Prediction and Analysis of Multiple Diseases Using Machine Learning Techniques

Abhishek Singh[1], Dr. SPS Chauhan[2]

[1]PG Student, Galgotias University, Greater Noida, India.
Assistant Professor CSE, Galgotias University, Greater Noida, India.

---***---

Abstract— *Taking health issue seriously in this busy lifestyle is now the most essential work for an individual and for the medical science. As we all know, medical technology nowadays has rapidly developed its growth to address all of the people's health and medical problems around the globe with all new features and technology improving day-by-day. It proves that in this era of medical science it can save time and money for any illness or treatment. Medical science now has all the possible results for the diagnosis of many diseases to evaluate and forecast diseases with the aid of a medical database of an individual. This is a smart card method based on smart analytical methods to predict and analyses all future and previous diseases with the help of KNN, Support Vector Machine, Random Forest, Decision Tree algorithms. With the help of this card people can move forward to better and innovative identification of their birth-related medical issues and health records. This helps you to connect with your doctor by scanning the details on your card in an easy way. This card is attachable to other government provided identities to make sure about patient's details and also provides all of the basic information at the time of emergency.*

Keywords— Machine Learning, KNN (K-Nearest Neighbors), Support Vector Machine, Random Forest, Decision Tree.

1. INTRODUCTION

Machine learning has been influential in finding hidden trends in large datasets in the world of medical science. As we know, medical science in the form of patient's information and medical history generates huge amounts of database. In this paper, we analyzed the different techniques of Machine Learning strategies to research and evaluate different disease behavior and forecast based on time and database.

The main motivation of this research is to provide better framework for an easy and accurate diagnosis of the disease. When doctors have to perform different number of tests manually to integrate all of the patient's necessary reports requiring a lot of resources, energy, commitment to treat and compile all the reports according to the period. Providing card as a work around for transporting an individual's entire

Medical record, we can save time for the patient and doctor to figure out and evaluate our current situation. Today machine learning is accessible anywhere, so that we can use it according to our dataset and requirement. People tend to neglect some of the signs of illness going on in their daily activities. The moment when it was still conscious of the epidemic has moved to a greater level of complexity. Also because of the strong likelihood of medical check-ups people usually ignore these things with time to carry on. The rising health problems in humans raise awareness of different methodologies that can be useful in predicting different disease analyses. By deploying machine learning algorithms we can use previous and ongoing database to predict future outcome illnesses. According to the 2014 study from the World Health Organization, cardiovascular diseases like heart attacks, persistent heart failure, cardiac arrest, etc. culminated in 17.5 million people being killed. Diseases in various regions that vary depending on the environment and living conditions of people with different lifestyles. Healthcare data today is large and so vast that it can be processed and stored for further prediction and analysis via Machine Learning and Big Data. Each parent wants to be safe and successful to their children or to future generations. We will do this by forecasting on the basis of possible result illnesses with the parents ' genetic dataset and further research. In brief, with fewer diseases, we can foresee the future generations and can work for better succession and generation growth. So in this paper we are not only analyzing and predicting diseases but also giving a new platform to store all the medical database and use it according to need. We are taking machine learning algorithms to determine and classify patient's medical records to examine and predict further.

So we're not only analyzing and predicting diseases in this paper but also providing a new platform to store and access all the medical database of patient's history according to need. For further investigation and forecasting, we take Machine Learning algorithms to access and identify medical patterns of patients.

2. Literature Survey

Analysis and prediction of future behavior with datasets is helping in various resources including human health. Human resource is one of the major and never lasting components on this planet. As we are growing digitally in every sector continuously though we are also growing in medical facilities and database of medical records. There are specific number of research have been done by analyzing and predicting the future diseases on the basic of datasets provided by patient's health records. By further records and work on medical technology we concluded some of the best methods for prediction and analysis.

[1][Deepika Verma 2017] concluded that they considered WEKA tool as one of the best tool for data mining classification. According to them J48 gives 74.28%accuracy rate on breast cancer dataset and SMO gives 76.80% accuracy rate on diabetes dataset.
[2][Saba Bashir, Zain Sikander Khan 2019] they performed some selection process on various algorithms and techniques to improve accuracy of different algorithms. After performing feature selection they get Decision Tree is 82.22%, Logistic Regression 82.56%, Random Forest 84.17%, and Naïve Bayes 84.24% and Logistic Regression SVM is 84.85.
[3][Prerna Jain, Amandeep Kaur 2018] they undertaken analysis of big data for prediction of coronary heart diseases using k-nearest and genetic algorithm. They concluded that from this given system they are able to predict diseases at early stage and this method helps to improve medical field for physician to analyze and predict easily.
[14][Poojitha Amin2019] they build a prediction model that uses data from wearable devices to perform diagnosis on cardiovascular problems. They obtained 87% accuracy with Logistic Regression Model. Additionally AUC score of 80% is obtained by measuring under ROC curve.
[15][Ch Anwar ul Hassan, Muhammad Sufiyan Khan, Munam Ali Shah 2018] they performed various comparison analysis of ML classifiers for prediction of heart and hepatitis diseases. They took six different classifiers including LR, DT, NB, KNN, SVM and RF on same database. After initializing all the algorithm on same dataset they observed that RF performed better and provide better prediction than all other classifiers.

TABLE -1. RELATED WORK ON VARIOUS CLASIFIER ACCORDING TO THE DISEASES AND YEAR OF IMPLEMENTATION.

REFERENCES	DISEASE	CLASSIFIER	YEAR
[1]	Breast Cancer, Diabetes	WEKA tool, J48, SMO	2017
[2]	Heart	DT , SVM , NB	2019
[3]	Coronary Artery	KNN , Genetic algorithm	2018
[4]	Multiple diseases	Genetics , naïve Bayes, Decision Tree	2017
[5]	Diabetes	Naive Bayes, SVM, Random Forest, Simple CART	2018
[6]	Breast Cancer	Naive Bayes, Bayesian Logistic Regression, Simple CART, J48	2018
[7]	Liver	KNN, Logistic Regression, SVM	2018
[8]	Chronic Disease	Naïve Bayes, KNN, Decision tree, SVM, RNN	2019
[9]	Diabetes	Random Forest	2019
[10]	Heart	Random Forest, Decision Tree, KNN	2019
[11]	Heart	Decision Tree, Random Forest, Neural Network	2019
[12]	Heart	Regression Model, Random Forest, Neural Networks	2019
[13]	Infectious diseases	Non- Invasive Logistic Regression Technique	2019
[14]	Heart	Logistic Regression, Gradient Boosting, Random Forest	2019

TABLE -2. RELATED WORK ON VARIOUS APPROACHES OF ANALYSIS AND PREDICTING FUTURE DISEASES

S NO	PAPER	OBJECTIVE	CONCLUSION
1	Analysis and Prediction of Breast cancer and Diabetes disease via Data mining classification Techniques.	They calculated all accuracy measures include WEKA tool for classification and prediction of breast cancer and diabetes.	J48 gives 74.28%accurate results and SMO gives 76.80% accurate results on diabetes dataset
2	Improving Heart Disease Prediction.	This research focuses on selection of techniques and algorithms for which multiple heart disease datasets are used for experimentation analysis.	The accuracy of Decision Tree is 82.22%, Logistic Regression 82.56%, Random Forest 84.17%, and Naïve Bayes 84.24% and Logistic Regression SVM is 84.85.
3	Designing Disease Prediction Model via Machine Learning.	This research proposed general disease prediction based on symptoms of the patient by using KNN and CNN.	Analyses and prediction system based on machine learning algorithm. Utilized that KNN and CNN algorithms to classify patient data because today medical data growing very vastly and that needs to process existed data for predicting exact disease based on symptoms.
4	Big Data Analysis for Prediction of Disease like Coronary Artery.	The analysis of huge amount of a patient by using data mining and machine learning algorithms improves a hospital administration. As a huge amount of the data in increase in every field, so it is difficult to analysis, extract, manage and store a structured and unstructured data, so that the big data technologies and tools are used.	A data set of patient is collected from different sources, then pre-process and predicts a heart disease based on symptoms and risk factors. We can conclude that through this predicted system we are able to predict any disease at early stage and this system is able to be use in medical field for physician to easily analysis and prediction of particular heart disease.
5	Smart Analytics And Predictions For Indian Medicare.	It comprises of a prediction system for predicting the diseases and disorders which use the concepts of machine learning and artificial neural networks & also provides online upload or download of medical reports. It also provides various functionalities to Indian citizens, who can get continuous notifications regarding dosage timings as prescribed by their doctors.	It comprises of a prediction system for predicting the diseases and disorders which use the concepts of machine learning and artificial neural networks & also provides online upload or download of medical reports. It also provides various functionalities to Indian citizens, who can get continuous notifications regarding dosage timings as prescribed by their doctors.
6	DATA MINING AND VISUALISATION FOR PREDICTION OF MULTIPLE DISEASES	In this paper, data mining methods namely Naïve Bayes and J48 algorithm are compared for testing their accuracy and performance on the training medical datasets.	The conclusion derived from the evaluation results of J48 and Naïve Bayes, it was observed that Naïve Bayes results is the best accuracy of prediction and even outclasses J48 in the latency analysis on the datasets.
7	Diabetes Disease Prediction using Machine Learning	The research hopes to recommend the best algorithm based on efficient performance efficient performance result for the prediction of diabetes disease. Experimental results of each algorithm used on the dataset was evaluated. It is observed that Support	The overall performance of Support Vector machine to predict the diabetes disease is better than Naïve Bayes, Random Forest and Simple Cart. Hence the effectiveness of the proposed model is

		Vector Machine performed best in prediction of the disease having maximum accuracy.	clearly depicted throughout the experimental results mentioned.
8	Using Data Mining Tools for Breast Cancer	The main goal is to classify data of both the algorithms in terms of accuracy. Their experimental result shows that among all the classifiers, decision tree classifier i.e. Simple CART (98.13%) gives higher accuracy.	The results obtained are compared and is found that Simple CART decision tree algorithm is the best classifier in terms of accuracy among all the classifiers used here in the research. The time complexity of simple CART is more. So in future we can work on that so more accurate results can be produced in less time.
9	Prediction of Liver Disease	The main aim is to predict liver disease using different classification algorithms. The algorithms used for this purpose of work is Logistic Regression, K-Nearest Neighbors and Support Vector Machines. Accuracy score and confusion matrix is used to compare this classification algorithm.	The main aim is to predict liver disease using different classification algorithms. The algorithms used for this purpose of work is Logistic Regression, K-Nearest Neighbors and Support Vector Machines. Accuracy score and confusion matrix is used to compare this classification algorithm.
10	High Quality Crowdsourcing Clinical Data For Prediction Of Diseases	They present how carefully chosen clinical features with our proper data cleaning method improves the accuracy of the Amyotrophic Lateral Sclerosis (ALS) disease progression and survival rate predictions. In addition, they present an incentive model which provides individual rationality and platform profitability features to encourage hospitals to share high quality data for such predictions.	The experimental results indicate that our proposed methods achieve good performance in ALS slope and survival predictions. As for future work, although the results indicate that a machine learning-based predictive model generated using existing patient records could aid in clinical care

3. CONCLUSION AND FUTURE WORK

In future these techniques can be applied on real time database of individual patient and can determine prediction of multiple diseases. The main propose of this paper is to determine a future work on medical database with the help of digital card to store information and analyze. This research paper focuses on analyzing and designing a system where patients ' real-time information can be processed and evaluated based on previous symptoms and on current symptoms for different diseases. By this paper we have concluded that KNN, Support Vector, Random Forest, Decision tree are the best algorithms with higher accuracy rate than others for predicting and analysis. So in future we can continue this paper by implementing these algorithms for better results and working model. This paper also outlines the technique to deploy this method to android and web platform to analyze and predict using real time data of users by collaborating with doctors and various medical organizations.

4. REFRENCES

[1] Deepika Verma AND Dr. Nidhi Mishra “Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques” ©2017 IEEE.

[2] Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, Khurram Bashir “Improving Heart Disease Prediction Using Feature Selection Approaches” Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 8th – 12th January, 2019.

[3] Prerna Jain , Amandeep Kaur “ Big Data Analysis for Prediction of Coronary Artery Disease” ©2018 IEEE.

[4] Anjinkya Kunjir , Harshal Sawant, Nuzhat F.sheikh “Data Mining and Visualization for Prediction of Multiple Diseases in Healthcare” ©2017 IEEE.

[5]Ayman Mir , Sudhir N. Dhage "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare" ©2018 IEEE.

[6]Dr. S. N. Singh , Shivani Thakral "Using Data Mining Tools for Breast Cancer Prediction and Analysis" ©2018 IEEE.

[7] Thirunavukkarasu K. , Ajay S. Singh , Md Irfan , Abhishek Chowdhury "Prediction of Liver Disease using Classification Algorithms" 2018 4th International Conference on Computing Communication and Automation (ICCCA) 978-1-5386-6947-1/18/$31.00 ©2018 IEEE.

[8] Arvindkumar.s ,Arun.P, Ajith.A ,"Prediction of Chronic Disease by Machine Learning' @IEEE 978-1-2-7281-1524-5.

[9] K. Vijiya Kumar ,B .Lavanya ,I.Nirmala , S.Sofia Carroline ,"Random forest Algorithm For The Prediction of Diabetes "IEEE 978-1-7281-1524-5.

[10] Divya Krishnani ,Akash Dewangan, Aditya Siingh ,Nenavath srrinivas Naik ,"Prediction of coronary Heart Disease Using Superviseed Machine Learning Algorithms" 978-1-7281-1895-6/19/$31.00©2019IEEE.

[11] Chidozie Shamrock Nwosu ,Soumyabrata Dev ,Peru Bhardwaj ,Bharadwaj Veeravalli ,Deepu John,"Preedicting Stroke From Electronic Health Records" 978-1-5386-1311-5/19-$31.00©2019 IEEE.

[12] Adela vrtkova ,Vsclsv Prrochazka ," Comparing The Performance of Rreegrression Models, random Foreests and Neural Networks for Stroke Patients ' Outcome Prediction "978-7281-1401-9/19-$31.00©2019 IEEE.

[13] Mr. Anirudhh Ravi ,Mr. Varun Gopal ,Dr. J. Prreetha Roselyn ," Detection of Infection Disease Using Non- Invasive Logistic Regression Technique "978-1-5386-9543-2/$31.00©2019 IEEEE.

[14] Poojitha Amin ,Nikhitha R .Anikireddypally ,Suraj Khurana ,"personalized Health Monitoring Using Predictive Analytics"978-72881-0059-3/19/$31.00©2019 IEEE .DI 10.1109?BigDataServices.2019.00048

[15] Chidozie Shamrock Nwosu, Soumyabrata Dev, Peru Bhardwaj, Bharadwaj Veeravalli, and Deepu John ," Predicting Stroke from Electronic Health Record" 978-1-5386-1311-5/19/$31.00 ©2019 IEEE.

[16] Ch Anwar ul Hassan, Muhammad Sufyan Khan, Munam Ali Shah," Comparison of Machine Learning Algorithms in Data classification"

[17] Adela vrtkova ,Vsclsv Prrochazka ," Comparing The Performance of Rreegrression Models, random Foreests and Neural Networks for Stroke Patients ' Outcome Prediction "978-7281-1401-9/19-$31.00©2019 IEEE.

[18] Mr. Anirudhh Ravi ,Mr. Varun Gopal ,Dr. J. Preetha Roselyn ," Detection of Infection Disease Using Non- Invasive Logistic Regression Technique "978-1-5386-9543-2/$31.00©2019 IEEEE

[19] Poojitha Amin ,Nikhitha R .Anikireddypally ,Suraj Khurana ,"personalized Health Monitoring Using Predictive Analytics"978-72881-0059-3/19/$31.00©2019 IEEE .DI 10.1109?BigDataServices.2019.00048

[20] Deepika Verma (Author), Dr. Nidhi Mishra (Author)," Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques "978-1-5386-1959-9/17/$31.00 ©2017 IEEE.

[21]] Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, Khurram Bashir "Improving Heart Disease Prediction Using Feature Selection Approaches" Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 8th – 12th January, 2019.

[22] Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. Ektaa Meshram," Designing Disease Prediction Model Using Machine Learning Approach" 978-1-5386-7808-4/19/$31.00 ©2019 IEEE.

[23] Prerna Jain , Amandeep Kaur " Big Data Analysis for Prediction of Coronary Artery Disease" ©2018 IEEE.

[24] Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. Ektaa Meshram "Designing Disease Prediction Model Using Machine Learning Approach" ©2019 IEEE.

[25] Anjinkya Kunjir , Harshal Sawant, Nuzhat F.sheikh "Data Mining and Visualization for Prediction of Multiple Diseases in Healthcare" ©2017 IEEE.

[26] Ayman Mir , Sudhir N. Dhage "Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare" ©2018 IEEE.

[27] Deepika Verma AND Dr. Nidhi Mishra "Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques" ©2017 IEEE.

[28] Thirunavukkarasu K, Ajay S. Singh* Md Irfan ,Abhishek Chowdhury#,"Prediction of Liver Disease using Classification Algorithms" 978-1-5386-6947-1/18/$31.00 ©2018 IEEE.

[29] Xie Shuang, Fan Huimin," Research on CNN to Feature Extraction in Diseases Prediction" 978-1-7281-3977-7/19/$31.00 ©2019 IEEE DOI 10.1109/ICCNEA.2019.00046.

[30] V.Nandhini, Dr.M.S.Geetha Devasena Professor," Predictive Analytics for Climate Change Detection and Disease Diagnosis" 978-1-5386-9533-3/19/$31.00 ©2019 IEEE.

[31] Arvindkumar.s ,Arun.P, Ajith.A ,"Prediction of Chronic Disease by Machine Learning' @IEEE 978-1-2-7281-1524-5

[32] K. Vijiya Kumar ,B .Lavanya ,I.Nirmala , S.Sofia Carroline ,"Random forest Algorithm For The Prediction of Diabetes "IEEE 978-1-7281-1524-5.

[33] Divya Krishnani ,Akash Dewangan, Aditya Siingh ,Nenavath srrinivas Naik ,"Prediction of coronary Heart Disease Using Supervised Machine Learning Algorithms" 978-1-7281-1895-6/19/$31.00©2019IEEE.