (614, 8)
y_train (614,)
X_test (154, 8)
y_test (154,)
```

Fig 7 Train Test Split

After that, training is performed with the hybrid machine learning algorithms. The machine learning algorithms used for hybrid model is K-Nearest Neighbour, Random Forest, Ada boost and XG boost.

Classification Report:

	precision	recall	f1-score	support
0	0.83	1.00	0.91	10
1	1.00	0.71	0.83	7
accuracy			0.88	17
macro avg	0.92	0.86	0.87	17
weighted avg	0.90	0.88	0.88	17

Fig 8 Classification report of K-Nearest Neighbour and Random Forest

Second hybrid algorithm used is Ada Boost and Random Forest. The below figure shows the confusion matrix for Ada Boost and Random Forest.

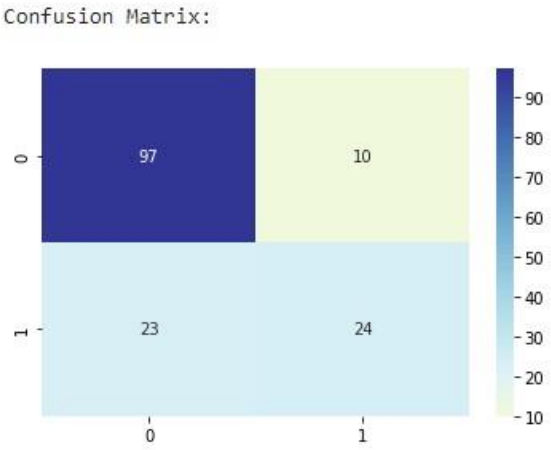

Fig 9 Ada Boost and Random Forest confusion matrix

The below figure shows classification report of Ada Boost and Random Forest.

Classification Report:

	precision	recall	f1-score	support
0	0.91	1.00	0.95	10
1	1.00	0.86	0.92	7
accuracy			0.94	17
macro avg	0.95	0.93	0.94	17
weighted avg	0.95	0.94	0.94	17

Fig 10 Classification report of Ada Boost and Random Forest

Third hybrid algorithm used is XG Boost, Ada Boost and Random Forest.The below figure shows the confusion matrix for XG Boost, Ada Boost and Random Forest.

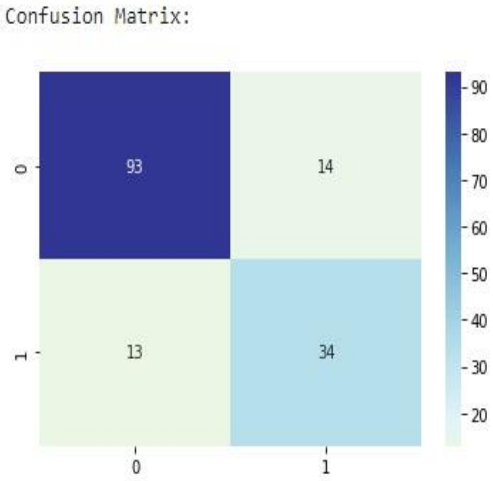

Fig 11 XG Boost, Ada Boost and Random Forest confusion matrix

The below figure shows classification report of XG Boost, Ada Boost and Random Forest.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10
1	1.00	1.00	1.00	7
accuracy			1.00	17
macro avg	1.00	1.00	1.00	17
weighted avg	1.00	1.00	1.00	17

Fig 12 Classification report of XG Boost, Ada Boost and Random Forest

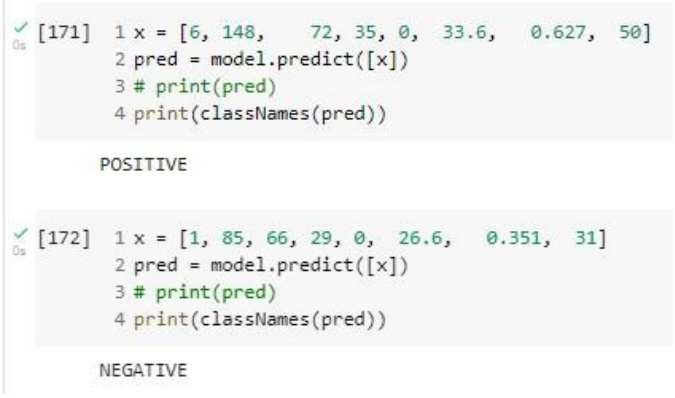

Fig 13 Prediction of Diabetes.

V. CONCLUSION

One of the major concern in the present medical field is not identifying the disease at early stage and not taking treatment at the right time. In this paper we endeavor various predicting techniques and determine the disease with high accuracy and reduce the computational cost. Based on the data investigated by (PID) Data set, we have promptly anticipated this disease. This method helps in giving productive treatment in a most efficient way and ultimately lessen the time expected for tracking down the infections. From the above obtained results we can see that the third hybrid model which consists of the XG Boost, Ada Boost and Random forest achieves the highest accuracy of 100% when compared to the other models, it can be used for real time disease prediction. Presently it is done physically which consumes additional time and furthermore includes human error. And also with these lines, by this venture This paper has suggested a method which decreases the time expected for manual arrangement and reduces the human mistake rate.

VI. FUTURE WORK

In the coming future, we will review the application of the diabetes forecast in the medical field and further develop it for recognizing different kinds of infections with more precision. In clinical field there are more opportunities to create or change our project in numerous ways. Hence, this undertaking has a proficient degree in coming future where manual methods can be automated in modest manner.

REFERENCES

[1] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, “Big Data Analytics in Healthcare,” Hindawi Publ. Corp., vol. 2015, pp. 1–16, 2015.
[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, “Big Data for Health,” IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, 2015
[3] E. Ahmed et al., “The role of big data analytics in Internet of
[4] Things,” Comput. Networks, vol. 129, no. December, pp. 459–471, 2017
[5] Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.
[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” Neurocomputing, vol. 237, pp. 350–361, May 2017.
[7] J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning for finance: deep portfolios,” Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.
[8] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, “Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings,” IEEE Trans. Autom. Sci. Eng.,
[9] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, “Localization Based on Social Big Data Analysis in the Vehicular Networks,” IEEE Trans. Ind. Informatics.
[10] P. A. Chiarelli, J. S. Hauptman, and S. R. Browd, “Machine Learning and the Prediction of Hydrocephalus,” JAMA Pediatr., vol. 172, no. 2, p. 116, Feb. 2018.
[11] A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, “Providing Healthcare-as-a-Service Using Fuzzy Rule- Based Big Data Analytics in Cloud Computing,” IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2018.
[12] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, “Predictive Methodology for Diabetic Data Analysis in Big Data,” Procedia Comput. Sci., vol. 50, pp. 203–208, Jan. 2015.
[13] J. Zheng and A. Dagnino, “An initial study of predictive machine learning analytics on large volumes of historical data for power system applications,” in 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 952–959.
[14] Journal of Advanced Computer andMathematical Sciences. Bi Publication-BioIT Journals, 2010.
[15] Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Infection Forecast by AI Over Huge Information From Medical care Networks," IEEE Access, vol. 5, pp. 8869-8879, 2017.
[16] R. A. Taylor et al., "Expectation of In-emergency clinic Mortality in Crisis Office Patients With Sepsis: A Neighborhood Huge DataDriven, AI Approach," Acad. Emerg. Medications., vol. 23, no. 3, pp. 269-278, Blemish. 2016
[17] S. Das and A. Thakral, "Prescient examination of dengue and intestinal sickness," in 2016 Worldwide Meeting on Registering, Correspondence and Mechanization (ICCCA), 2016, pp.172-176.
[18] M. S. Simi, K. S. Nayaki, M. Parameswaran, and S. Sivadasan, "Investigating female barrenness utilizing prescient scientific," in 2017
[19] IEEE Worldwide Helpful Innovation Gathering (GHTC), 2017, pp. 1-6.R. Lafta, J. Zhang, X. Tao, Y. Li, and V. S. Tseng, “An Intelligent Recommender System Based on Short-Term Risk Prediction for Heart Disease Patients,” in 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, pp.102–105.
[20] S. T. Prasad, S. Sangavi, A. Deepa, F. Sairabanu, and R. Ragasudha, “Diabetic data analysis in big data with predictive method,” in 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), 2017, pp. 1–4.
[21] W. H. S. . Gunarathne, K. D. . Perera, and K. A. D. C. . Kahandawaarachchi, “Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD),” in 2017 IEEE (BIBE), 2017, pp. 291–296.
[22] S. Jhajharia, S. Verma, and R. Kumar, “Predictive Analytics for Breast Cancer Survivability,” in Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies - ICTCS ’16, 2016, pp. 1–5.

[23] J. Finkelstein and I. cheol Jeong, "Machine learning approaches to personalize early prediction of asthma exacerbations," Ann. N. Y. Acad. Sci., vol. 1387, no. 1, pp. 153–165, Jan. 2017.
[24] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," Proc. Annu. Symp. Comput. Appl. Med. Care, pp. 261–265, 1988.
[25] B. M. K. Prasad, K. K. Singh, N. Ruhil, K. Singh, and R. O'Kennedy, Communication and Computing Systems :
[26] Proceedings of the International Conference on Communication and Computing Systems (ICCCS 2016), Gurgaon, India, 9-11 September, 2016. CRC Press, 2017
[27] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics Med. Unlocked, vol. 10, pp. 100–107, Jan. 2018.
[28] K. Kayaer and T. Yildirim, "Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks," Iternational Conf. Artif. Neural Networks Neural Inf. Process., pp. 181–184, 2003.
[29] Rahul Joshi and Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm": Ensemble approach,
[30] International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct -2017
[31] Zhilbert Tafa and Nerxhivan Pervetica, "An Intelligent System for Diabetes Prediction", 4th Mediterranean Conference on Embedded Computing MECO – 2015 Budva, Montenegro.
[32] Sumi Alice Saji and Balachandran K, "Performance Analysis of Training Algorithms in Diabetes Prediction", International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India 2015.