International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Predicting Heart Disease using Machine Learning Techniques

Velicheti Srinivas,
Associate Professor, S.D College of Information Technology, Old Town, Tanuku, West Godavari Dist., Andhra Pradesh
velicheti.srinivas469@gmail.com

Abstract:

Cardiovascular disease (CVD)/heart disease has been a common cause of high mortality worldwide over the past decades and has emerged as the most life-threatening disease not only in India but worldwide. Therefore, there is a need for a reliable, accurate and workable system for diagnosing such diseases in time for appropriate treatment. Machine learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Recently, many researchers have adopted several popular machine learning techniques to help the medical industry and professionals diagnose heart disease. This white paper provides an overview of various models based on such algorithms and techniques and analyzes their performance. The models used are based on supervised learning algorithms, Decision Trees (DT), Naive Bayes, K Nearest Neighbors (KNN), Random Forest (RF) and other ensemble models, which are very popular and influential.
Keywords: Diagnosis, Heart Disease, Health Informatics, Machine learning, Prediction.

1. Introduction

The heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it doesn't work properly, the brain and various other organs stop working and within minutes a person dies. Lifestyle changes, work-related stress, and poor diet all contribute to an increase in some heart diseases. Heart disease is one of the leading causes of death worldwide. According to the World Health Organization, heart-related diseases kill 17.7 million people each year, accounting for 31% of all deaths worldwide. Heart disease is also a leading cause of death in India [1]. In 2016, he killed 1.7 million Indians from heart disease, according to the 2016 Global Burden of Disease Report, published on 15 September 2017. Heart-related diseases increase health care costs and reduce personal productivity. According to World Health Organization (WHO) estimates, India lost up to US$237 billion due to heart or cardiovascular disease between 2005 and 2015[2].

Therefore, feasible and accurate prediction of heart disease is of great importance. Healthcare organizations around the world collect data on a variety of health-related topics. This data can be leveraged using various machine learning techniques to generate useful insights. However, the data collected is very large and often this data can be very noisy. These datasets are too overwhelming for the human brain, but can be easily explored using various machine learning techniques. Thus, these algorithms have recently become very useful in accurately predicting the presence or absence of heart disease.

2. Dimensionality Reduction

Dimensionality Reduction includes choosing a mathematical illustration such that you will relate the bulk of, however now no longer all, the variance in the given statistics, thereby together with simplest maximum sizable information. The statistics taken into consideration for a

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

project or a problem, can also additionally includes numerous attributes or dimensions, however now no longer all of those attributes can also additionally similarly affect the output. A big range of attributes, or features, can also additionally have an effect on the computational complexity and can even cause over-becoming which ends up in bad results. Thus, Dimensionality Reduction is a completely crucial step taken into consideration whilst constructing any model. Dimensionality Reduction is commonly executed via way of means of methods -Feature Extraction and Feature Selection.

A. Feature Extraction

The new feature set is derived from the original feature set. Feature extraction involves transforming features. This conversion is often irreversible because most or perhaps a lot of useful information is lost in the process. [3] and [4] use principal component analysis (PCA) for feature extraction. Principal Component Analysis is a widely used linear transformation algorithm. In the feature space, find the direction that maximizes the variance and find the directions that are orthogonal to each other. This is a global algorithm for optimal reconstruction..

B. Feature Selection

A subset of the original feature set has been selected. In [5], important features are selected by a combination of CFS (correlation-based feature selection) subset evaluation and a best-first search method for dimensionality reduction. [6] uses the chi-square statistical test to select the most important features.

3. Algorithms and Techniques Used & Architecture

A. Decision Tree

Decision trees are one of the most popular supervised learning algorithms. This technique is mainly used in classification problems. Easy to use continuous and categorical attributes. This algorithm divides the population into two or more similar sets based on the most important predictor variables. The decision tree algorithm first computes the entropy of each attribute. The dataset is split using the variable or predictor with the highest information gain or lowest entropy. These two steps are performed recursively on the remaining attributes.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

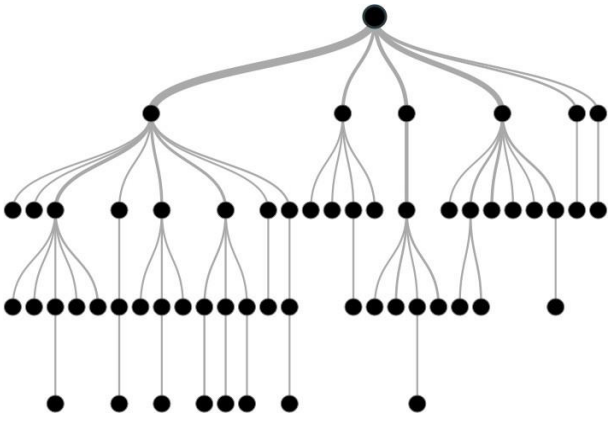


Fig. 1: Structure of a Decision Tree

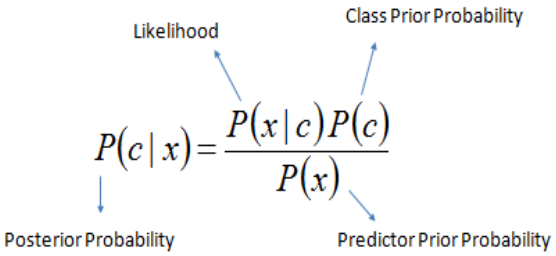
In [10] selection tree has the worst overall performance with an accuracy of 77.55% however while selection tree is used with boosting approach it plays higher with an accuracy of 82.17%.In [9] selection tree plays very poorly with a successfully labeled example percent of 42.8954% while in [14] additionally makes use of the identical dataset however used the J48 set of rules for imposing Decision Trees and the accuracy accordingly acquired is 67.7% that's much less however

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

nevertheless an development at the former. Renu Chauhan et al. have acquired an accuracy of 71.43% [15]. M.A. Jabbar et al. have used alternating selection bushes with precept thing evaluation to attain an accuracy 92.2% [16]. Kamran Farooq et al. have executed the exceptional effects on the usage of selection tree-primarily based totally classifier mixed with ahead choice which achieves a weighted accuracy of 78.4604% [17].

B. Naïve Bayes

Naive Bayes is a easy however an powerful class approach that's primarily based totally at the Bayes Theorem. It assumes independence amongst predictors, i.e., the attributes or functions have to be now no longer correlated to each other or have to now no longer, in anyway, be associated with every different. Even if there's dependency, nevertheless a lot of these functions or attributes independently make a contribution to the chance and this is why it's miles known as Naïve.



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

In [7], Naive Bayes has executed an accuracy of 84.1584% with the ten maximum good sized functions which might be decided on the usage of SVM-RFE (Recursive Feature Elimination) and benefit ratio algorithms while in[8], Naive Bayes has executed an accuracy of 83.49% while all thirteen attributes of the Cleveland dataset[23] are used.

C. K – Nearest Neighbor

In 1951, Hodges et al. delivered a nonparametric approach for sample class that's popularly recognized the K-Nearest Neighbor rule [11]. K-Nearest Neighbor approach is one of the maximum primary however very powerful class strategies. It makes no assumptions approximately the records and is typically be used for class responsibilities while there's very much less or no previous know-how approximately the records distribution. This set of rules includes locating the ok nearest records factors with inside the schooling set to the records factor for which a goal price is unavailable and assigning the common price of the discovered records factors to it. In [10] KNN offers an accuracy of 83.16% while the price of ok is same to nine at the same time as the usage of 10-go validation approach. In [12] KNN with Ant Colony Optimization plays higher than different strategies with an accuracy of 70.26% and the mistake fees is 0.526.Ridhi Saini et al. have acquired a performance of 87.5% [13], which could be very good.

D. Random Forest

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Random Forest is likewise a popularly supervised device studying set of rules. This approach may be used for each regression and class responsibilities however typically plays higher in class responsibilities. As the call suggests, Random Forest approach considers a couple of selection bushes earlier than giving an output. So, it is largely an ensemble of selection bushes. This approach is primarily based totally at the perception that greater variety of bushes could converge to the proper selection. For class, it makes use of a vote casting gadget after which makes a decision the elegance while in regression it takes the imply of all of the outputs of every of the selection bushes. It works properly with massive datasets with excessive dimensionality.

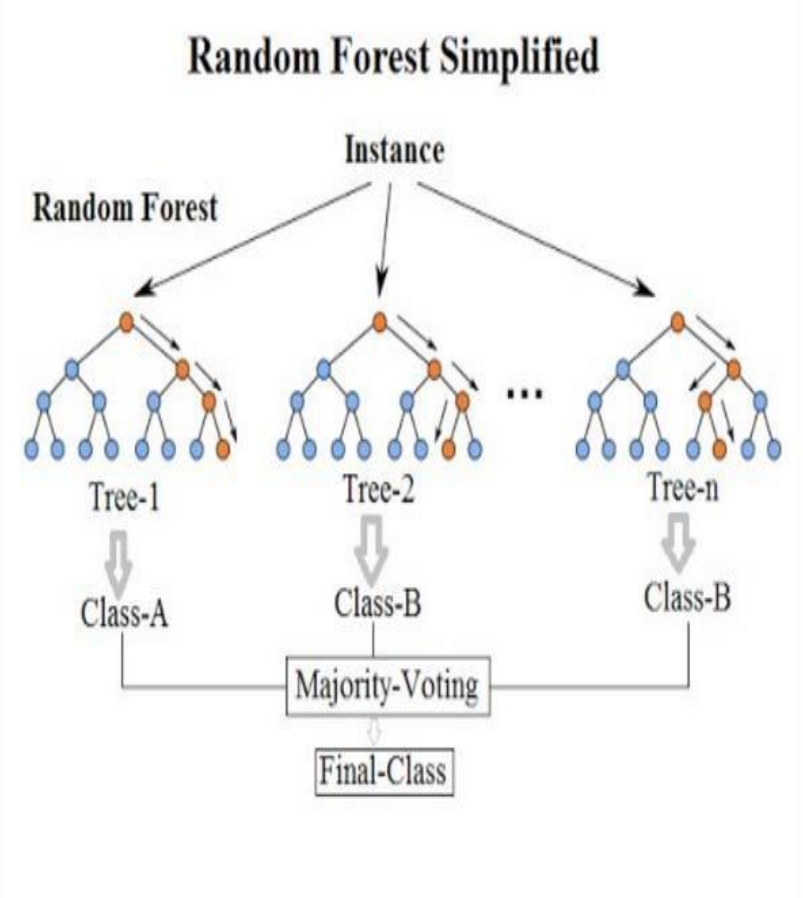


Fig. 2: Random Forest

In [5], random woodland plays distinctly properly. In Cleveland dataset, random woodland has a drastically better accuracy of 91.6% than all of the different methods. In People's Hospital dataset, it achieves an accuracy of 97%. In [18] [26-32] random woodland has executed an f-degree of 0.86. In [19], random woodland is used to are expecting coronary heart disorder and it obtains an accuracy of 97.7%.

E. Ensemble Model

In ensemble modeling or greater associated however distinctive analytical fashions are used and bring their outcomes are blended right into an unmarried score.

Tahira Mahboob et al. [20] have used an ensemble of KNN and ANN to gain an accuracy of 94.12%. The Majority vote-primarily based totally version as established with the aid of using Saba Bashir et al. [21] which incorporates of Naïve Bayes, Decision Tree and Support Vector Machine classifiers, gave an

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

accuracy of 82%, sensitivity of 74% and specificity of 93% for UCI coronary heart sickness dataset. In [22-25] an ensemble version, which includes Gini Index, Naïve Bayes classifiers, has been proposed which gave an accuracy of 98% in predicting Syncope disease.

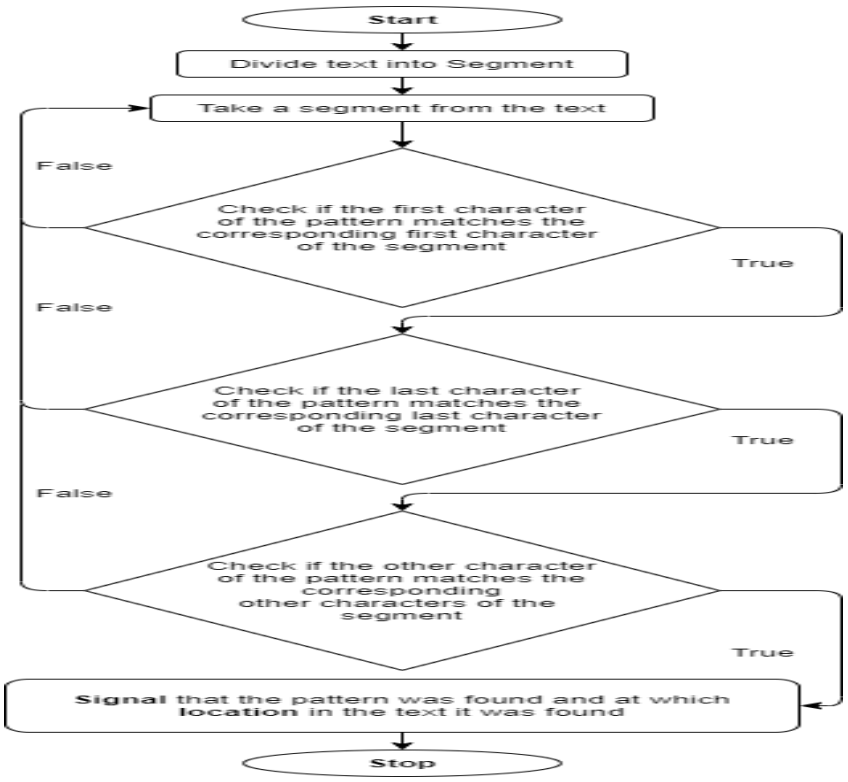


Fig. 3: Algorithm Flow Diagram

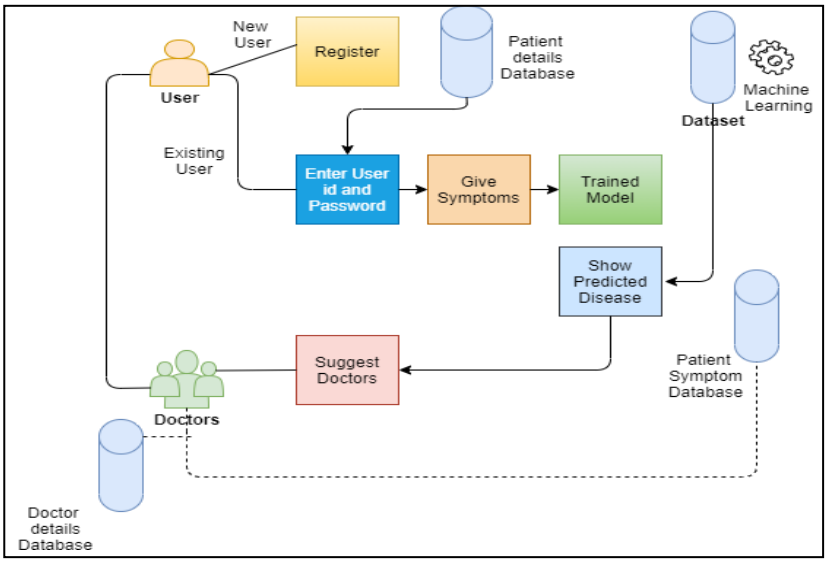


Fig. 4: General Application Architecture

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4. Conclusion

Based at the above review, it is able to be concluded that there's a big scope for device mastering algorithms in predicting cardiovascular illnesses or coronary heart associated illnesses. Each of the above-referred to algorithms has finished extraordinarily nicely in a few instances however poorly in a few different instances. Alternating choice timber whilst used with PCA, have finished extraordinarily nicely however choice timber have finished very poorly in a few different instances which may be because of over-becoming. Random Forest and Ensemble fashions have finished thoroughly due to the fact they resolve the trouble of over-becoming with the aid of using more than one algorithm (more than one Decision Trees in case of Random Forest). Models primarily based totally on Naïve Bayes classifier have been computationally very rapid and feature additionally finished nicely. Systems primarily based totally on device mastering algorithms and strategies were very correct in predicting the coronary heart associated illnesses however nonetheless there is lots scope of studies to be executed on a way to manage excessive dimensional records and over-becoming. A lot of studies also can be executed on the ideal ensemble of algorithms to apply for a selected form of records.

References

[1] Ramadoss and Shah B et al.“A. Responding to the threat of chronic diseases in India”. Lancet. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.

[2] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011

[3] Dhomse Kanchan B and Mahale Kishor M. et al. “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis”, 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.

[4] R.Kavitha and E.Kannan et al. “An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining “, 2016

[5] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework”, 2017 IEEE 2nd International Conference on Big Data Analysis.

[6] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis and Vijay K. Mago et al. “ Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map”, 978-1-5090-0626-7/16/$31.00 c 2016 IEEE.

[7] Kanika Pahwa and Ravinder Kumar et al. “Prediction of Heart Disease Using Hybrid Technique For Selecting Features”, 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).

[8] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. “ A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease”, 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017

[9] Hanen Bouali and Jalel Akaichi et al. “Comparative study of Different classification techniques, heart Diseases use Case.”, 2014 13th International Conference on Machine Learning and Applications

[10] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez et al. “ A Comprehensive Investigation and Comparison of Machine Learning

Techniques in the Domain of Heart Disease", 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops - ICTS4eHealth 2017

[11] J. Hodges et al. "Discriminatory analysis, nonparametric discrimination: Consistency properties," 1981.

[12] S.Rajathi and Dr.G.Radhamani et al. "Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO ", 2016.

[13] Puneet Bansal and Ridhi Saini et al. "Classification of heart diseases from ECG signals using wavelet transform and kNN classifier", International Conference on Computing, Communication and Automation (ICCCA2015).

[14] Simge EKIZ and Pakize Erdogmus et al. "Comparitive Study of heart Disease Classification", 978-1-5386-0440-3/17/$31.00 ©2017 IEEE.

[15] Renu Chauhan, Pinki Bajaj, Kavita Choudhary and Yogita Gigras et al. "Framework to Predict Health Diseases Using Attribute Selection Mechanism", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIA Com).

[16] M.A.JABBAR , B.L Deekshatulu and Priti Chndra et al. "Alternating decision trees for early diagnosis of heart disease", Proceedings of International Conference on Circuits, Communication, Control and Computing (I4C 2014).

[17] Amir Hussain, Peipei Yang, Mufti Mahmud and Jan Karasek et al. "A Novel Cardiovascular Decision Support Framework for effective clinical Risk Assessment.", 978-1-4799-4527-6/14/$31.00 ©2014 IEEE.

[18] Quazi Abidur Rahman, Larisa G. Tereshchenko, Matthew Kongkatong, Theodore Abraham, M. Roselle Abraham, and Hagit Shatkay et al. "Utilizing ECG-based Heartbeat Classification for Hypertrophic Cardiomyopathy Identification", DOI 10.1109/TNB.2015.2426213, IEEE Transactions on Nano Bioscience TNB-00035-2015.

[19] Ahmad Shahin, Walid Moudani, Fadi Chakik, Mohamad Khalil et al. "Data Mining in Healthcare Information Systems: Case Studies in Northern Lebanon", ISBN: 978-1-4799-3166-8 ©2014 IEEE.

[20] Tahira Mahboob, Rida Irfan and Bazelah Ghaffar et al. "Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics", 978-1-5090-4815-1/17/$31.00 ©2017 IEEE.

[21] Saba Bashir, Usman Qamar, M.Younus Javed et al. "An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis" International Conference on Information Society (i- Society 2014).

[22] Ammar Asjad Raja, Irfan-ul-Haq , Madiha Guftar Tamim Ahmed Khan and Dominik Greibl et al. "Intelligent Syncope Disease Prediction Framework using DM-Ensemble Techniques", FTC 2016 - Future Technologies Conference 2016.

[23] CI Education, Heart Disease Data Set [OL]. http://archive.ics.uci.edu/ml/datasets/Heart+Disease CHDD.

[24] T. Padmapriya and V.Saminadan, "Handoff Decision for Multi- user Multiclass Traffic in MIMO-LTE-A Networks", 2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016) – Elsevier - PROCEDIA OF COMPUTER SCIENCE, vol. 92, pp: 410-417, August 2016.

[25] S.V.Manikanthan and D.Sugandhi "Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel" International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume- 7, Issue 1 –MARCH 2014.

[26] Sri Hari Nallamala, Dr. Suvrnavani Koneru, "An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records", (IJET) (UAE), Vol. 7, No 2.7, SI7, ISSN: 2227 – 524X,

2018.

[27] Sri Hari Nallamala, Dr. Suvrnavani Koneru, "A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment", (IJET) (UAE), Vol. 7, No 2.7, SI7, ISSN: 2227 – 524X, 2018.

[28] Sri Hari Nallamala, Dr. Pragnyaban Mishra, Dr. Suvrnavani Koneru, International Journal of Advanced Trends in Computer Science and Engineering, "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", IJATCSE, ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, 2019.

[29] Sri Hari Nallamala Dr. Pragnyaban Mishra, Dr. Suvrnavani Koneru, International Journal of Recent Technology and Engineering "Breast Cancer Detection using Machine Learning Way", IJRTE, ISSN: 2277-3878, Vol-8, Is-2S3, July, 2019.

[30] Sri Hari Nallamala Dr. Pragnyaban Mishra, Dr. Suvrnavani Koneru, "Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment", (IJSTR) International Journal of Scientific and Technology Research, ISSN: 2277-8616, Vol. 8, Issue 11, November, 2019.

[31] Sri Hari Nallamala, Dr. D. Durga Prasad, Ranga Rajesh Jallipalli, Dr. Pragnaban Mishra, Sushma Chowdary Polavarapu, A Review on "Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management", TEST Engineering & Management, ISSN: 0193-4120, Vol.83. No.3, May-June 2020.

[32] Sushma Chowdary Polavarapu, Umamaheswari K, Sri Hari Nallamala, "RFID based automatic tollgate collection", (IJET) (UAE), Vol. 7, No 2.1, 2018, ISSN: 2227 – 524X, 2018.