ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse119952020.pdf
https://doi.org/10.30534/ijatcse/2020/119952020

Prognosis of Chronic Diseases Using Different Machine Learning Models

Anand Javali[1], Suchithra R.[2]
[1] Ph.D. Research Scholar, Department of Computer Science and Engineering, Jain Deemed University, India
anand.javali@gmail.com
[2] Director Department of Computer Science and Engineering, Jain Deemed University, India
suchithra.suriya@gmail.com

ABSTRACT

Chronic diseases are the diseases that persist for a long time these are classified as the conditions persist for one year or more and require regular monitoring and ongoing medical attention on a day-to-day basis. Prediction of such chronic disease in human beings can be very useful for preventing these situations and know the complication in advance; the prediction model would give information such as the disease that patients might prone to, the more likelihood of the disease that might encounter in near future. The basic idea of this prediction model is to analyze and find out various chronic diseases based on various contributing parameters. This methodology of predicting chronic disease provides real-time assistance on diagnosis and prediction of the major chronic disease that is Diabetes. In this research data analysis and prediction, modeling is done by using the large datasets consisting of larger sets of different types of attributes. Research is done to identify which model works better, from the list of many machine learning algorithms; few algorithms were explored in the current paper. We are proposing a system that mainly works on five machine learning algorithms, those are, Naïve Bayes, Decision Tree, KNN Classifier, Random Forest, and SVM. The study of the prediction of chronic disease also involves the comparison of these different models to seek efficient results. Achieving high accuracy is an important aspect, based on the model that provides higher accuracy and efficacy the model will be selected as the final solution. In this research work, multiple machine learning algorithms are analyzed to predict the Chronic Diseases, among these analysis models Naïve Bayes will give the best results. In future work, this research would be extended to predict more such chronic disease and combine all of those models to create a better health prediction system. This Research can be applied in the medical field to diagnose and predict chronic diseases, this helps to reduce the dependency of medically trained professionals, and this will reduce the cost, effort, and time and enables the early detection of chronic diseases.

Key words: Chronic Disease, Diabetes, Disease Prediction, Health Informatics, ML Models, Medical Informatics.

1. INTRODUCTION

Chronic diseases are one of the major concerns in the medical field. A disease that will remain to show its effects for long periods is termed as a chronic disease. Many prolonged diseases lead to major deaths worldwide. In recent years, people of all age groups are suffering from different types of chronic diseases. Sometimes it becomes difficult for people themselves to identify whether they are prone to any such disease.

The main problem in the existing system is identifying the diseases and rectifying methods that are expensive and time-consuming. To avoid these problems various Machine Learning algorithms are used based upon the sample datasets results are proved. The objective of the research work aims to predict all the chronic diseases in human beings. Predictions of Diabetes for the person will be identified by various factors. Using Machine Learning Algorithms and sample datasets results are predicted [3]. Different types of chronic diseases include Various Cancers, Alzheimer's disease and dementia, Arthritis, Asthma, Diabetes, Kidney Disease, and HIV/AIDS, etc. According to many Healthcare institutions and Medical, Researchers diabetes is one of the top trending diseases. Nowadays, many people are suffering from diabetes, it will be termed to prolong period once the person encounters in their lifetime.

Diabetes is majorly categorized into two types, Type1 and Type2. Type1 diabetes occurs mostly in human beings in early life or teenage. People encounter with Type2 diabetes in their later part of life, adult age. A large range of factors defines the existence of diabetes in a human. Some important factors causing diabetes are Age, Gender, Obesity, Lifestyles, heredity (family background), BP (Blood Pressure).

A complete Prognosis and Recommender system for predicting chronic disease in human beings should predict all disease signs encounters to the human being. However, according to Medical Science, many diseases are interconnected and interlinked. Likely Chronic Disease is

also interconnected; suppose a person has Diabetes then that patient will have the risk of getting affected with heart-related problems. Similarly, a patient suffering from heart disease might be stated at risk of Kidney failure due to improper blood pumping.

In this research, we will first work on one chronic disease. As mentioned above Diabetes is one of the top trending diseases. Diabetes is taken as the first disease to be predicted. A primary dataset of Women patients suffering from diabetes mellitus is taken from The National Institute of Diabetes and Digestive and Kidney Diseases which consists of 9 different parameters that decide whether that patient has diabetes or not. The Machine Learning Classification models used in this paper for prediction are Naive Bayes, Random Forest, Decision Tree, and SVM. The dataset is first pre-processed and data cleaning methods are applied before training it with the machine learning models. These will help in achieving high and constant accuracies.

The Prognosis of diabetes will help a person to know if he/she will get the disease based on many factors like in the data sets mentioned above. The Machine Learning models are trained by the data set (train data and test data) and the outcome will be either yes or no whether a person is affected with diabetes. In future work, the research would be continued for other chronic diseases as well. The dataset will be huge and it might consist of all the parameters that might be enough to detect which chronic disease a human being will get affected and the risks of getting affected with those other interlinked chronic diseases.

2. LITERATURE REVIEW

Related work describes the background work done on various techniques and concepts related to the proposed system. The purpose is to select the documents related to our work. This provides new ideas, information, data, and evidence to fulfill certain aims or views of the nature of the topic. Wide applications of the deep mining used in applications like alphaGo[1] can be used in the medical field as well. In clinical experiments, input factors are not completely independent as there will be complex interaction[2] hence outcomes and predictions might not require to specify in the hypothesis.

The author, Akbar Waljee[3], et.al, in this research they proposed the medical diagnosis for various diseases through the Statistical model. These models are unable to predict accurate results. Kim, S.; Chang, Y [4] et.al. Describes the cohort study for the patients during the particular period. The connection between asymptomatic hyperuricemia and the development of nephrolithiasis is unidentified. This study uses basic methods to observe the patients it is a time consuming also, in this study, serum uric acid level increased was diffidently and associated nephrolithiasis higher risk in a dose-response method for healthy person.

Sangwoo Lee[5] et.al., determine a comparative study on different ML algorithms KNNC, DTC, NBC, RFC, SVMC, and DAC for the hyperuricemia prediction using EHR data. Some concepts of these works are used in our research work in the study on diabetes patients. Yang Guo[6], et.at., presents the study on Diabetes mellitus, it is a chronic disease and a health challenge global. Using data mining methods to aid people to predict diabetes is a widespread topic. In this paper, the Bayes Network was proposed to predict patients with developing Type-2 diabetes. Here study was limited to only one algorithm.

Chen, Min [7], et al., defined the diseases related to different regions, which exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. Here a new Convolutional neural network (CNN)-based multimodal disease threat estimate algorithm using structured and unstructured data from the hospital. They used both data sets for the prediction of the best results. Asmaa S. Hussein [10] et al., describes a CDD recommender system approach for the hybrid method, applying different Unified Collaborative Filtering (CF) also classifications. Multiple classifications are decision tree algorithms applied to build an efficient predictive model. This model provides higher accuracy and effective endorsements to guide patients in controlling their health conditions and help patients through the remote monitoring system.

Efficient Pre-processing is the key process to get high accuracy. Jathin Desan [19], The segmentation by k-means clustering, Gaussian Filtering, extracting features by DWT and GLCM followed by KNNRBF classification. the extracted features with KNNRBF is most optimal for patient data processing

In this paper, Models Learning for identifying contributing factors from Patient's Historical Data. Collecting datasets and then pre-processing & cleaning it. Then we use ML algorithms to train data and predict whether the person has that particular chronic disease. In this paper, we are doing it only for diabetes. Future work we will work on other chronic diseases and interconnect them to form a complete system of prediction for chronic disease in human beings.

3. MATERIALS AND METHOD

The proposed work uses several Machine Learning models to predict Chronic Diabetes Disease (CDD) [8, 9]. The models used to make the predictions are Naive Bayes, Decision Trees, Random Forest, and SVM. The data set for this project is taken from Kaggle. The data set contains 769 different

instances and has 8 attributes namely Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function, Age, and outcome. The data obtained were then pre-processed to fill the missing data and convert the data to a simpler format. The processed data is then used to train the machine learning model to predict the CDD [11]. The models are evaluated based on the accuracy and confusion matrix. The data is analyzed for different chronic diseases. Preprocessing and data integration is done for the datasets.

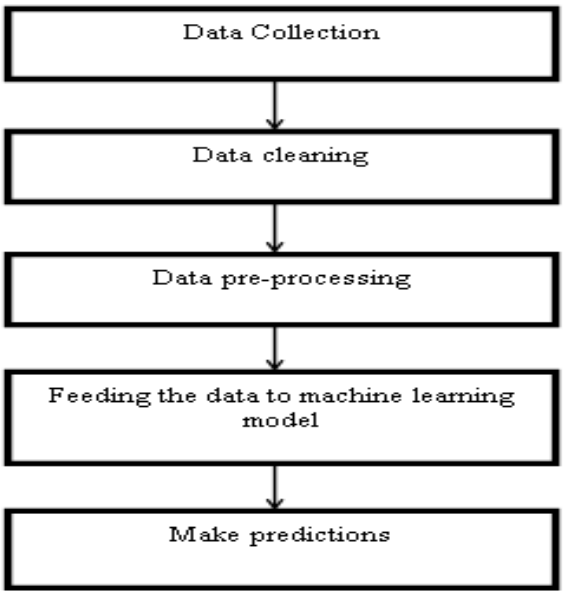

Figure 1: Machine Learning Prediction Model

Accuracy: It is the number of exactly classified cases divided by the total number of cases present in the data set.
Accuracy = exactly classified cases/ total number of cases.

Entropy: this relays to machine learning, it is a measured data randomness processed set of data [13]. The higher the entropy, the harder it is to draw any conclusions from that data.

Formula: (log(n/(p+n))) x (-p/(p+n)) x (log(p/p+n)) (n/(p+n))

(Here p and n are no of classes belonging to the positive and negative class in the data set.) Information gain formula: It is the difference in entropy before and after splitting data set on an attribute.

Gain = Entropy-Average information

In this proposed Model, Dataset is collected from Kaggle with 769 instances and 8 different attributes. These datasets are segregated first and in the next stage, data is preprocessed for feeding the model. By using the various libraries and methods results are predicted, finally, these results are compared for analysis. To increase the accuracy the did explored and used similar method as increasing the performance of [20] ELSA (Efficient Linear Selection of Adaptive Features) algorithm like PSO (Particle Swarm Optimization), BAT, ACO (Ant colony Optimization) along with the extractors of features this will provide higher accuracy in medical applications.

For the diagnosis, we can think of using impactful factors such as in case of Alzheimer Disease the Spectral-Domain Optical Coherence Tomography (SD-OCT) used for daignosis [21] in such case retinas inner layer segments from retina images which are colored images using chromatography processes. On same aspects we can use blood glucose, body mass index etc. for identifying the patterns
The following Machine Learning models are used to implement the Chronic Disease Detection (CDD).

3.1 Naive Bayes Model

The Naïve Bayes classifier is a machine learning model which is used for classification task based on the attributes. The classifier is based on the Bayes Theorem. This machine learning model is fast and easy to implement. It is a classification algorithm for both binary and multi-class classification. The representation of naive Bayes is probabilities. The probabilities for each attribute is are calculated and then the model classifies or predicts the output.

Naive Bayes formula: P(A/B)=P(B/A)*P(A)/P(B)

Predictions of diabetes disease based on various attributes in the dataset are done, using the Naïve Bayes model. It produces the accuracy of output with 84.374 (84%). In this model gaussianNB() function is used for the train_x and train_y trained datasets.

3.2 Decision Trees

The decision tree is a popular and most powerful tool for classification and prediction. It has a flowchart tree-like structure that has a decision node having branches (the outcome of the test) and a leaf node (terminal node). Decision trees have a clear indication about which attribute is important for prediction or classification.

A decision tree is very useful in the field of medical science where several parameters are involved in the classification of the data set [15]. The decision tree shows the important attributes among all the parameters and it gives the result based on the entropy and information gain, hence the data set plays a very important role. In diabetes, the decision tree can show the parameters that affect the patients such as Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function, and Age.

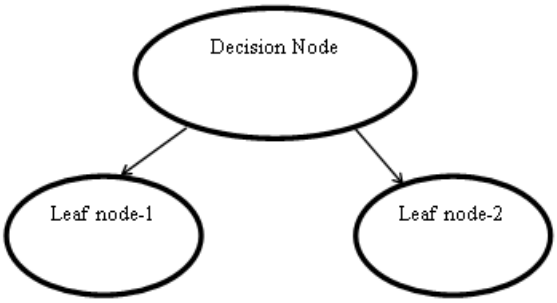

Figure 2: Decision Tree with Nodes

Prediction of diabetes disease based on various attributes in the dataset, using the Decision tree model. It gives the accuracy of output with 71.43(71%). It uses the DecisionTreeClassifier() function with criterion, random_state, max_depth and min_samp as a parameters.

3.3 Random Forest

The random forest has an ensemble learning approach towards classification and prediction. It consists of a large number of decision trees where each tree in the random forest gives an output prediction and the class with the most number of votes is considered [12]. It uses feature randomness when building each decision tree to create uncorrelated trees in the forest. Indecision trees overfitting is a concern, but in Random Forest overfitting can be handled. It will be used for classification and regression problems [16]. In this model sns.distplot() function is used to predict the results, here actual results are compared with the target results.

A Random forest would look like with two trees

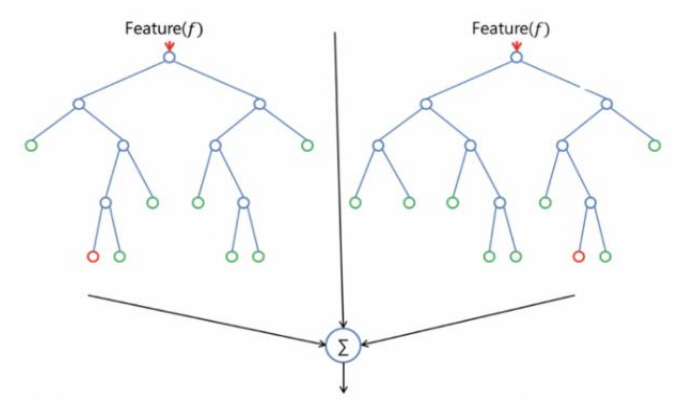

Figure 3: Random Forest with Two Trees

3.4 Support Vector Mechanism

Support Vector Mechanism is a supervised machine learning algorithm that can be used for classification and regression [17]. The SVM model is mostly used for the classification problem. In this algorithm, the data points are plotted on the n-dimensional space and we perform Classification by choosing the hyper-plane that divides the two classes very well. The arbitral decision line is drawn somewhere between the nearest point of the two data. These points act as the support vector.

Using the SVM model, the Prediction of diabetes is based on various attributes and instances in the dataset. It produces the prediction result of 81.17 (81%).

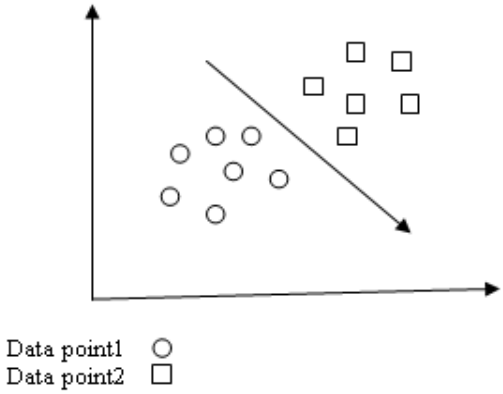

Figure 4: Model for Support Vector Mechanism

3.5 KNN Classifier

K-Nearest Neighbor (KNN) algorithm produces the results with Using different test sizes and n_neighbors. It is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution[18]. KNN will be used for both Classification and Regression predictive problems. It is widely used for Industry problems. KNeighborsClassifiers(n_neighbor) function will be used to classify and predict the data as shown in Figure 5.

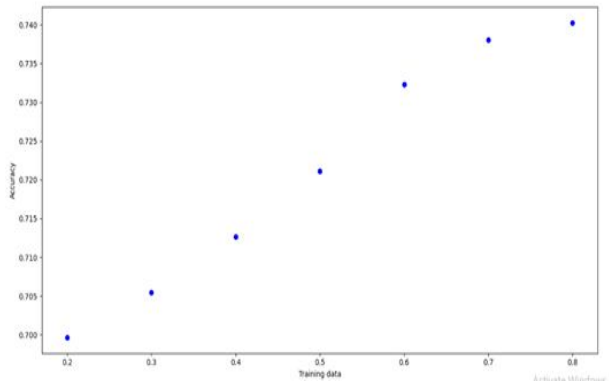

Figure 5: Predicted results from the KNN Classifier

4. RESULTS AND DISCUSSION

For the diagnosis of diabetes disease, several Machine Learning algorithms are used. SVM and Naïve Bayes are the common machine-learning models being used in recent

times. For our analysis, we use two other ML algorithm's namely decision trees, KNN, and random forest. To visualize the relationship between Prediction and Target Data. Actually to predict the unseen objects using Prediction Data.

In Dataset, 9 instances are used for Training and Outcome labels are used to testing Linear Relationship between Target and Prediction Data as defined in figure 5.

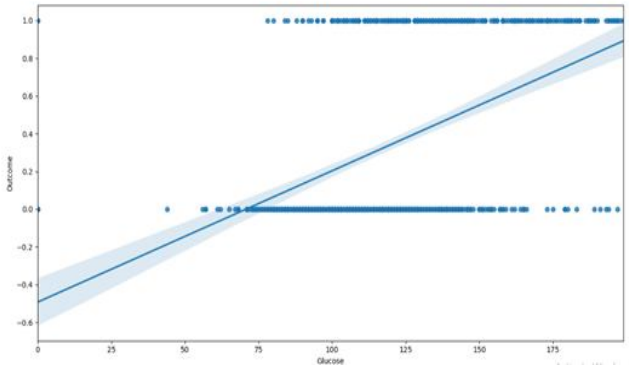

Figure 5: Testing Linear Relationship between Target and Prediction Data

The below results show, Prediction Model of the Naïve Bayes Algorithm. Here actual data is compared with predicted data. This results in the red color line show the target data and the blue color shows actual results as represents in figure 6.

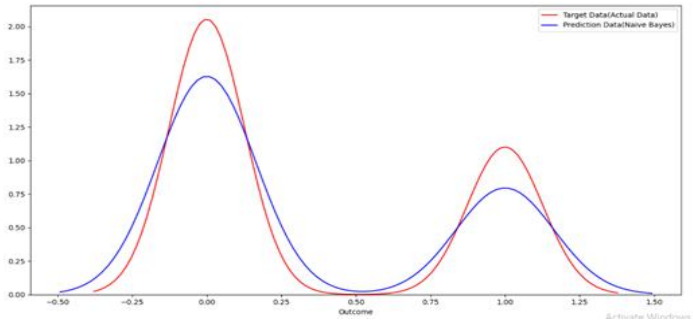

Figure 6: Prediction Model of Naïve Bayes Algorithm

Figure 7 explains about the visualization of all models explained in Table 1. By using the accurate results of all models, the graph has been generated. Red color * (asterisk) shows the accuracy points.

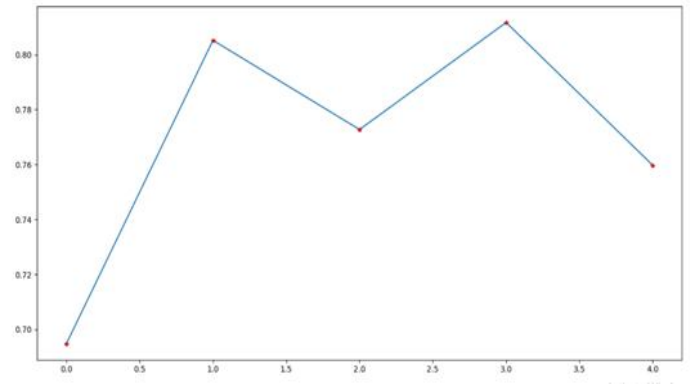

Figure 7: Visualization of all Models

It includes the Decision Tree, Logistic regression, and Naïve Bayes, SVM, and KNN Classifier algorithm with the main function for analysis of results. Among these algorithms, Naïve Bayes and SVM produces high accuracy results.

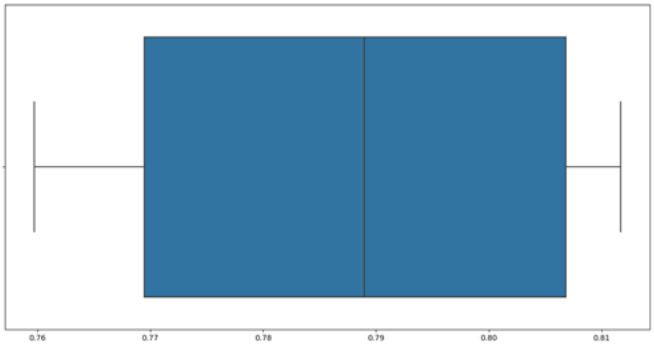

Figure 8: Total project Accuracy Visualization

Figure 8 shows the Total Project Accuracy Visualization. After analyzing the accuracy of different ML algorithms, we can conclude that 84.17% of Naïve Bayes is the best-suited algorithm for predicting chronic diabetes disease followed by Support Vector Mechanism with 81.17%. The following Table 1 shows the comparison between the accuracy of different ML algorithms used for the analysis.

Table 1: Comparison of different Machine Learning Algorithms

Model Name	**Model Constructor**	**Accuracy**
Decision Tree Classifier	dtf = DecisionTreeClassifier()	0.7143
Random Forest	rf = RandomForestRegressor.Predict(Fearueset)	0.8052
Naïve Bayes	nb= GaussianNB()	0.84374
SVM	svc= SVC(kernel='linear')	0.8117
KNN Classifier	knn=KNeighborsClassifier(n_neighbors=n)	0.7597

5. CONCLUSION AND RECOMMENDATIONS

Using Machine Learning techniques, we can extract the knowledge with an efficient way to produce the results from the EHR, which contains the patient's historical data and the current patient details. These mechanisms will help doctors and other medical staff to make correct and more appropriate decisions on diagnosis and treatment, which leads to better health care service improvement [14]. ML-based EHR Model is trained using datasets from one EHR and used for predicting the patient outcome.

In this research, we have applied ML modeling on clinical informatics big data analysis to predict chronic diseases such as Diabetes leveraging the programming paradigm, concepts of Artificial Intelligence, and Data Science. Since prevention is always better than cure, this research enables the model to the prediction of chronic diseases that will help patients to take precautionary majors and try to prevent the occurrence. We have multiple algorithms built around Machine Learning and advanced analytics concepts to provide statistical analysis and clinical outcomes. The main data source used is EHR containing patient historical data and real-time data. Data sets include diagnosis, procedures, medications, clinical trials, and the vital hence this is the best single source widely available in hospitals for doctors and medical staff. Hence, we can conclude that 84.17% of Naïve Bayes is the best-suited algorithm for predicting chronic diabetes disease followed by Support Vector Mechanism with 81.17% and this approach is would be the most efficient model for the Chronic Disease Prediction.

REFERENCES

1. **Artificial intelligence: Google's alphago beats go master lee se-dol.** BBC News, 12 March 2016.
2. Sakr, S.; Elshawi, R.; Ahmed, A.M.; Qureshi, W.T.; Brawner, C.A.; Keteyian, S.J.; Blaha, M.J.; Al-Mallah, M.H., **Comparison of machine learning techniques to predict all-cause mortality using fitness data,** The Henry ford exercise testing (FIT) project. BMC Med. Inform. Decis. Mak. 2017, 17, 174.
3. Waljee AK, Higgins PD., **Machine learning in medicine: a primer for physicians,** Am J Gastroenterol. 2010;105(6):1224-1226. doi:10.1038/ajg.2010.173
4. Kim, S.; Chang, Y.; Yun, K.E.; Jung, H.S.; Lee, S.J.; Shin, H.; Ryu, S., **Development of Nephrolithiasis in Asymptomatic Hyperuricemia: A Cohort Study**, Am. J. Kidney Dis. 2017, 70, 173–181.
5. Sangwoo Lee, Eun Kyung Choe, and Boram Park, **Exploration of Machine Learning for Hyperuricemia Prediction Models Based on Basic Health Checkup Tests,** 2019 2. doi: 10.3390/jcm8020172. PMCID: PMC6406925.
6. Yang Guo, Guohua Bai, and Yan Hu, **Using Bayes Network for Prediction of Type-2 diabetes,** 2012 International Conference for Internet Technology and Secured Transactions, London, 2012, pp. 471-472.
7. Chen, Min et al., **Disease Prediction by Machine Learning Over Big Data From Healthcare Communities,** IEEE Access 5 (2017): 8869-8879.
8. Manish Kumar, **Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm**, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.2, February- 2016, pg. 24-33.
9. Parul singh, Poonam Singh, **Comparative study of Chronic Kidney Disease Prediction using KNN and SVM,** department of Information technology, Barkatullah Institute of Technology, Bhopal, India,
10. Asmaa S. Hussein, Wail M. Omar, Xue Li**, Efficient Chronic Disease Diagnosis Prediction and Recommendation System**, School of IT and Electrical Engineering The University of Queensland Queensland, Australia asmaa.hussein@uqconnect.edu.au, 978-1-4673-1666-8/12/.
11. YanHu Yang Guo Guohua Bai, **Using Bayes Network for Prediction of Type-2 Diabetes,** School of Computing Blekinge Institute of Technology Karlskrona, Sweden School of computing, Sweden.
12. Sahil Sharma, Vinod Sharma, Atul Sharma, **Performance-Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis,** Department of Computer Science & IT University Of Jammu Jammu, India Department Of Internal Medicine Government Medical College Jammu, India.
13. Ms. Soumya.N.S, Mrs. Prabha.R, **Cloud Computing: Data Security Using RSA**, IJLTEMAS, Volume IV, Issue X, October 2015, ISSN 2278 – 2540.
14. Soumya N.S, Divya K. S, Deva Kumari A, Soumya. K, Sreeparna Chakrabarti, **Secure and Multi Copy Dynamic Information Possession in Cloud System**, 'International Journal of Recent Technology and Engineering (IJRTE)', ISSN: 2277-3878 (Online), Volume-8 Issue-5, January 2020. Page No. 4234-4238.
15. Zriqat, Esraa & Altamimi, Ahmad & Azzeh, Mohammad. **A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods**, 2017.
16. Chen, Hung-Chia & Kodell, Ralph & Cheng, kuang-Fu & Chen, James, **Assessment of performance of survival prediction models for cancer prognosis.** 2017 BMC medical research methodology. 12. 102. 10.1186/1471-2288-12-102.
17. Lee, Sangwoo & Choe, Eun & Park, Boram, **An Exploration of Machine Learning for Hyperuricemia Prediction Models Based on Basic Health Checkup Tests. Journal of Clinical Medicine",** 2019. 8. 172. 10.3390/jcm8020172.

18. Hathaway QA, Roth SM, Pinti MV, et al., **Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics,** Cardiovasc Diabetol. 2019;18(1):78. Published 2019 Jun 11. doi:10.1186/s12933-019-0879-0
19. Jathin Desan, **Diagnosis of Brain Hemorrhage Using KNN based Radial Basis Classifier,** International Journal of Advanced Trends in Computer Science and Engineering, 8(6), November - December 2019, 3658– 3664
20. B. Buvaneswari *et al*., **ELSA- A Novel Technique to Predict Parkinson's Disease in Bio-Facial Recognition System**, International Journal of Advanced Trends in Computer Science and Engineering, 8(1), January – February 2019, 12- 17
21. P. S. Ramesh et al., **Analysis of Various Methods for Diagnosing Alzheimer Disease and their Performances**, International Journal of Advanced Trends in Computer Science and Engineering, 8(3), May - June 2019, 755 – 757