International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume- 8, Issue- 4, July- 2020
https://doi.org/10.21276/ijircst.2020.8.4.14
www.ijircst.org

A Detailed Review on Disease Prediction Models that uses Machine Learning

Md. Ehtisham Farooqui, Dr. Jameel Ahmad

ABSTRACT- Human body is guarded by the immune system, but sometimes this immune system alone is not capable of preventing our body from diseases. Environmental conditions and living habits of people are the cause of many diseases that are the main reason for a huge number of deaths in the world, and diagnosing these diseases sometimes becomes challenging. We need an accurate, feasible, reliable, and robust system to diagnose diseases in time so that these can be properly treated. With the growth of medical data, many researchers are using these medical data and some machine learning algorithms to help the healthcare communities in the diagnosis of many diseases. In this paper a survey of various models based on such algorithms, techniques is presented and their performance is analyzed. Researches have been conducted on various models of supervised learning algorithms and some of them are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Naïve Bayes and Random Forest (RF).

KEYWORDS- Decision Tree, Machine Learning, Naïve Bayes, Random Forest

I. INTRODUCTION

According to McKinsey's report, 50% of Americans have multiple chronic diseases. Due to living habits that people have, the chances of chronic disease is increasing [1]. In India, as the lifestyle of people has improved, the frequency of diseases is also increased. Approx. 60% of death has happened due to non-communicable diseases like heart disorder, cancer, and diabetes. These diseases are often caused by environmental conditions and living habits that people have [2].

Continuous growth in medical data gave us a way to extract the required information to predict the disease. Data Science and Big Data can be applied to detect various types of diseases by using past health data collected from the patient. These disease prediction models are very important to know the presence of disease [3].

For the detection of the diseases we require machine learning techniques like supervised, semi-supervised, unsupervised learning, etc. and raw medical data [4]. This raw data could easily obtained from famous government hospitals. Machine learning techniques can use the raw data for the learning process and based on that learning they can predict the disease later.

This paper contains information about, what machine learning techniques other researchers have used and what the accuracy of their proposed system was.

II. LITERATURE REVIEW

There have been numerous studies done related to predicting the disease using different machine learning techniques and algorithms which can be used by medical institutions. This paper reviews some of those studies done in research papers using the techniques and results used by them. Reviews are given below:

A. Reviews

MIN CHEN et al, [1] proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, and Decision Tree. This proposed system had an accuracy of 94.8%.

Sayali Ambekar et al, [2] recommended Disease Risk Prediction and used a convolution neural network to perform the task. In this paper machine learning techniques like CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are used. The system uses structured data to be trained and its accuracy reaches 82% and achieved by using Naïve Bayes.

Naganna Chetty et al, [3] developed a system that gives improved results for disease prediction and used a fuzzy approach. And used techniques like KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier. In this paper diabetes disease and liver, disorder prediction is done and the accuracy of Diabetes is 97.02% and Liver disorder is 96.13.

Dhiraj Dahiwade et al, [4] designed a model for prediction of the disease using approaches of machine learning and used techniques like KNN and CNN. This

Manuscript received July 22, 2020

Md. Ehtisham Farooqui, Student, Department of Computer Science and Engineering, Integral University, Lucknow, India, (e-mail: mefxe01@gmail.com)

Dr. Jameel Ahmad, Assistant Professor, Department of Computer Science and Engineering, Integral University, Lucknow, India.

paper suggests disease prediction i.e. based on patient's symptoms. The accuracy of KNN is 95% and the accuracy of CNN is 98%.

Lambodar Jena et al, [5] focused on risk prediction for chronic diseases by taking advantage of distributed machine learning classifiers and used techniques like Naive Bayes and Multilayer Perceptron. This paper tries to predict Chronic-Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively.

Dhomse Kanchan B. et al, [6] studied special disease prediction utilizing principal component analysis using machine learning algorithms involving techniques like Naive Bayes classification, Decision Tree, and Support Vector Machine. The accuracy of this system is 34.89% for Diabetes and 53% for Heart disease.

Pahulpreet Singh Kohli et al, [7] suggested disease prediction by using applications and methods of machine learning and used techniques like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for Heart disease.

Deeraj Shetty et al, [8] studied the uses of data mining for diabetes disease prediction by using Naïve Bayes and KNN algorithms. This system predicts diabetes and accuracy obtained by KNN are better than Naïve Bayes.

Rashmi G Saboji et al, [9] tried to find a scalable solution that can predict heart disease utilizing classification mining and used Random Forest Algorithm. This system presents a comparison against Naïve-Bayes classifier but Random Forest gives more accurate results with accuracy 98%.

Rati Shukla et al, [10] suggested prediction and detection for breast cancer by utilizing machine learning techniques like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network, and KNN. In this system, the Support Vector Machine gives more accurate results than all other algorithms.

Senthilkumar Mohan et al, [11] focused on hybrid techniques in machine learning that can be used for effectively predicting heart disease and used algorithms like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN. The accuracy of this system is 88.47%.

Anjan Nikhil Repaka et al, [12] designed and implemented a prediction model for heart disease using naive Bayesian. Any user can use this system using any smartphone device and get the prediction results. The accuracy of this system is 89.77%.

Aakash Chauhan et al, [13] proposed a disease prediction model for heart disease by utilizing evolutionary rule learning. Association Rule is used in this proposed system. This system is not very efficient because it has an accuracy of 53%.

Aditi Gavhane et al, [14] suggested prediction for heart disease that utilizes Machine Learning. Multi-Layer Perceptron model is used in this system. This system predicts heart disease based on basic symptoms like age, sex, pulse rate, etc. The accuracy of this suggested system is 91%.

Ankita Dewan et al, [15] recommended a disease prediction system that uses data mining classification hybrid technique for predicting heart disease. This system is using techniques like Neural Network, Decision Tree, and Naive Bayes. The accuracy of this system is 87%. .

B. A comparative study using various algorithms in the literature review

Table 1: Comparative study using various algorithms in the literature review

Year	Author	Purpose	Techniques Used	Accuracy
2017	MIN CHEN et al, [1]	Proposed a disease prediction system in his paper where he used machine learning algorithms.	CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, Decision Tree	94.8%
2018	**S**ayali Ambekar et al, [2]	Recommended Disease Risk Prediction and used a convolution neural network to perform the task	CNN-UDRP algorithm, Naive Bayes and KNN algorithm	The highest accuracy of 82% is achieved by Naïve Bayes.
2015	Nagana Chetty et al, [3]	Developed a system that gives improved results for disease prediction and used a fuzzy approach	KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier	Diabetes: 97.02 %
				Liver disorder: 96.13 %
2019	Dhiraj Dahiwade et al, [4]	Designed a model for prediction of the disease using approaches of machine learning	K-Nearest neighbor (KNN) and Convolutional neural network (CNN)	KNN: 95%
				CNN: 98%
2017	Lambodar Jena et al, [5]	Focused on risk prediction for chronic diseases by taking advantage of	Naive Bayes	95%
			Multilayer Perceptron	99.7%

		distributed machine learning classifiers		
2016	Dhomse Kanchan B. et al, [6]	Studied special disease prediction utilizing principal component analysis using machine learning algorithms	Naive Bayes classification, Decision Tree and Support Vector Machine	Diabetes Disease: 34.89 % Heart Disease: 53%
2018	Pahulpreet Singh Kohli et al, [7]	Suggested disease prediction by using applications and methods of machine learning	Logistic Regression	Breast Cancer: 95.71 % Diabetes: 84.42 % Heart Disease: 87.12 %
			Decision Tree	Breast Cancer: 94.29 % Diabetes: 74.03 % Heart Disease: 70.97 %
			Random Forest	Breast Cancer: 97.14 % Diabetes: 81.82 % Heart Disease: 77.42 %
			Support Vector Machine	Breast Cancer: 97.14 % Diabetes: 85.71 % Heart Disease: 83,87 %
			Adaptive Boosting	Breast Cancer: 98.57 % Diabetes: 80.52 % Heart Disease: 83.87 %
2017	Deeraj Shetty et al, [8]	Studied the uses of data mining for diabetes disease prediction	Naïve Bayes and KNN	KNN gives better accuracy, compared to Naïve Bayes.
2017	Rashmi G Saboji et al, [9]	Tried to find a scalable solution that can predict heart disease utilizing classification mining	Random Forest Algorithm	98%
2019	Rati Shukla et al, [10]	Suggested prediction and detection for breast cancer by utilizing machine learning	Naive Bayes Classifier, Logistic Regression, Support Vector Machines	SVM provides a more accurate result compar

		techniques	(SVM), Artificial Neural Networks and K-Nearest Neighbor	ed to others.
2019	Senthilkumar Mohan et al, [11]	Focused on hybrid techniques in machine learning that can be used for effectively predicting heart disease	Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN	88.4%
2019	Anjan Nikhil Repaka et al, [12]	Designed and implemented a prediction model for heart disease using naive Bayesian	Naïve Bayes	89.7%
2018	Aakash Chauhan et al, [13]	Proposed a disease prediction model for heart disease by utilizing evolutionary rule learning	Association Rule	53%
2018	Aditi Gavhane et al, [14]	Suggested prediction for heart disease that utilizes Machine Learning	Multi-Layer Perceptron	91%
2015	Ankita Dewan et al, [15]	Recommended a disease prediction system that uses data mining classification hybrid technique	Neural Network, Decision Tree and Naive Bayes	87%

III. CONCLUSION

To predict the diseases using multiple data mining and machine learning techniques and algorithms have been summarized. Each algorithm has its disease prediction performance and one can apply the proposed system according to his or her needs. It is also possible to improve the performance and accuracy of the algorithm if independent variables or features are selected more correctly. After studying these methods it has been found that if we have a structured dataset then the accuracy of prediction is improved. If we can collect millions of structured datasets for a particular disease then that disease can be predicted with the highest accuracy and data mining can help us collecting such datasets.

These reviews have shown as that any machine learning model can be improved through multiple revisions and by changing the algorithms that they use. Sometimes it is good for the model but sometimes it reduces the performance of the model.

In conclusion, by using a literature survey it has been identified that a single algorithm cannot perform very well but if it is combined with other algorithms then the accuracy can have huge improvements. So the combination of these algorithms should be used in multiple sequences and a comparison must be collected to check which of these combinations are performing with more accuracy than the previous ones in predicting the disease.

There are so many possibilities ahead that could be utilized to improve the performance of these prediction systems and increase the scalability and accuracy of the system. It is not possible to explore all the options within this limited time, the following research options can be performed in the future. Multiple classification techniques and regression techniques should be combined, different types of decision trees and neural networks should be used to check how much accuracy has been improved.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities" IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2] **S**ayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network" IEEE, 978-1-5386-5257-2/18, 2018.

[3] Naganna Chetty, Kunwar Singh Vaisla and Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach" IEEE, DOI 10.1109/ICACCE.2015.67, pp. 569-572, 2015.

[4] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach" IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4, pp. 1211-1215, 2019.

[5] Lambodar Jena and Ramakrushna Swain, "Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers" IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.

[6] Dhomse Kanchan B. and Mahale Kishor M., "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis" IEEE, 978-1-5090-0467-6/16, pp. 5-10, 2016.

[7] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction" IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, " Diabetes Disease Prediction Using Data Mining" IEEE, 978-1-5090-3294-5/17, 2017.

[9] Rashmi G Saboji and Prem Kumar Ramesh,"A Scalable Solution for Heart Disease Prediction using

Classification Mining Technique" IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.
[10] Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.
[11] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019.
[12] Anjan Nikhil Repaka, Sai Deepak Ravikanti and Ramya G Franklin,"Design And Implementing Heart Disease Prediction Using Naives Bayesian" IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8, pp. 292-297, 2019.
[13] Aakash Chauhan, Purushottam Sharma, Vikas Deep and Aditya Jain, "Heart Disease Prediction using Evolutionary Rule Learning" CICT 2018.
[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, "Prediction of Heart Disease Using Machine Learning" IEEE Xplore ISBN: 978-1-5386-0965-1, pp. 1275-1278, 2018.
[15] Ankita Dewan and Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification" IEEE, 978-9-3805-4416-8/15, pp. 704-706, 2015.

ABOUT THE AUTHORS

Md. Ehtisham Farooqui has done B.Tech in Computer Science from Institute of Technology and Management, GIDA, Gorakhpur, India and currently perusing M.Tech from Integral University, Lucknow, India

Dr. Jameel Ahmad is Assistant Professor, Department of Computer Science and Engineering in Integral University and has more than 18 years of teaching experience.

.

