Analytical Approach towards Prediction of Diseases Using Machine Learning Algorithms

Ayushi Grover, Anukriti Kalani, Sanjay Kumar Dubey
Department of Computer Science Engineering, ASET
Amity University Uttar Pradesh, Sector 125, Noida, India

Abstract— **Healthcare is a human right and in this complex technology driven world, healthcare industry is equipped with modern technology for the solution of disease but struggles when it comes to prevent them beforehand. Machine learning can transform healthcare industry. Machine Learning provides a wide scope of apparatuses, strategies and structures to address difficulties like electronic record the executives, information combination, PC supported judgments and disease expectation. This research paper aims to predict disease accurately according to the symptoms of patients and helps doctor in better diagnosis, further reducing the cost of treatment and improving quality of life. It includes the comparative study of the outcomes and time required for analysis and prediction of disease by various machine learning algorithms and contribute towards research in healthcare department.**

Keywords—healthcare; data mining; machine learning; algorithms; big data

I. INTRODUCTION

Recently the report generated by the World Health Organization stated that the numerous diseases are caused by unhealthy lifestyles which includes smoking, alcoholism, lack of exercise, excessive salt intake and many more leading to huge number of deaths across the world. Among these deaths, more than 40% are related to people under the age of 60 years [1]. Due to lack of knowledge and busy schedule People often ignore various symptoms occurred, leading to more critical conditions. Hence there is an urgent need for the early prediction and accurate detection of diseases.

The healthcare industry is an ever-evolving industry. This industry generates large amount of data. The amount of medical or healthcare information being created is growing at a rate of 48% annually. The healthcare informatics along with the data mining techniques is a field which generally deals with storage, retrieval, and how to gain knowledge of various diseases [2]. Effective analysis of such data results in early detection and prevention of diseases which could further aggravate, resulting in the growth of survival rate.

Algorithms in machine learning are utilized for detection of infections based on some symptoms. These algorithms are used to maintain the complete hospital data and for processing both structured and unstructured data. To defeat the trouble of the missing data it uses a "latent factor model". Models manufacture utilizing AI calculation productively use and investigates the information and convey the outcomes utilizing both past and continuous data. These algorithms have effectively improved the medical standards empowering the specialists to take a decent choice on patient's analyses and treatment choices [3]. As far as we could possibly know none of the current work concentrated on the both data types structured and unstructured. The comparative study of the result generated through these algorithms is depicted.This review paper focuses on exploring and drawing a comparative analysis between the various techniques used in the previous research in this particular field and aims to tackle two main questions:

RQ1. Are machine learning algorithms better equipped with predicting diseases as compared to traditional ways used by various doctors and hospitals?
RQ2. What are some of the best algorithms of Machine Learning to predict diseases?

II. RESEARCH METHODOLOGY

Research methodology depicted is based on the digital library and Google Scholar that focused on various researches. Searched various different research papers and considered the journals beneficial with reference to the topic. While writing this paper the main objective is to present the inference of various papers and find the suitable algorithms for the accurate and early prediction. Numerous previous research papers are referred at different stages. While preparing the review phase, the need for writing this review is considered, the related research questions are explained, and the review standard is identified. Finally while generating the review, the initial researches are chosen, the attributes evaluated during studies are defined, the data abstraction & examination is carried out and the data generated is then evaluated. While stating the review phase the publishing mechanism are identified and the review report is generated. In this review, examined the papers from 2010-2019 as more of the work with advanced technologies in this field is done in the recent years.

III. LITERATURE REVIEW

In 2010, W. J. Roy and W. F. Stewart predicted a model which uses the Electronic Health Record databases and applied various ML algorithms on this data to predict heart disease.

978-1-7281-2791-0/20/$31.00

Boosting, Support Vector Machine, Logistic Regression were used to identify patients at high risk for critical conditions and by comparing their performances found that Logistic regression algorithm when used with the model selection based on Bayesian information criterion generated the most accurate result and SVM had the poorest performance among all due to imbalanced data [4]. In 2014, W. Raghupathi and V. Raghupathi presented a paper stating how the techniques of the big data analytics can alter the healthcare industry. For evaluation of data the criteria followed should incorporate accessibility, progression, usability, versatility, capacity to control at various degrees of granularity, protection and security enablement, and quality affirmation. Paper includes discussion on various advantages, characteristics and architectural structure of big data techniques in medical industry [5]. In 2016, S. Patel and H. Patel presented a survey paper regarding data mining techniques in medical domain for accurately using the large data and make accurate prediction about the human diseases. They gave a summarized information for predicting specific diseases using specific technique like coronary heart diseases we can use Naïve Bayes and many more diseases and techniques. This can help in research towards using intelligence in medical industry [6]. In 2016, M. Chen, Y. Ma, J. Song, C. Lai, B. Hu proposed a system called "Smart clothing". This paper presents design details, key advancements and hands-on for smart clothing framework. The "Smart clothing" system can be help in monitoring various symptoms and provide personal services on basis of physio-information received through clothes with the help of sensors, cloud computing, big data and machine learning algorithms [7]. In 2017, Vidya Zope, Pooja Ghatge, Aaron Cherian, Piyush Mantri, Kartik Jadhav proposed a system using Association rule mining, clustering algorithm like K-means and Classification Algorithm like Decision tree, KNN, logistic regression, random forests and Naïve Bayes. They performed study and researched mainly about the comparison between Naïve Bayes and logistic regression. The proposed architecture used R, python, Django, jQuery, JavaScript, HTML and CSS. They successfully found out that Naïve Bayes algorithm gave a better result in terms of accuracy and time [8]. In 2017, B. Nithya, Dr. V. Ilango presented a paper in which they depicted the extensive study of machine learning tools and techniques used in healthcare. They discussed various Machine Learning processes like supervised learning (predictive model), unsupervised learning (descriptive model), Semi Supervised and Reinforcement Learning and ML tools like platform, library, GUI, API, CLI, Local and remote. Prediction Models helps in diagnosis and prediction of chronic diseases like cardiovascular diseases, diabetes, hepatitis, cancer etc [9].In 2017, M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn proposed a Wearable 2.0 human services framework to improve Experience Quality and Service Quality of the future medical industry. In the proposed framework, washable smart clothing, which comprises of sensors, terminals, and wires, which collects information and status gave by cloud-based machine insight [10]. In 2018, Shraddha Subhash Shirsath, Prof. Shubhangi Patil proposed a system using CNN-MDRP (unstructured and structured data) over CNN-UDRP (structured data) CNN-MDRP also resolve the issues with CNN-UDRP and gave more accurate results as compared to previous prediction algorithm. They proposed a system which also compared two types of algorithm Naïve Bayes for structured data and CNN-MDRP which resulted in CNN-MDRP giving more accurate results over large data collected from hospital [11]. In 2018, Harini DK and Natesh M proposed the system that combined the structured and unstructured data in medical field to survey the danger of ailment. They used latent factor model to complete the missing records collected from a hospital and by using statistical knowledge they determined the major interminable infections in the specific locale. Structured data can be processed using KNN, Naïve Bayesian and Decision Tree. For unstructured text data CNN algorithm was used. They proposed CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm considering both structured and unstructured data [12]. In 2018, Vinitha S, Sweetlin S, Vinusha H and Sajini S proposed a cost-effective system in which the analysis accuracy of medical big data is increased using Machine Learning algorithms such as Decision Tree and Map Reduce. Analysis of disease is weakened due to different diseases occur in different regions and have unique symptoms. The existing work mostly considered the structured data. Decision Tree algorithm is used to predict diseases as well as the sub diseases and Map reduce is used to speed up the operational efficiency by reducing the query retrieval time. Their proposed system reaches 94.8% accuracy with a regular speed which is faster than that of CNN unimodal disease risk prediction (CNN-UDRP) [13]. In 2018, Mr. Santosh A. Shinde, Dr. P. Raja Rajeswari presented a research paper to serve as a guideline in research utilization of machine learning strategies in the healthcare prediction, and gives further research headings required into wellbeing expectation framework utilizing ML. They did an extensive research and reviewed on different aspects of prediction of diseases using binary classification, multi class and mental health analysis [14]. In 2018, Mr. Chala Beyene, Prof. Pooja Kamat presented a survey paper in which they did an extensive study of different algorithms on prediction of heart diseases specifically which can further help in better decisions and selection method for rightly predicting the disease. They concluded that J48 (Decision tree algorithm), Naïve Bayes and Support Vector machine algorithm helped in early automatic diagnosis and reduced the time to get the results and complete accuracy [15]. In 2019, Chieh-Chen Wu, Wen-Chun Yeh, Wen-Ding Hsu presented a paper that talks mainly about early prediction of Fatty liver Diseases (FLD) and to develop an architecture using ML algorithms to further help doctors, physicians in preventing, diagnosing and manage FLD. They used classification algorithm like random forest, Naïve Bayes, logistic regression and artificial neural networks and observed that random forest algorithm showed an accuracy of 87.48% which is better as compared to other algorithms [16]. In 2019, J Clin Med presented a in which the Comparison is done using scikit-learn library and auto-sklearn on the cardiovascular

disease dataset. He successfully founded that automatic machine learning produce classifiers performed better results than machine learning algorithms [17].

IV. COMPARITIVE ANALYSIS

Table 1:Summary of Literature Review

S.No	Year	Author	Keywords	Model / Algorithms	Analysis
1.	2016	M. Chen, Y. Ma, J. Song, C. Lai, B. Hu	Smart clothing, Health monitoring, Wearable cloud computing	–	Using bigdata and machine learning algorithms, sensors smart clothing system can help in early diagnosis and personalized services
2.	2016	S. Patel and H. Patel	Data Mining, Health Care, Classification, Clustering, Association	–	A detailed description about using specific algorithms for specific diseases increases accuracy.
3.	2017	Vidya Zope, Pooja Ghatge, Aaron Cherian, Piyush Mantri, Kartik Jadhav	Data Mining, Healthcare, Prediction System	K-means, logistic regression, and Naïve Bayes	Naïve Bayes algorithm gave a better result like efficient time & precise prediction
4.	2017	B. Nithya, Dr. V. Ilango	Predictive Analytics, Predictive Tools, Machine predictive techniques	–	extensive study of machine learning tools, techniques and processes.
5.	2017	M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn	IOT, Smart clothing	Wearable 2.0 machine healthcare system	Proposed a launderable savvy apparel framework, comprises of sensors, anodes and wires to gather user's physiological data and analyze the result provided by cloud-based intelligence.
6.	2018	Harini DK and Natesh M	Disease Prediction, Convolution Neural Network MDRP	CNN-MDRP, KNN, Naïve Bayesian, Decision Tree	System combined both structured and unstructured data.
7.	2018	Vinitha S, Sweetlin S, Vinusha H and Sajini S	Big data analytics, healthcare.	Decision Tree, Map Reduce, CNN-UDRP	Cost-effective system with 94.8% accuracy with a faster than CNN-UDRP.
8.	2018	Mr. Santosh A. Shinde, Dr. P. Raja Rajeswari	EHR, Big data analytics	–	reviewed on different aspects of prediction of diseases using binary classification, multi class and mental health analysis.
9.	2018	Mr. Chala Beyene, Prof. Pooja Kamat	KFold Cross-Validation, Heart Disease, Machine Learning.	Decision Tree, Naïve Bayes and Support Vector Machine	Heart diseases can be accurately predicted using ML algorithms.

10.	2018	Shraddha Subhash Shirsath, Prof. Shubhangi Patil	Data analytics, Health care data, Machine learning	Naïve Bayes, CNN-UDRP, CNN-MDRP	CNN-MDRP gave 94.8% accuracy in results & work on both unstructured and structured data.
11	2019	Chieh-Chen Wu, *et al*	Disease, fatty liver different grade, random forest machine learning	Random forest, Naïve Bayes, regression, classification and ANN	Random forest algorithm showed an accuracy of 87.48% to predict fatty liver disease which is comparatively better.
12.	2019	J Clin Med	Heart Disease, Machine Learning Algorithms, AI	Least Squares Twin Support Vector Machines, KNN, Naïve Bayes, Decision Tree, Auto-ML	automatic machine learning produce classifiers that perform preferable outcomes over the ones utilizing AI calculations.

V. Discussion

Machine learning algorithms can be differentiated on the basis of their accuracy and interpretability for intended purpose of this paper which is depicted in the following figure 1 :

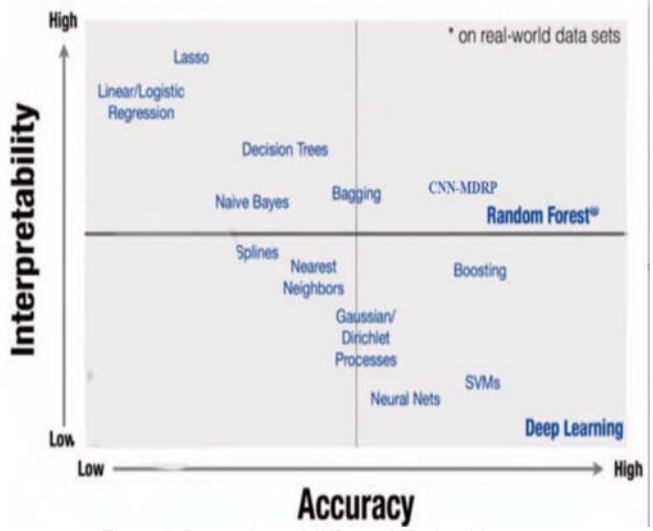

Figure 1:Comparison of different ML algorithms

After extensive research, we can answer the research questions that were put forward by this paper in the beginning:

A. Research Question-1

Through various observations which explored the traditional ways of data collection which includes structured data only required for predicting diseases, there was this direct understanding that the data analysis wasn't accurate and difficult to obtain frequently. Machine Learning algorithms and Data Mining Techniques have been used to complete the missing data through various sources whereas traditional survey data collection is otherwise expensive, incomplete and time consuming.
Predictive Analysis model can generate predictions using both structured and unstructured data whereas traditional model uses only structured data.

B. RESEARCH QUESTION-2

Predictive analysis approach equipped with new convolutional neural network based multimodal disease risk prediction uses both types of data namely structured & unstructured to predict various diseases and give us an account of the digitalization in prediction of healthcare. Through extensive research we observed that this new CNN MDRP precision is 94.8% than CNN-UDRP algorithm. It also helps in early and accurate prediction according to the symptoms of patients and helps doctor in better diagnosis and treatment, further improving quality of life. After extensive study of research papers it is observed that research in this area is constantly increasing as well as the medical data. From 1998-2018 the related work using machine learning techniques is rising shown in Fig.2. Graph indicates the rising research work in healthcare industry[18].

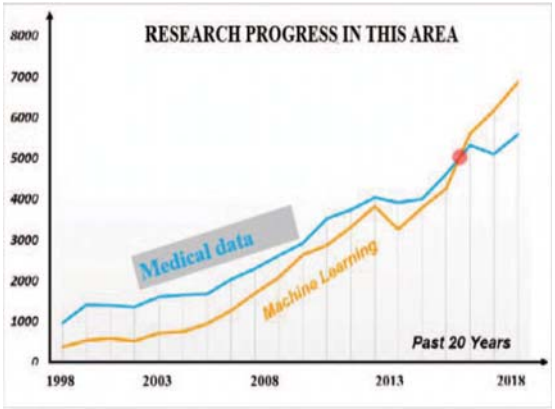

Figure 2:Research progress in healthcare domain

VI. CONCLUSION AND FUTURE WORK

After the extensive research done on different machine learning algorithm, big data analytics and data mining tools and techniques we have successfully compared the results generated by different machine learning algorithms based on precision and effective time taken for analysis of the patients data in healthcare domain which can help in better future of healthcare industry and could also decrease the rate of fatality. This paper can help towards the early and accurate prediction of diseases and gives a direction for further research. This may further help as a guide to other researches as a basis for comparative analysis of different algorithms and which algorithm to choose to achieve the goal in different situations.

VII. REFERENCES

[1]Min Chen ; Yin Zhang ; Meikang Qiu ; Nadra Guizani ; Yixue Hao, "SPHA: Smart Personal Health Advisor Based on Deep Analytics, " IEEE Communications Magazine Volume: 56 , Issue: 3 , pp. 164 – 169, March 2018.

[2] W. Yin and H. Schutze,"Convolutional neural network for paraphrase identification," Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 901–911,2015.

[3] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," Scientific Reports volume 6, 2016.

[4] W. J. Roy and W. F. Stewart, "Prediction modelling using ehr data: challenges, strategies, and a comparison of machine learning approaches," Medical care, volume 48, Issue 6, 2010, pp. 106- 113.

[5] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potentials," Health Information Science and Systems, volume 2, Issue1, 2014.

[6] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.

[7] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring," ACM/Springer Mobile Networks and Applications' Vol. 21, No. 5, pp.825C845, 2016.

[8] Vidya Zope, Pooja Ghatge, Aaron Cherian, Piyush Mantri, Kartik Jadhav, "Smart Health Prediction using Machine Learning," IJSRD - International Journal for Scientific Research & Development Vol. 4, Issue 12, ISSN (online): 2321-0613, 2017.

[9] B. Nithya, Dr. V. Ilango, "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques," International Conference on Intelligent Computing and Control Systems ICICCS 2017.

[10] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Communications., vol. 55, no. 1, pp. 54_61, Jan. 2017.

[11] Shraddha Subhash Shirsath, Prof. Shubhangi Patil, "Disease Prediction Using Machine Learning Over Big Data," International Journal of Innovative Research in Science, Engineering and Technology, Vol. 7, Issue 6, June 2018 .

[12] Harini D K, Natesh M, "PREDICTION OF PROBABILITY OF DISEASE BASED ON SYMPTOMS USING MACHINE LEARNING ALGORITHM," International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 05 ,May-2018.

[13] Vinitha S, Sweetlin S, Vinusha H and Sajini S, "DISEASE PREDICTION USING MACHINE LEARNING OVER BIG DATA," Computer Science & Engineering: An International Journal (CSEIJ), Vol.8, No.1, February 2018.

[14] Mr. Santosh A. Shinde, Dr. P. Raja Rajeswari, "Intelligent health risk prediction systems using machine learning: a review", International Journal of Engineering & Technology, 7 (3) 1019-1023, 2018.

[15] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques," International Journal of Pure and Applied Mathematics, 2018.

[16] Chieh-Chen Wu, *et al* "Prediction of fatty liver disease using machine learning algorithms," Computer Methods and Programs in Biomedicine, Volume 170, Pages 23-29 March 2019.

[17] J Clin Med, "Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction," vol. 8(7), PMC6678298, 18 July 2019.

[18] Ge Wang, "A Perspective on Deep Imaging," IEEE Access 2016, DOI:10.1109 / ACCESS.2016.2624938, 2016.