JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Review on Machine Learning Algorithms and Classification Techniques in Diabetes Medical Diagnosis and Healthcare Systems

G.Vinoda Reddy, Ch. Rekha, M.Sreenu Naik, Ms.Syeda Sumaiya Afreen

Professor, Department of CSE (AI & ML), CMR Technical Campus, Hyderabad.
Assistant professor, Department of CSE, CMR Technical Campus, Hyderabad.
Assistant Professor, Department of CSE (AI & ML), CMR Technical Campus, Hyderabad.
Assistant Professor, Department of CSE (AI & ML), CMR Technical campus, Hyderabad.

Abstract: Machine Learning spreads its wings to not only technical domains and also all domains including medical diagnosis of various chronic diseases. Among which diabetes is one of the major chronic diseases on the globe now a days. Diagnosis of diabetes is an open challenge in spite of in these advanced generations too. In this paper we attempted to have a review on machine learning algorithms and classification techniques for diabetes medical diagnosis. This study focuses on new developments in machine learning which have made significant impacts in the detection and diagnosis of diabetes. In this work, the machine learning algorithms are used to classifying diabetes patients. The classification accuracy is achieved by the classifying diabetes patients. Authors perform a review of the literature on machine models and suggest an intelligent framework for diabetes prediction based on their findings. The idea that had motivated us to present a review of various diabetic prediction models is to address the diabetic prediction problem by identifying, critically evaluating, and integrating the findings of all relevant, high-quality individual studies. In this paper, we have analysed the work done by various authors for diabetes prediction methods. Our analysis on diabetic prediction models was to find out the methods so as to select the best quality researches and to synthesize the different researches. Analysis of diabetes data disease is quite challenging because most of the data in the medical field are nonlinear, correlation structured, and complex in nature. Machine learning-based algorithms have been ruled out in the field of healthcare and medical imaging. We strongly recommend our study because it comprises articles from various sources that will help other researchers on various diabetic prediction models.

Key words: Diabetes, classification, prediction, machine learning, accuracy, healthcare.

I.INTRODUCTION

Diabetes is an infection that is caused due to over the top measure of glucose in it. Our body needs energy, and glucose is one of the primary wellsprings of energy to assemble the muscles and tissues of the body. By and large, unfortunate way of life and absence of activity are the primary drivers of type 2 diabetes in individuals. The presence of a lot of sugar in the blood causes diabetes. In some cases, the pancreas can't change over the food into insulin; in this way, sugar remains unabsorbed, which causes diabetes. Diabetes can influence kidney, eyes, sensory system, blood vessels, etc. Diabetes is of three sorts.

First is adolescent diabetes [1], which happens for the most part in kids and obliterates the cells which produce insulin in the pancreas. Second is type 2 diabetes, which by and large occurs after the age of 40 on the grounds that of absence of activity and unfortunate way of life. Diabetes is a sort of sickness that can't be turned around yet can be controlled with the help of drugs, customary walk and exercise, and a legitimate diet. Type 2 diabetes [2] is otherwise called insulin-autonomous diabetes since patients are not infused with insulin after a hole of normal stretches, however on account of type 1 diabetes, insulin is infused at a standard time span to the patient, so this also otherwise called insulin-subordinate diabetes.

The third sort of diabetes [3] is incubations, which happens during pregnancy due to the difference in chemicals, and this by and large vanishes later the conveyance. There is another condition, or at least, pre diabetes, in which the admission levels of sugar are on the fringe, and this condition can be switched with the assistance of ordinary activity also, solid way of life. In this paper, we have attempted to anticipate diabetes utilizing AI. AI is a part of man-made consciousness in which the machine attempts to foresee the result in view of specific information and past results. AI is of two kinds. First is directed learning, in which information go about as an educator and the model is worked around the dataset. Second is solo learning, in which information trains itself by tracking down specific examples in the dataset and naming them. As of late, many creators have distributed what's more, introduced their work on diabetes forecast by utilizing AI calculations. In this paper, we have considered different diabetes expectation techniques utilizing AI what's more, introduced a near investigation of a couple of strategies in our paper.

II. RELATED WORK

Healthcare frameworks offer modified administrations in broad ranging regions to help patients in coordinating themselves into their standard schedules of life. Diabetes mellitus is among the most critical extreme issues in the clinical calling. Order is among the most huge decision making strategies in the present commonsense conditions. The essential objective is to classify the information as diabetes or non diabetic what's more, increment the arrangement exactness. Machine learning in the determination of diabetes is for the most part about understanding designs from the diabetes dataset which would be given. AI lately has forever been the creating, reliable, and strong innovation in the clinical area.

This study is centred around the recognizable proof of diabetes kinds of patients in light of individual and clinical data using AI classifiers. This segment contains a synopsis of the works proposed by various specialists during the last 10 years. It is gainful to recognize the deficiencies of recommended works in the field of diabetic patients' treatment routine AI classifiers. Conclusion of diabetes is a developing area of study.

Sun and Zhang [1] have examined a couple of profound learning strategies and grouping techniques, for example, counterfeit brain network, choice trees, arbitrary timberland, and backing vector machine.

Qawqzeh et al. [4] have executed a strategic relapse characterization procedure for the order of diabetes information. Preparing information incorporates 459 patients, and testing information incorporates 128 patients. Order precision accomplished by the creators was 92% utilizing calculated relapse.

The significant detriment of the model was that it was not contrasted and the other diabetic expectation models and consequently couldn't be approved.

Tafa et al. [5] isolated the dataset into half preparation set and half testing set. The model was proposed utilizing a mix of naiıve Bayes and support vector machine calculations for diabetes expectation. Dataset was gathered from three distinct areas, and the proposed model was approved on this dataset. Eight ascribes were available inside the dataset, and it comprised of 402 patients, among which 80 patients were type 2 diabetic. Troupe of naive Bayes and backing vector machine has accomplished the precision of

97.6%, which is far superior than the calculations when run alone on the dataset, that is to say, Naıve Bayes accomplishing an exactness of 94.52 and support vector machine accomplishing 95.52%.

The creators have not referenced any pre processing procedure to sift through any undesirable values from the dataset.

Karan et al. [6] showed a new technique for diabetes conclusion by planning a scattered end to- end three-level inescapable medical care framework engineering using fake brain organization (ANN) registering. At the most fundamental level, sensors and wearable gadgets are utilized to screen crucial markers on the human body. At level 2, client-side gadgets, for example, PDAs and PCs act as a judge furthermore, communicator between both the essential and last levels. The third level end incorporates strong work area servers that give clients social government assistance organizations furthermore, data set activities. Utilizations of a fake brain network are applied to analyze ailments at both the following and ensuing levels. Fake brain organization calculations make the client and server model ward on them. This strategy propels computations and frameworks interchanges on the client and server sides by contingent upon the idea of ailments.

Sisodia and Sisodia [7] have applied Naive Bayes, choice trees, and backing vector AI calculations on the Pima Indians Diabetes Dataset, and the greatest precision to anticipate the diabetes was accomplished by Naive Bayes classifier.

The objectives of our study are as follows:

1) To deepen ourselves with the different diabetic prediction models.
2) To evaluate and discuss the existing models based on classification accuracy.
3) To discuss the various attributes required for the prediction of diabetes.
4) To identify the research gaps in the existing literature.
5) To present a comparative study of various diabetic prediction models.
6) To collect more and more information about the prediction of diabetes in the primitive stage.

The idea that had motivated us to review the various diabetic prediction model is to address the diabetic prediction problem by identifying, critically evaluating, and integrating the findings of all relevant, high-quality individual studies. To achieve our motivation for this review process, we have studied various articles on diabetic prediction models, and we have taken those articles in this review process that have satisfied the following criteria:

- Article must have discussed various predictive methods and machine learning algorithms for the classification of diabetes data.
- Article must have discussed various pre-processing techniques to filter the noisy data.
- Authors have validated their model against a few performance parameters such as sensitivity, specificity, accuracy, true positive rate, and true negative rate.
- Predictive models were compared with the other existing diabetic prediction models.

III. MACHINE LEARNING ALGORITHMS

Machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the extract information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise. Many industries apply machine learning to extract relevant data. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves without being explicitly programmed. Many mathematicians and programmers apply several approaches to find the solution of this problem which are having huge data sets.

Table 1. ML algorithms for various model building approaches

Learning type	Model building	Examples
Supervised	Algorithms or models learn from labelled data (task-driven approach)	Classification, regression
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, associations, dimensionality reduction
Semi-supervised	Models are built using combined data (labelled + unlabeled)	Classification, clustering
Reinforcement	Models are based on reward or penalty (environment-driven approach)	Classification, control

Supervised learning: It consists of a given set of input variables (training data) which are pre labelled and target data [5]. Using the input variables it generates a mapping function to map inputs to required outputs. Parameter adjustment procedure continues until the system acquired a suitable accuracy extent regarding the teaching data.

Unsupervised learning: In this algorithm we only have training data rather a outcome data. That input data is not previously labelled. It is used in classifiers by recognizing existing patterns or cluster in the input datasets [4].

Reinforcement learning: Applying this algorithm machine is trained to map action to a specific decision hence the reward or feedback Signals are generated. The machine trained itself to find the most rewarding actions by reward and punishment using past experience

There are massive numbers of algorithms used by machine learning are designed to erect models of machine learning and implemented in it [4]. All algorithms can be grouped by their learning methodology, as follows:

Regression algorithms: In Regression algorithms predictions are made by the model with modelling the relationship between variables using a measure of error[25]. Continuously varying value is predicted by the Regression technique. The variable can be a price, a temperature.

Instance based learning algorithms: In the algorithms which based on Instance, decision problem is a issue with illustration of training data build up a database and compare test data then form a prediction. Instance-based learning method is famous as lazy learner.

Algorithms using Decision Tree: Algorithms using Decision trees are used mainly in classification problem. They splits attributes in two or more groups by sorting them using their values. Each tree have nodes and branches [4]. Attributes of the groups are represented by each node and each value represented by branch [5].

Baysian algorithms: Machine Learning is multidisciplinary field of Computer Science like Statistics and algorithm. Statistics manages and quantifies the uncertainty and are represented by bayesian algorithms based on probability theory and Bayes' Theorem.

Data Clustering algorithms: This algorithm split items into different types of batches. It groups the item set into clusters in which each subset share some similarity. It is unsupervised learning method and its methods are categorized as hierarchical or network clustering and partitioned clustering.

Learning algorithms using Association Rule: Learning algorithms using Association rule are generally utilized by the organization commercially when multidimensional datasets are huge in size. They are used as extraction methods that can explore observed relationships between variables and data.

Algorithms using Artificial Neural Network: Artificial neural networks models are based on the biological neuron structure and uses supervised learning. It consists of artificial neurons which have weighted interconnections among units. They are also well known by parallel distributed processing networks.

***Deep Learning algorithms*:** Deep Learning methods upgraded the artificial neural networks They are more complex neural networks are large in size.

Algorithms using Dimensionality Reduction: Dimensionality reduction method is widely used in case of large number of dimensions, large volume of space concerned. Then that problem requires a statistical significance. Dimensionality reduction methods used for minimizing the number of dimensions outlined the item and removes unrelated and un essential data which lessen the computational cost. Some of these methods are used in classifying and regression.

Ensemble Algorithms: They are based on unsupervised Learning. It groups the teaching data into many types of classes of data. Self-supporting models for learning are built for those groups. To make correct hypothesis all learning models are combined.

Table 2. Summary of ML algorithms and examples

Type of Algorithm	Example
Regression algorithms	***Linear Regression algorithm*** ➢ Ordinary Least Squares Regression ➢ Multivariate Adaptive Regression Splines ➢ Logistic Regression ➢ Locally Estimated Scatter plot Smoothing ➢ Stepwise Regression
Instance based learning algorithms	***Learning Vector Quantization*** ➢ Self-Organizing Map ➢ k-Nearest Neighbor ➢ Locally Weighted Learning
Algorithms using Decision Tree	➢ Iterative Dichotomized 3 ➢ M5 ➢ Chi squared Automatic Interaction Detection ➢ C5.0 and C4.5 (various versions of a powerful approach) ➢ Decision Stump ➢ Classification and Regression Tree ➢ Conditional Decision Trees
Baysian algorithms	***Bayesian Belief Network (BBN)*** ➢ Multinomial Naive Bayes Bayesian Network (BN) ➢ Averaged One-Dependence Estimators (AODE) ➢ Gaussian Naive Bayes ➢ Naive Bayes
Data Clustering algorithms	***K Means*** ➢ Expectation Maximisation (EM) ➢ K Medians ➢ Hierarchical-Clustering
Learning algorithms using Association Rule	➢ Eclat algorithm ➢ Apriori algorithm
Algorithms using Artificial Neural Network	***Radial Basis Function Network (RBFN)*** ➢ Back-Propagation ➢ Perceptron ➢ Hopfield Network

Deep Learning algorithms	***Deep Belief Networks*** ➢ Stacked Auto Encoders ➢ Deep Boltzmann Machine (DBM) ➢ Convolution Neural Network (CNN)
Algorithms using Dimensionality Reduction	***Partial Least Squares Regression*** ➢ Multidimensional Scaling ➢ Principal Component Analysis ➢ Flexible Discriminant Analysis ➢ Mixture Discriminant Analysis ➢ Sammon Mapping ➢ Projection Pursuit ➢ Linear Discriminant Analysis ➢ Principal Component Regression ➢ Quadratic Discriminant Analysis
Ensemble Algorithms	***Gradient Boosting Machines*** ➢ Boosting ➢ Gradient Boosted Regression Trees ➢ Bagging ➢ Bootstrapped Aggregation ➢ Stacked Generalization (blending) ➢ AdaBoost ➢ Random Forest

Table 3 A comprehensive study of the machine learning methods done by some researchers

Algorithm	Method used/innovation	Application and future work	Results and limitations
J48, AdaBoost, And bagging on base classifier [2]	The model was performed on Canadian Primary Care Sentinel Surveillance Network dataset with several features to train on. The author used ensemble methods AdaBoost on base classifier J48 DT.	The author claimed that these ensemble algorithms can be used on other disease datasets to increase accuracy.	The AdaBoost algorithm with the J48 as the base classifier showed the maximum accuracy followed by bagging and then the J48 classifier. The AROC was used as the parameter.
NB with clustering [3]	Dataset used was the PIMA Indians Diabetes Dataset with eight attributes. The model is NB performed on prior clustering. This model is compared with only the NB model. Five hundred and thirty-one instances of data were divided into 5 clusters. The fourth cluster was the only one used for testing, which consisted of 148 instances.	By collecting a large amount of data for training, the accuracy can be increased by many-fold, helping people by developing a system that gives them a correct prediction without having to consult a doctor.	The parameters used for evaluation are accuracy, sensitivity, and specificity. The model with clustering showed a 10% increased accuracy, rise in sensitivity by 53.11% but the imitation caused here was the fall of specificity by 10.99% and also a reduced amount of dataset.
DTs, LR, and NB with bagging And boosting	Initial datasets were collected from primary care units, which (through further changes) consisted of 11 features and a data of 30122 people. The three algorithms are used along with bagging and boosting methods,	The final model obtained with highest accuracy was deployed on a commercial web application.	The following data shows the accuracy with bagging and boosting. DT 85.090, LR 82.308, NB 81.010, Bagging with DT (BG+DT) 85.333, bagging with LR (BG+LR) 82.318, bagging with NB (BG+NB) 80.960, boosting with DT (BT+DT) 84.098, boosting with LR (BT+LR) 82.312,

[4]	which are to decrease overfitting and increase accuracy.		and boosting with NB (BT+NB) 81.019. RF 85.558 shows the maximum accuracy. The ROC was used for final validation.
LR, KNN, SVM, LDA, NB, DT, and RF[12]	The author collected a raw dataset from Noakhali medical hospital containing 9843 samples with 14 attributes. Eighty percent of the data was taken for training and the rest for testing.	The author proposed that we can enhance the accuracy of early treatment to lessen the suffering of patients.	The RF classifier was the algorithm that performed the best in classifying data and LR showed the worst performance. Although machine learning classifiers are widely used, they still lack in terms of accuracy against deep learning models.
LR and DTs [12]	The dataset was prepared using a questionnaire carried out for 1487 individuals in which 735 were diabetic and the remaining 752 negatives. A Pearson chi-square test was carried out on all the characteristics. The models' performance was evaluated on three parameters: accuracy, sensitivity, and specificity.	Recently, many researchers have been implementing various algorithms and networks to compare them and find out the most feasible one. DTs and LR are among the ones that are most used.	LR achieved a ACC of s76.54%,sensitivity of 79.4%, and specificity of 73.54% on the testing data while the DT gained an accuracy of 76.97%, sensitivity of 78.11%, and specificity of 75.78%. Overall, the DT model performed better than the LR model. The model poses a limitation of the dataset. It is collected only from one area of China, if it had been collected from different regions, the model implementation could be more practical.
SVM and LR [13]	Practice fusion de-identified dataset was used for the study taken from Kaggle containing data of approximately 10000 patients.	LR is a model that is widely used in public health and clinical practice for disease detection and to calculate risks.	On using a smaller subset of features, the LR model performed slightly better than the SVM model.
DTs and NB[15]	The dataset taken for consideration was the PIMA Indian diabetes database. On applying feature selection, the author obtained five features. 10-fold cross validation was used for data preparation after which the J48 algorithm – DTs and NB is applied.	The author proposed to gather information for the dataset from different people to make a more representative model. The work can be further enhanced to include automation.	Using a percentage split of 70:30, the J48 DT algorithm correctly classified 177 instances (76.95%) whereas the NB got an accuracy of 79.56%. The accuracy obtained performed better on the percentage split, which shows
RF and XG Boost [16]	The author used the PIMA diabetes dataset. Using Jupyter Notebook as an IDE, the author trains the model using 8 attributes of the total 9 provided in the dataset.	The author suggests the use of more algorithms in this branch of machine learning like hybrid model for better accuracies.	The accuracy gained on the RF classifier came out to be 71.9%. The hybrid model proposed through XG boost gained an accuracy of 74.1%.
DT (J48) and	The author used the PIMA Indian diabetes dataset with 8 attributes, which was reduced to 5 based on the feature selection. The pre-	In future, it is planned to gather the information from different locales over the	The J48 algorithm was 76.95% accurate with other parameters like kappa statistic, MAE, RMSE, relative absolute error, and root relative absolute error. The NB algorithm was

NB[17]	processing was performed used the WEKA using 10-fold validation. The model was created using the 70% dataset and the rest was used fortesting.	world and make a more precise and general prescient model for diabetes conclusion.	accurate up to 79.56%. Since this model is not optimally configured, a developed model would require more training data for creation and testing.
LR, DT, RF, SVM [19]	The author used the PIMA Indian women dataset concerned with women's health with 8 attributes. Different models were trained for this dataset under different hyperparameters.	The author proposed to create advanced models on RF because of its highest accuracy and ability to overcome overfitting.	Different models were compared on basis of accuracy. RF gained the highest accuracy with 77.06% followed by SVM.

Table 4. Comparative analysis of diabetes prediction using machine learning methods

S.No	Method Name	Dataset	Size	Classifier used	Feature selection method	Speed	Classification accuracy
1	Kamrul Hasan	PID	768	NN, DT, RF, MLP,AB, XB, and NB	CA, ICA, and CRB	Slow	78.9%
2	Quan Zou	uzhou and PIDD	68994, 768	N J4	PCA and mrM	Slow	80.84%
3	Nishith Kumar	PIDD	768	GPC, LDA, QDA, and NB	Kernels	Fast	91.97%
4	Maniruzzuman	NHANES	9858	NB, DT, RF, and AB	LR	Slow	92.75%
5	V. Jackins	PIDD	768	B and RF	CRB	Fast	74.46%
6	N. Sneha	PIDD	2500	SVM, RF,NB, DT, and KNN	CRB	Slow	82.3%
7	S. Mohapatr	PIDD	768	MLP	None	Fast	77.5%
8	D. Sisodia	PIDD	768	NB, SVM,and DT	None	Fast	76.3%
9	Orabi	Egyptian National Research Centr	Not mentioned	DT	Not mentioned	Slow	84%
10	O. M. Alad	PIDD	768	NN	None	Fast	Prediction

V. OBSERVATIONS

After conducting a survey of various articles on diabetic prediction models, we strongly recommend our study because of the following reasons:

a) We have included recent articles.
b) We have presented a comparative statement of major diabetic prediction models which will help other researchers to understand and evaluate the models.
c) Advantages and disadvantages have been presented.
d) Various Strategies to predict diabetes have been discussed in the paper.

All the methods proposed so far for the prediction of diabetes are focusing more towards feature selection strategy and few machine learning methods such as random forest, naive Bayes, support vector machine, and decision trees, whereas only a few features are to be selected for prediction purpose. While studying all of these articles, the challenges that we faced are as follows:

a) The major challenge in prediction purpose was the absence of a larger dataset since the publicly available dataset contains only nine attributes, one being the class attribute. Time and effort are being spent on those features that have no potential to be selected for prediction purposes.

b) Most of the authors have dropped missing values from the standard dataset, which can affect the results as the size of the dataset decreases.

c) General machine learning algorithms are applied to the dataset; only one author has made use of AdaBoost and gradient boost technique. None of the authors has made use of the recurrent neural network or deep learning technology, which can help in increasing the efficiency. So, a method needs to be developed which can deliver more accurate results, has to be fast in terms of processing, and is more effective for the prediction purpose.

VI. CONCLUSION

This study focused on machine learning classification algorithms for predicting diabetes disease with more accuracy. In this study, various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms. Diabetes can be devastating after a certain period if not detected or diagnosed correctly. Many machine learning methods have been discussed, starting from different basic algorithms such as the LR, SVM, DTs, to further classification including the ID3, C4.5, C5.0, J48 and CART and NB. Ensemble methods, such as bagging, boosting, and RF regressors, are further used to enhance the accuracy and performance of models. Most researchers have implemented several algorithms in both machine and deep learning to compare their performance on the data, while others have combined two or three methods to gain more accuracy on a single system.

VII. REFERENCES

1. Perveen S, Shahbaz M, Guergachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci 82:115–121.

2. Khan NS, Muaz MH, Kabir A, Islam MN (2019) A machine learning-based intelligent system for predicting diabetes. Int J Big Data Anal Healthc 4(2):1.

3. Nai-Arun N, Moungmai R (2015) Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci 69:132–142.

4. Roshi Saxena,"A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey", Hindawi, Journal of Healthcare Engineering Volume 2022, Article ID 8100697.

5. Mahesh Kotha, "A Survey on Predicting Uncertainty of Cloud Service Provider Towards Data Integrity and Economic" 2019 IJSRST | Volume 6 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X.

6. Kocher T, Holtfreter B, Petersmann A, Eickholz P, Hoffmann T, Kaner D et al (2019) Effect of periodontal treatment on HbA1c among patients with prediabetes. J Dent Res 98(2):171–179.

7. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q (2013) Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 29(2):93–99.

8. Sheikhi G, Altınçay H (2016) The cost of type II diabetes mellitus: a machine learning perspective. In: Kyriacou E, Christofides S, Pattichis CS (eds) XIV mediterranean conference on medical and biological engineering and computing 2016. IFMBE proceedings, vol 57. Springer, Cham, pp 818-821.

9. Mahesh Kotha, "Predict Trustworthiness of Cloud Services Using Linear Regression Model", International Journal of Advanced Science and Technology Vol. 29, No. 4, (2020), pp. 5737 – 5745.

10. Raja Krishnamoorthi,"A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques", Hindawi, Journal of Healthcare Engineering Volume 2022, Article ID 1684017.

11. Ravindra Changala, "Automated Health Care Management System Using Big Data Technology" ,Journal of Network Communications and Emerging Technologies (JNCET), Volume 6, Issue 4, April (2016), 2016, pp.37-40,ISSN: 2395-5317, ©EverScience Publications.

12. Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Manag Process 5(1): 1–14.

13. Barik S, Mohanty S, Mohanty S, Singh D (2021) Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In: Mishra D, Buyya R, Mohapatra P, Patnaik S (eds) Intelligent and cloud computing. Smart innovation, systems and technologies, vol 153. Springer, Singapore, pp 399–409.

14. Ravindra Changala, "DEVELOPMENT OF PREDICTIVE MODEL FOR MEDICAL DOMAINS TO PREDICT CHRONIC DISEASES (DIABETES) USING MACHINE LEARNING ALGORITHMS AND CLASSIFICATION TECHNIQUES", ARPN Journal of Engineering and Applied Sciences, VOL. 14, NO. 6, MARCH 2019, ISSN 1819-6608

15. Ephzibah EP (2011) A hybrid genetic-fuzzy expert system for effective heart disease diagnosis. In: Wyld DC, Wozniak M, Chaki N, Meghanathan N, Nagamalai D (eds) Advances in computing and information technology. first international conference, ACITY 2011, July 2011. Communications in computer and information science, vol 198. Springer, Berlin, Heidelberg, pp 115-121.

16. M.Rajeswari,"A Review of Diabetic Prediction Using Machine Learning Techniques", International Journal of Engineering and Techniques - Volume 5 Issue 4,July 2019.

17. Ravindra Changala, "Statistical Models in Data Mining: A Bayesian Classification" in International Journal of Recent Tredns in Engineering & Research (IJRTER), volume 3, issue 1, pp.290-293. in 2017.

18. Zheng T, Xie W, Xu LL, He XY, Zhang Y, You MR et al (2017) A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 97:120–127.

19. Md. Kamrul Hasan, Md. Ashraful Alam, D. Das, E. Hussain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Acess, vol. 8, pp. 76516–76531, 2020.

20. Umair Muneer,"Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", Hindawi,Journal of Healthcare Engineering Volume 2021, Article ID 9930985.

21. V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," Fe Journal of Supercomputing, vol. 77, no. 5, pp. 5198–5219, 2021.

22. Ravindra Changala, "EVALUATION AND ANALYSIS OF DISCOVERED PATTERNS USING PATTERN CLASSIFICATION METHODS IN TEXT MINING" in ARPN Journal of Engineering and Applied Sciences,Volume 13, Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018.

23. Aishwarya M,"Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science 165 (2019) 292–299.

24. N. Sneha and T. Gangil, “Analysis of Diabetes Mellitus for Early Prediction Using Optimal Feature Selection,” Journal of Big Data, vol. 6, 2019.

25. Ravindra Changala, “Integrating Different Machine Learning Techniques for Assessment and Forecasting of Data” in Springer series, August-2015.