Cardiovascular Disease prediction by using Big Data Tool and Machine Learning Approaches

By

Abhishek S P

Abstract

This paper investigates the prediction of Cardiovascular Diseases , a serious cardiovascular condition, by the combination of big data technologies and machine learning techniques. The study combines several datasets related to heart disease and uses algorithms including Naive Bayes, KNN, Random Forest, and Logistic Regression for predictive modelling by utilising an organised health information system architecture. The extensive literature study highlights the value of big data analytics and machine learning in the healthcare industry, including prior successes with risk assessment and illness classification. The project intends to analyse key risk factors linked with Cardiovascular Diseases using a combined dataset from many sources, and evaluate algorithmic performance using measures such as accuracy and precision. In addition to addressing issues like dataset variety and healthcare dynamics, the work addresses therapeutic implications for early detection and treatment.

Keywords: Cardiovascular Disease, Big Data Analytics, Machine Learning, Health Information System, Predictive Modeling, Data Integration, Algorithmic Comparison, Performance Measures, Healthcare Technology

Introduction

Throughout history, health has always been of utmost importance, even before the development of modern technologies. After years of substantial growth, the healthcare sector now provides a large study area. It is critical to keep improving healthcare technology, especially since that patient data and medical outcomes from advanced diagnostic equipment are digital. But understanding and interpreting the vast amount of data produced by this information revolution is extremely difficult. The field of healthcare uses big data analytics, a complete strategy that can handle a high amount, diversity, and complexity of data connected to health, to meet this difficulty. Volume, Variety, and Validity (3Vs) are the three criteria that Doug Laney identified in 2001 as characterising big data. Applications for this new paradigm in data processing and analysis may be found in a number of fields, including science, engineering, business, social sciences, finance, and—most importantly—healthcare. Hadoop and Spark are two of the most popular technologies for processing and analysing data among the many others. Hadoop, which consists of the Hadoop Distributed File System (HDFS) for distributed storage and MapReduce for distributed processing, has emerged as a key component for managing large and varied information.

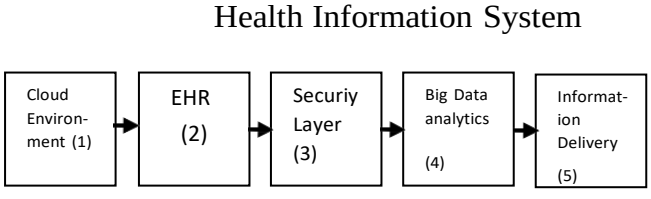

Fig 1.Health Information System Proposed by Yousuf, 2014

Ahmed Yusuf presented a Big Data Analytics-based structured architecture for health information systems in 2020. There are five main components to this framework, which is mainly intended to help patients and make organising and analysing large amounts of healthcare data easier. The Cloud Environment is the fundamental layer that facilitates the provision of various services and grants users access to data. Patient data is gathered from several sources via the Electronic Health Record (EHR) component. Security issues are handled by a security layer using encryption and authentication techniques. The fourth component, big data analytics, uses extensive analytics tools for thorough data analysis. Information delivery, the last layer, makes it easier to gather health-related data from diverse sources, which enhances a range of healthcare services.

Literature Review

Computer technology advancements have completely changed the healthcare industry, especially when it comes to machine learning and big data analytics (ML and BDA). With an emphasis on tackling analytical issues for developing intelligent healthcare systems, this literature review seeks to highlight the integration of machines and predictive models in the healthcare sector [3].

One of the main drivers of motivation is the discovery of patterns in data, which is a basic component of data analytics [3]. Hughes Health data is recognised for its significant contribution to the comprehension of correlations in data patterns, providing more insights that are essential for well-informed decision-making [4]. Interestingly, the Nearest Neighbour and Decision Tree algorithms—two data mining techniques—are suggested for use in the categorization of diseases, including diabetes [5].

When it comes to probabilistic data collecting, Prasanna Kumar et al. [6] focus on the examination of reciprocal relationships in the data that is gathered. The result of their efforts is the creation of a stochastic prediction model that extrapolates information about diseases based on an individual's present state of health. Using a variety of data sources, such as environmental sensors and Twitter data, Sudha Ram et al. [7] provide a method useful in predicting emergency patient visits due to asthma.

The suggested data analytics framework by Yichuan Wang et al. [8] adds to the body of research specifically for the healthcare industry. They include Pattern Analysis, Unstructured Data Analysis, Decision Support, Predictive Analysis, and Traceability as the five main components of big data analytics. A healthcare paradigm utilising Smart Home Big Data is presented by Abdulsalam Yassine et al. [9], which successfully teaches and identifies patterns of human behaviour. Their method includes cluster analysis of patterns, frequent pattern mining, and occupant behaviour prediction.

A testing theory concentrating on the use of big data characteristics for illness management and diagnosis is presented by Javier Andreu-Perez et al. [10]. Their work encompasses a variety of health data categories, such as sensor informatics, imaging informatics, health informatics, and

classical bioinformatics. A model for determining the likelihood of developing diabetes mellitus is put out by Nongyao [11], who makes use of machine learning techniques such logistic regression, artificial neural networks, and decision trees.

Cardiovascular Diseases is a class of disorders that affect the heart muscle. Genetic abnormalities, viral infections, autoimmune diseases, exposure to toxins, pregnancy, and other medical conditions are some of the causes that can lead to the development of Cardiovascular Diseases . The widespread prevalence of cardiovascular illnesses makes the use of machine learning models essential for timely identification and efficient treatment. This review of the literature provides background for our work, which uses a heterogeneous dataset combined from five different heart disease datasets to predict Cardiovascular Diseases [context].

TABLE I
COMPARISON OF PROPOSED WORK DONE INLITERATURE REVIEW

	Author	Proposed Work
1	Prasanna Kumar et. al	Proposed probabilistic data collection, which performs ananalysis of the mutual relationship between the data collected. And developed stochastic predictivemodel
2	Yichuan Wang et. al	data analytics structure which identified five big data analytics entities like Pattern's Analysis, unstructured Data Analysis,Decision Support, Predictive and traceability.
3	AbdulsalamYassine et al.,	Developed a model that discover human activity patterns with the help of Smart home big data for health care
4	Javier Andreu-Perez et.al	They introduced a theorem about the diagnosis and disease management for treatment using the feature of big data.
5	Nongyao	Model for the risk of diabetes by using four famous machine- learning algorithms such as the Decision Tree, Artificial Neural Network, and Logistic Regression

Objectives

- Analyse Risk variables: Using a large dataset, determine and evaluate the major risk variables connected to Cardiovascular Diseases .
- Dataset Integration: To build a reliable dataset for Cardiovascular Diseases prediction, integrate several heart disease datasets.
- Algorithmic Comparison: Use machine learning methods to assess how well they predict Cardiovascular Diseases , such as Naive Bayes, KNN, Random Forest, and Logistic Regression.
- Performance measures: Assess an algorithm's performance using measures such as recall, accuracy, and precision, among others.
- Clinical Implications: Talk about the possible uses and clinical ramifications of prediction models for Cardiovascular Diseases early identification and treatment.
- Future Improvements: Make suggestions on how to improve predictive accuracy in clinical scenarios going forward, such as adding more features and real-time data.

Research Methodology

Data Collection

Overview of Dataset- Cardiovascular Disease Dataset: The dataset utilised combines data from five distinct cardiac datasets, namely Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog, to provide a comprehensive understanding of cardiovascular disease (CVD). This is one of the most extensive databases on heart disease that is available for study, including 11 characteristics. Important characteristics in the dataset include age, sex, type of chest pain, maximum heart rate achieved, exercise-induced angina, serum cholesterol, fasting blood sugar, resting blood pressure, old peak measurement, slope of the peak exercise ST segment, and the output class indicating the presence or absence of heart disease.

Cleveland: 303 observations were added to the dataset, Hungarian: 294 records, Switzerland: One hundred and thirty-three notes, 200 observations in Long Beach, VA, Stalog, or Heart 270 observations make up the data set. 1190 observations in all.

272 observations were duplicated.

918 observations make up the final dataset after duplicates are eliminated. The scope and diversity of characteristics are improved by this dataset merging, allowing for a more thorough investigation of cardiovascular disease prediction modelling.

Dataset Source: The UCI Machine Learning Repository's Index of Heart Disease datasets provides access to each of the utilised datasets separately. Databases on Heart Disease. archive.ics.uci.edu/ml/machine-learning-databases/heart-disease

precision, and recall matrix as they are widely used in standard data mining fields [15].

Based on the confusion matrix, it will be very easy to calculate the accuracy of proposed algorithm. Accuracy can be finding by using following formula.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots (1)$$

Where TP=True PositiveTN= True Negative, FP= False Positive FN=False Negative

Precision can defined using the above equation where the total numbers of correctly classified positive samples are dividing by the total number of true positive samples.

$$Precision = \frac{TP}{TP + FP} \dots\dots (2)$$

Recall is define in equation 3 as the total number of correctly classified positive samples divided by the total number of predicted positive samples.

$$Recall = \frac{TP}{TP + FN} \dots\dots (3)$$

Naïve Bayes Algorithm: It works through the Probability Major, which chase a distinct order for execution. This method implemented using the following formula:

$$Posterior\ Probability\ P(clx) = \frac{Likelihood\ P(xlc)\ X\ Class\ Prior\ Probability\ P(c)}{Predictor\ Prior\ Probability\ P(x)} \dots\dots (4)$$

This method uses the following formula for implementation: For the Navy Baas, here the dataset divided in the ratio of80: 20, where the training set is around 80% while the testset is 20%. The Gaussian algorithm chosen to create themodel that is the simplest classifier model [15].

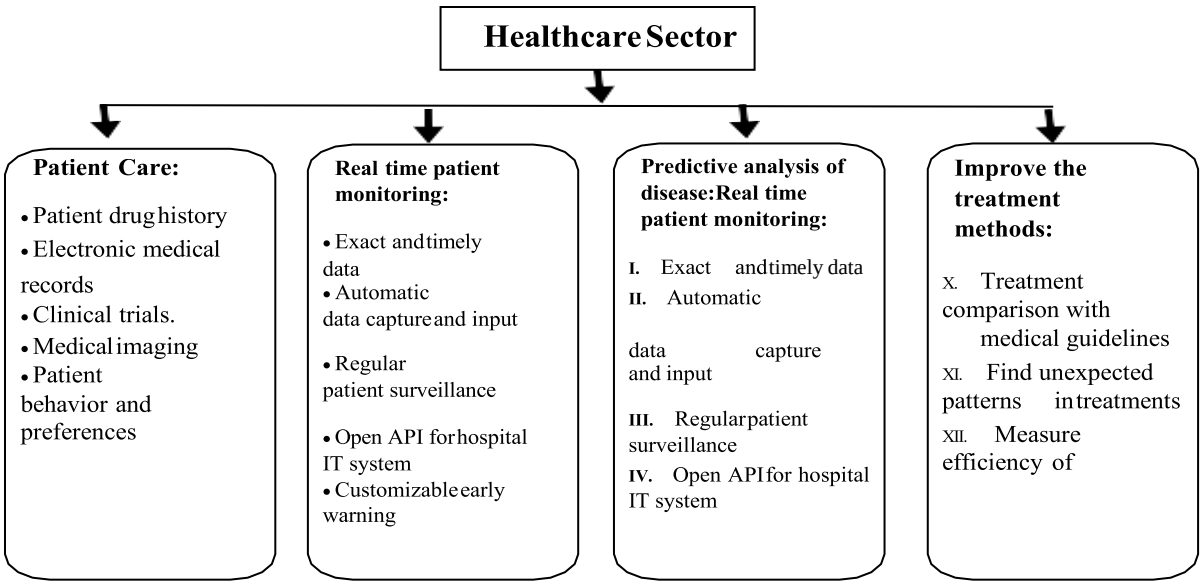

Fig 2 Big data in Healthcare sector

Fig. 2 Big data in the healthcare industry By using expressive, predictive, and reusable massive data analysis approaches, each of these four pillars of value healthcare may be closely monitored.

A. Patient-centered care: Based on clinical data and medication dosage limitation, it assists the patient in light of the initial phase's distance from the results. Reduced readmission rates in hospital clinics and lower patient costs are further benefits of this.

B. Predictive Analysis of Diseases: based on real-time analysis, vaccine the viral issues at the outset before they spread. This may be determined by looking at the social logs of the patients who are ill in a certain location. This further motivates medical professionals to take necessary preventative action before casualties arise.

C. Real-time patient monitoring: this checks to see if hospital arrangements meet the Indian Clinical Committee's standards. This kind of recurring registration aids in the government's ability to take the required action to close the facility.

D. Redesign the therapy System: Drug analyses, which are subject to quick changes, provide the basis for investigating the course of therapy for a patient who was previously prescribed medication. The patient's investigation information, which is derived from their symptoms, aids the physician in prescribing new patients with appropriate medications [12].

Flow Chart & Prosed Methodology

This section provides a quick overview of the advancements made in the technology. The suggested classifier model accepts input into the diabetes data set and mostly alerts patients who have the condition. Various machine learning models, including random forest, logistic regression, KNN algorithm, and naive bayes, have been evaluated using input data sets, and the findings derived from these models have been compiled based on the outcomes of the experiments. The algorithm that performs the best has been selected based on its accuracy in accurately predicting the condition that precedes diabetes. Figure 3 outlines the approach taken to build the model and compute its comparative analysis in order to accurately forecast the onset of diabetes [13].

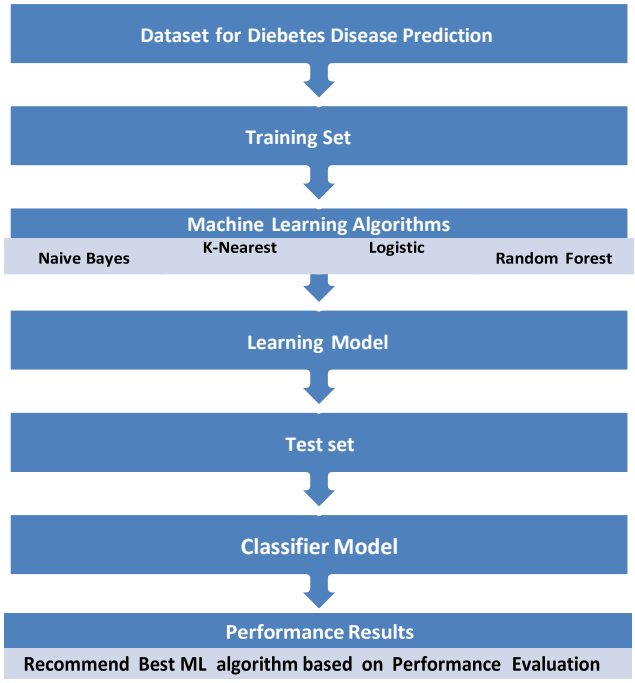

Fig 3 Proposed Methodology Flowchart

The stages related to the Fig. 3 processes are shown in the following. Suggested Classifier Techniques In steps The Proposed Methodology's Procedure

Stage 1: Using a Python programme, the data collection was preprocessed for diabetic illness.

Step 2: Following the first phase, the data sets were split 80:20 between training and testing sets.

Step 3: Several machine learning techniques, including random forest, logistic regression, naive bayes, and KNN algorithms, are chosen for testing in this step.

Step 4: Based on a data set, an ML model was created for a machine learning algorithm at this step.

Step 5: The model was tested on the testing set once it was created.

Step 6: Comparative assessment of the classifier's experimental output.

Step 7: A comparative examination of the experimental performance results obtained from the classifier model is carried out, and the most effective algorithms are chosen according to the accuracy and precision of their findings.

The suggested classifier model was created with the aid of a Python programme and is dependent on the successful completion of experimental procedures. This has the ability to estimate test outcomes.

Limitations

Although the goal of this research is to improve Cardiovascular Diseases detection prediction by using big data and machine learning, there are a few constraints that need be taken into account. First off, the model's generalizability may be limited since the dataset, albeit large, might not fully capture all differences in Cardiovascular Diseases. Furthermore, the intrinsic diversity present in the source datasets adds unpredictability that might compromise the resilience of the model.

The dynamic character of healthcare, with its ever-changing diagnostic standards and treatment options, is another factor to take into account. Over time, the model's relevance may be affected by the dataset's static nature, which may not accurately reflect these improvements. Furthermore, even if the process of integrating datasets is extensive, it poses problems with data harmonisation and possible biases from different datasets.

Conclusion

With an emphasis on Cardiovascular Diseases specifically, this study concludes by highlighting the critical role that big data analytics and machine learning play in improving the prediction of cardiovascular illness. A thorough grasp of predictive modelling is facilitated by the integration of various datasets related to cardiac disease and algorithmic comparisons utilising Naive Bayes, KNN, Random Forest, and Logistic Regression. The study emphasises how important it is to pinpoint the main risk factors for Cardiovascular Diseases , providing insightful information for prompt diagnosis and effective treatment. Although there is potential in the suggested approach, it is considered to have inherent limitations due to the diversity of datasets and the changing nature of healthcare standards. In spite of these obstacles, the study establishes a framework for upcoming improvements by recommending the incorporation of new elements and real-time data to improve forecast accuracy.

References

[1]. Prableen Kaura, Manik Sharma, Mamta Mittal "Big Data and Machine Learning Based Secure Healthcare Framework" International Conference on Computational Intelligence and Data Science (ICCIDS 2018) 10.1016/j.procs.2018.05.020

[2]. Thérence Nibareke and Jalal Laassiri " Using Big Data- machine learning models for diabetes prediction and flight delays analytics" J Big Data (2020) 7:78 https://doi.org/10.1186/s40537-020-00355-0 Springer

[3]. Rahul C. Basole, Mark L. Braunstein, And Jimeng Sun, "Data and Analytics Challenges for a Learning Healthcare System", ACM Journal of Data and Information Quality, Vol. 6, No. 2–3, Article 10,

Publication date: July 2015

[4]. Emrana Kabir Hashi, Md. Shahid Uz Zaman , Md. Rokibul Hasan, "An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques", International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16 -18, 2017, IEEE

[5]. Md. Golam Rabiul Alam, Rim Haw, Sung Soo Kim, Md. Abul Kalam Azad, Sarder Fakhrul Abedin, Choong Seon Hong, "EM-Psychiatry: An Ambient Intelligent System for Psychiatric Emergency", IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 12,

NO. 6, DECEMBER 2016

[6]. PRASAN KUMAR SAHOO, SUVENDU KUMAR MOHAPATRA,

SHIH-LIN WU " Analyzing Healthcare Big Data With Prediction for Future Health Condition", Vol-4 20176 IEEE Digital Object Identifier 10.1109/ACCESS.2016.2647619

[7]. Ram, S., Zhang, W., and Williams, M., Predicting Asthma-Related Emergency Department Visits Using Big Dat a. IEEE Journal 19(4): 1216–1223, 2015.

[8]. Wang, Y., and Kung, L. A., Terry Anthony Byrd, " Understanding itscapabilities and potential benefits for healthcare organizations". Journal of Technological Forecasting and Social Change 126:3 –13, 2018.

[9]. Abdulsalamyassine, S., Mining Human Activity Patterns From Smart Home Big Data for Health Care Applications. IEEE Access 5:13131 – 13149, 2017.

[10]. Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang, Fellow, " Big Data for Health" IEEE, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4, JULY 2015

[11]. Nongyao Nai-aruna,*, Rungruttikarn Moungmaia " Comparison of Classifiers for the Risk of Diabetes Prediction" (http://creativecommons.org/licenses/by-nc-nd/4.0/) Procedia Computer Science 69 (2015) 132

[12]. Archenaa J. et al. (2015) " A Survey of Big Data Analytics in Healthcare and Government." Procedia Computer Science. 50: 408 – 413.

[13]. Ayman Mir, Sudhir N. Dhage " Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare" 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 978-1-5386-5257-2/18/$31.00 c 2018 IEEE

[14]. Senthilkumar SA, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran, Chandrakumarmangalam " Big Data in Healthcare Management: A Review of Literature" American Journal of Theoretical and Applied Business 2018; 4(2): 57-69

http://www.sciencepublishinggroup.com/j/ajtab doi: 10.11648/j.ajtab.20180402.14 ISSN: 2469-7834 (Print); ISSN: 2469-7842 (Online)

[15]. K. Shailaja, B. Seetharamulu, M. A. Jabbar " Machine Learning in Healthcare: A Review" Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record 42487; IEEE Xplore ISBN:978-1-5386-0965-1

[16]. Usha Nandhini, Dr. K. Dharmarajan "Diabetic Analysis on Big data and Machine Learning - A Literature Review" Parishodh Journal ISSN NO:2347-6648 2020.