IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Prediction of Heart Disease Using AI and NLP Techniques

S. Sabeena[1]*, V. Sujitha[2]

[1]Dept. of Computer Applications, Pioneer College of Arts and Science, Coimbatore, India
[2]Dept. of Computer Applications, Pioneer College of Arts and Science, Coimbatore, India

*Corresponding Author: sabeena.mphil@gmail.com

Available online at: www.ijcseonline.org

Received: 23/Jan/2018, Revised: 31/Jan/2018, Accepted: 15/Feb/2018, Published: 28/Feb/2018

Abstract — The leading cause of death in recent days due to heart diseases for both men and women. When the heart is too weak to pump blood throughout the body, heart failure occurs. The heart attack happens suddenly when there is a total blockage of an artery which supplies the blood to the heart. The chances of survival after the attack happen to a person is very low and most cases death will be an ultimate result if exact first step measures are not taken immediately. Most of the people who are in heart failure die within five years of diagnosis. Thus, we had proposed AI techniques to analyze the patient information in this paper including natural language processing methods to obtain accuracy in prediction. We used AI and NLP techniques in which decision tree algorithm is used to extract information from unstructured data such as doctor's notes by analyzing the information and data that includes risk criteria or other types of symptoms. The aim of this paper is to predict whether a patient is likely to affect heart disease or not in early. This intelligent system decides the line of treatment to be followed by suitable databases obtained.

Keywords— Heart disease; Artificial Intelligence AI; Natural language processing NLP

I. INTRODUCTION

A paradigm shift in healthcare provided by Artificial intelligence (AI) which targets to copycat human perceptive function driven by increasing availability of healthcare data and speedy progress of analytics techniques. AI can be applied to various types of structured and unstructured healthcare data. A technique includes natural language processing for unstructured data [1]. Most of the disease prediction that uses AI devices includes cancer, neurology, and cardiology. The AI technique used in stroke are in three major areas, early detection, diagnosis and treatment, outcome prediction and prognosis evaluation.

The alterations in stress using AI have been discussed in the journal of cardiovascular research. AI uses sophisticated algorithms to 'learn' features to assist clinical practice. Based on feedback, improve its accuracy by equipping with learning and self-correcting abilities. By providing up-to-date medical information from journals, textbooks and clinical practices an AI system can assist doctors to inform proper patient care. In addition to this, it can help to reduce diagnostic and therapeutic errors in the human clinical practice [2].

Moreover, health risk alert and health outcome prediction assisted by making real-time inferences using an AI system extracts which gives useful information from a large patient population [3]. Healthcare data are generated from clinical activities, such as screening, diagnosis, treatment assignment. These types of medical data often present in, but it is not restricted to the form of clinical notes, electrical recordings from the medical instruments, laboratory findings, and physical diagnosis.

AI devices include natural language processing (NLP) methods which extract information from unstructured data such as hospital records of patients and medical journals to additive and raise structured medical data. Even though AI techniques are as powerful can be but they have to be encouraged for the problems faced by the clinic and can be used in the form of assisting hospital practice in the end [4]. The main work of NLP procedures is that they convert the written form of texts to the structured data in machine-readable format which can then be analyzed through ML techniques.

For the better understanding, the following flowchart of figure 1 defines the hospital data generation in the form of the route map to NLP data, where the route map begins and ends with the clinical activities [5].

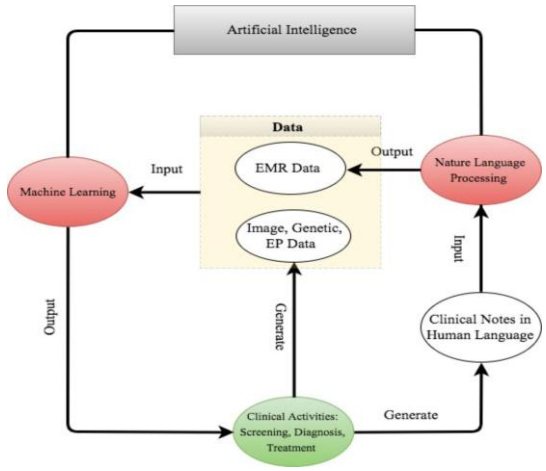

Figure 1. The route map from hospital data generation to NLP data.

This research work is organized as follows. Rest of the paper is organized as follows, Section I contains the introduction of AI and NLP Processing, Section II contains the related work of Natural Language Processing and the methods, section III explain the methodology with flow chart, Section IV describes results and discussion, Section V concludes research work and Section VI contain references to the research work.

II. NATURAL LANGUAGE PROCESSING

After proper preprocessing or quality control processes, the ML algorithms can be directly performed. The image, EP and genetic data are machine-understandable. However, large proportions of patients' information are in the narrative text, such as physical examination, lab reports, doctor's notes and discharge summaries, and are unstructured and unreadable for the computer program [6]. Thus, NLP extracts useful data from the narrative text to assist hospital decision making.

An NLP pipeline consists of two main components: (a) text processing and (b) classification. The NLP identifies a series of disease-relevant keywords in the hospital notes through text processing based on the historical databases. Then a dataset of the keywords is selected through examining their effects on the classification of the normal and abnormal cases. Then the validated keywords are entered and enrich the structured data to support hospital decision making [7].

The NLP pipelines have been developed to assist hospital decision making in arranging treatment arrangements, monitoring untimely effects and other data. Fiszman et al introduce NLP for reading the chest X-ray reports. This would assist the antibiotic assistant system which alerts doctors to the possible need for anti-infective therapy. The laboratory-based adverse effects are automatically monitored by NLP used by Miller et al. The NLP pipelines can help with diagnosing the diseases. Castro et al identified cerebral aneurysms disease associated variables through implementing NLP on the doctor notes. The resulting variables classify the normal patients and the patients with cerebral, accuracy rates on the training and validation samples.

III. AI APPLICATIONS IN STROKE

Stroke is a common and frequently occurring disease. It affects more than 500 million people. Therefore, research has great significance on prevention and treatment for stroke. Nowadays, more and more stroke-related studies have been used in AI techniques. The three main areas of stroke are early prediction and diagnosis, treatment, outcome prediction and prognosis evaluation.

Materials and Methods

A. Data

Data is important for risk prediction and further analysis. Most of the data available in an unstructured format it is also known as doctors notes which are unstructured. The doctor's report contains a rich and diverse source of information. Challenges for handling doctors report grammatical, short phrases, Abbreviations, Misspellings, Semi-structured information.

Unstructured patient data is also known as corpus. It is available from the informatics for combined biology and the bedside which is used for identifying the risk assessment. Both text and attributes are in XML data. These elements tell about patient information on present and past status as well as physical and hospital parameters. Data available from informatics for integrating biology & bedside some files already having CAD (abnormal) and remaining are normal.

Each data file contains patient details such as medication, laboratory results, medical history and personal information (age, weight) [8]. This analytics can be used in healthcare. Model of Health Care Analytics CAD risk parameters such as personal information, laboratory results and medical history taken from these data files using natural language processing toolkit.

B. Methods

This method is an automatic prediction of cardiac arrestment which obtains clinical examinations, physical diagnosis, medical history of the patient depending on age, cholesterol, Systolic Blood Pressure (BP), platelets (WBC, RBC).

Each parameter has some score points and numbers of points are based on a range of the parameter and again these are

different for men and women. The total number of points is determined by adding all points and the final risk score is obtained. The automatic Technique extracts the physical and doctor's parameters from a data file. It contains gender may be male and female age also represented either year Y.

Reynolds risk score can be determined using a computational formula for men and women. A 10-year cardiovascular disease for men can be estimated using equations for evaluating the risk for women.

Where B = 4.385× ln (age) + 2.607× ln (BP) + 0.963×ln (Total cholesterol) – 0.772 × ln (HDL) + 0.405 (if current smoker) + 0.102 × In (HSCRP) + 0.541 (Parental History).

Where B = 0.0799× (age) + 3.137× In (BP) + 1.382× In (Total cholesterol) – 1.172 × In (HDL) + 0.818 (if current smoker) + 0.180 ×ln (HSCRP) + 0.438 (Parental History).

Another simple method for calculating risk is 10-year prospective cardiovascular monster (PROCAM) study based on age, blood pressure, LDL cholesterol and HDL cholesterol and triglycerides.

All scores are categorized into three groups, low if it is 20%.It finds characteristic words and expressions of a text [9]. To identify and extract defined words and also use regular expressions to search, match and select the specific text present in unstructured data, NLTK is used.

IV. METHODOLOGY

A. Decision Tree

The most famous technique for diagnosis the cardiac disease is decision tree. It is very important to research different areas related to glitches by using accessible data to build a decision tree. According to the flowchart, each non-leaf node shows a test on a specific attribute. Each branch shows the result of the test. Each leaf node needs a class tag. A root node is an uppermost node. The research analysis for computing conditional probability is the utmost usage of a decision tree is in processes.

Decision Tree is easily understandable, simply interpret, perform well in huge dataset and knobs both category and numerical data. Computational efficacy is enhanced by the structures that convey supreme information chosen carefully for classification while other features are put off.

B. Data Source

The detection of heart disease using Decision Tree algorithm has been carried out by the experiments. The dataset contains 52 instances. Only 8 attributes are taken for experimental work such as age, chest pain, blood pressure, blood sugar that achieved, angina electrocardiogram.

SPSS has been used for calculation and analysis of data. This analysis, finds patterns, analysis and good prediction.

C. Data Set

Selection of data sets is very important. Based on the data sets, the experiments and results are obtained. The latest, accurate and clean data set could be obtained based on accuracy.210 instances are taken from the patient database.

Table 1. Dataset and Its Description

Attribute	Description
Age	30 t0 50 =1 , 51 to above = 2
Chest pain	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
Blood Pressure	Normal (80 to 120) =1 High (above 120) =2
Blood Sugar	False = 0, True = 1
ECG	0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophy
Heart Rate	Normal (60 to 100) =1 High (above 100) = 2
Angina	Yes = 1, No = 0
Class Attribute	
Disease	1: 50% diameter narrowing (Heart Disease) 0: No Heart Disease

From 210, 8 attributes are selected for experiments. The dataset contains 117 patients without heart disease and 92 patients with heart disease. Diagnosis class having value 1 with heart disease and value 0 with no heart disease. The selected attributes and their description are shown in Table 1.

D. Proposed Model

The proposed model gives finest result and perfections over previous models. The first step is the selection of data which is the data source. After sourcing, field option is used. A type field is selected which allows field metadata that is used to

determine and controlled. The modeling phase occurs in which algorithm C5.0 is selected [10]. It is to construct a predictive Decision Tree depends on our own choice and nature of data. After executing the predicted model, performance analysis is performed. Hence, the performance of the algorithm can be evaluated.

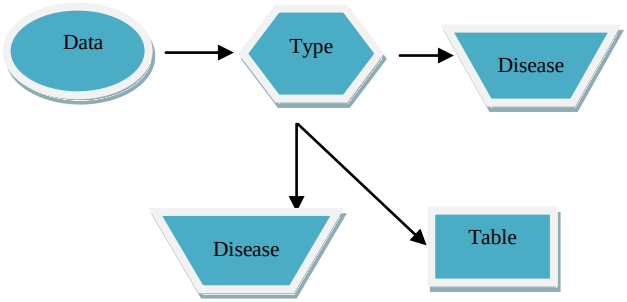

Figure 2. Classification model

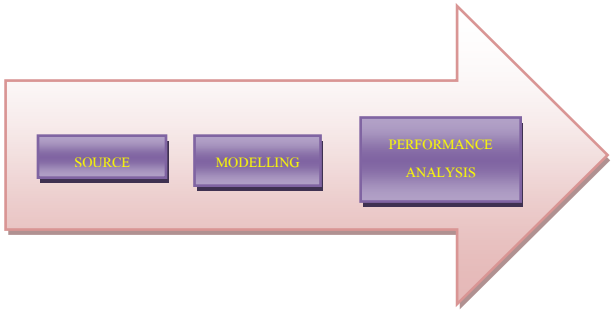

Figure 3. Block diagram of proposed model

V. RESULTS AND DISCUSSION

Experiments were made using Clementine tool. Data Set contains 210 Patients with 8 medical attributes. Attributes are in discrete form and the discrepancies are resolved among them. By using 8 attributes, decision Tree performs best having a good estimation probability of 79.9%. In total 8 attributes, 7 are numeric and the last two values 0 and 1 (0 means negative and 1 means positive disease) is my class attribute. The results of predicted disease are shown in a tree diagram. A class attribute has two child nodes which are exercise angina and chest pain which is maximum 2 depth tree.

The class attribute, Disease 1 shows the presence of heart disease and 0 shows the absence of heart disease. Initially, 92 persons are found to be infected by heart disease out of 209 patient's records. Rest of them took further tests and observations of patients having angina during exercise. Add exercise angina attribute has two nodes with value 1 and value 0. In node 1, there is a probability of having the heart disease is 83% while rest 17% are not the victims of heart disease. In node 2 out of 137 people, 23% individuals that are 32 people would be infected by heart disease based on the test of chest pain.

Those who have chest pain type typical angina and asymptomatic angina that is typical angina and non-anginal pain are 52 percent and 8 percent respectively. The remaining 105 people are not infected by the heart disease.

From the doctor's point of view, whenever a patient attends the clinic with chest pain, it is a common practice to carry out all the tests. But to reach the conclusion whether the patient has a heart disease or simply he suffers from the muscular pain it takes much time. It is also very costly to patients. With the help of classification tree, numbers of diagnostic tests are reduced which also helps to reduce cost significantly.

VI. CONCLUSION

The classification is the most widely used technique of data mining in the healthcare sector. In this research, the extensive classification method used for the prediction of heart disease is the decision tree. Sometimes poor observations lead towards death. The main purpose of this research is to diagnose the heart patients more precisely and more accurately with a minimum number of tests and reduction of attributes. Due to this effective technique applied in the diagnosis which helps in an earlier prediction of cardiac arrestment before the initial stage of a heart attack so through this prior warning, we can start our treatment immediately without time delay before the initial symptoms shown. Because of an earlier prediction, now the chances of survival have been increased, which reduces the death rate. This research plays a major role in the cost reduction of treatment, diagnose disease and additional enhancement of the medical studies. The purposed research work can further be boosted and expended for the prediction of various types of heart diseases.

VII. REFERENCES

[1] Kuldip Singh, Singh G, "*Alterations in Some Oxidative Stress Markers in Diabetic Nephropathy*", Journal of Cardiovascular Disease Research, Vol. 8, No. 1, pp.24-27, 2017.

[2] Nagre SW, "*Mobile Left Atrial Mass – Clot or Left Atrial Myxoma*", Journal of Cardiovascular Disease Research, Vol. 8, No. 1, pp.31-34, 2017

[3] Krishnamurthy VT, Venkatesh SA, "*Negative Pressure Pulmonary Oedema after Sedation in a Patient Undergoing Pacemaker Implantation*", Journal of Cardiovascular Disease Research, Vol. 8, No. 1, pp.28-30, 2017.

[4] Paryad E, Balasi LR, Kazemnejad E, Booraki S, "*Predictors Of Illness Perception In Patients Undergoing Coronary Artery Bypass Surgery*", Journal of Cardiovascular Disease Research, Vol. 8, No. 1, pp.16-18, 2017.

[5] Jiang F, Jiang Y, Zhi H, *et al*, "*Artificial intelligence in healthcare: past, present and future*", *Stroke and Vascular Neurology*, 2017.
[6] Dhuper S, Buddhe S, Patel S, "*Managing cardiovascular risk in overweight children and adolescents*", Pediatric Drugs, Vol.15, Issue. 3, pp.181-190, 2013.
[7] Dinarević S, Hasanbegović S, "*Problem of obesity in children and youth in Canton Sarajevo*", Pediatr Res, 68:1091, 2010.
[8] McNiece KL, Gupta-Malhotra M, Samuels J, Bell C, Garcia K, et al, National High Blood Pressure Education Program Working Group: "*Left ventricular hypertrophy in hypertensive adolescents: Analysis of risk by 2004 National High Blood Pressure Education Program Working Group staging criteria*", Hypertension, Vol. 50, No.2, pp.392-395, 2007.
[9] Torrance B, McGuire KA, Lewanczuk R, McGavock J "*Overweight, physical activity and high blood pressure in children: a review of the literature*", Vasc Health Risk Manag, Vol. 3, No. 1, pp.139-149, 2007.
[10] Jiber H, Blitti MC, Bouarhroum A, "*Acute type B Aortic Dissection Complicated by Acute Limb Ischemia: Case Report*", Journal of Cardiovascular Disease Research, Vol. 7, No.2, pp.97-99, 2016.
[11] Muhammad Subhi Al- Batah, "*Testing the probability of Heart Disease using Classification and Regression Tree Model*", Annual Research & Review in Biology, Vol. 4, Issue. 11, pp.1713-1725, 2014.
[12] Setty HS, Hebbal VP, Channabasappa YM, Jadhav S, Ravindranath KS, Patil SS, et al, "*Assessment of RV function following Percutaneous Transvenous Mitral Commissurotomy (PTMC) for rheumatic mitral stenosis*", Journal of Cardiovascular Disease Research, Vol. 7, No. 2, pp.58-63, 2016.
[13] Patnaik L, Pattanaik S, Sahu T, Panda BK, "*Awareness of symptoms and risk factors of Myocardial Infarction among adults seeking health care from a rural hospital of India*", Journal of Cardiovascular Disease Research, Vol. 7, No. 2, pp.83-85, 2016.

Authors Profile

*Ms. S. Sabeena B*achelor of Science from SriKrishna College, Coimbatore in 2012 and Master of Science from Avinashilingam University in year 2014. Master of Philosophy from Avinashilingam University in year 2015 and currently working as Assistant Professor in Departmet of Computer Applications, Pioneer College of Arts and Science, Coimbatore since 2017. She has published more than 2 research papaers in reputed International Journals including Scopus. Her main research work focuses on Feature Selection in Data Mining.

Ms. V. Sujitha pursuing Bachelor of Science from Pioneer College of Arts and Science, Affiliated to Bharathiar University, Coimbatore. She has published 2 research papers in reputed International Journal.

