DISEASE STATUS IDENTIFICATION SYSTEM IN ELECTRONIC HEALTH RECORDS USING MACHINE LEARNING

Prof. R. A. Deshmukh
Professor
Rajarshi Shahu College of Engineering
Maharashtra, INDIA
Pune - 411033
radesh19@gmail.com

Hrushikesh S. Aher
Undergraduate Student
Rajarshi Shahu College of Engineering
Maharashtra, INDIA
Pune – 411033
hrushi.aher@gmail.com

Vaishnavi R. Ahirrao
Undergraduate Student
Rajarshi Shahu College of Engineering
Maharashtra, INDIA
Pune - 411033
ahirrao.vaishnavi09@gmail.com

Vipul G. Depolkar
Undergraduate Student
Rajarshi Shahu College of Engineering
Maharashtra, INDIA
Pune - 411033
vipul.depolkar29@gmail.com

Mayur D. Bhole
Undergraduate Student
Rajarshi Shahu College of Engineering
Maharashtra, INDIA
Pune – 411033
mayurbhole605@gmail.com

Abstract

These days dependent on accessible clinical examination specialist is utilizing experimentation approach for anticipating infections. To foresee the infection is one of the real difficulties in past years and today too. Among the different maladies, heart disease is normal nowadays due to pushing the issue and remaining burden. On the basis of available symptoms and patients health, there is a great need of some system that predicts the diseases at an early stage. This research paper provides a survey of different data mining techniques which have already used for heart disease prediction and understand how efficiently Electronic Health Record (EHRs) offer clinicians functions that could never be achieved with written records. We are proposing the Disease status identification system by the Machine Learning algorithm. By using this algorithm we are going to predict Heart Disease based on their symptoms.

Keywords: ***Electronic health records system, KNN, Logistic Regression, Random Forest, Disease Prediction, Privacy, Security, and data mining.***

I. INTRODUCTION

The previous decades has seen a soaring increment of information in the Electronic Health Record (EHR) systems. Structured patient information such as demographics, disease history, and lab results, procedures and medications, and unstructured information such as progress notes and discharge notes are collected during each clinical encounter [1]. As EHRs have evolved, attempts have been made to improve the efficiency of electronic documentation. In recent studies, many promising results have generated using EHR and deep learning models to predict clinical events.

An EHR is an electronic version of a patient's medical records, which is maintained by the provider over time. It may include all of the key administrative clinical data which is relevant to that persons care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. The EHR framework computerizes access to data and can possibly streamline the clinician's work process. The EHR also have the ability to support other care-related activities directly or indirectly through various interfaces, including evidence-based decision support, quality management, and outcomes reporting [5]. EHRs are the next step in the continued progress of healthcare data that can strengthen the relationship between patients and clinicians. The data and the timeliness and availability of it will enable providers to make better decisions and provide better solution or care.

EHR selection for social insurance is consoling, it is imperative that information keep on having an auxiliary use for quality enhancement and research that enhances quiet consideration and possibly limit medicinal services costs [12]. Over the years, EHR data have been used with the intent to improve care, increase patient engagement, perform quality improvement, build shared models and standardization across institutions, create new knowledge, conduct research in a "real-world" settings instead of in controlled trials, enable public health surveillance and intervention, and facilitate personalized care and decision-making. The ultimate goal is to create a continually learning healthcare infrastructure with real-time knowledge production and create an ecosystem that is predictive, preventive, personalized, and participatory.

The implementation of EHR systems including techniques and evaluation, bioinformatics articles without significant clinical emphasis (genomics, genome-wide association studies), public policy, user-interface design, and medical imaging [11]. By utilizing EHRs, specific phenotypes within disease categories have been identified in conditions such as diabetes, heart disease, cancer, and neuropsychiatric disorders. EHRs have been utilized to retrospectively assess treatment effectiveness in real-world settings, quality of care and cost. Personal monitoring devices and social media may eventually be integrated into EHRs to enhance predictions.

Nowadays Heart disease is greatest reason for death. Manifestations like Blood weight, cholesterol, and heartbeat rate are the real explanation behind the heart disease. Some non-modifiable factors are there like smoking, drinking also the reason for heart disease [3]. As we have known that the heart is an operating system for the human body. If the function of the heart is not done properly means, it will affect other human body parts as well. Some risk factors of heart disease are Family history, High blood pressure, Cholesterol, Age, Poor diet, Smoking etc.. When blood vessels are overstretched, the risk level of the blood vessels is increased this leads to the blood pressure. Blood pressure is measured in two ways such as systolic and diastolic [2]. The systolic indicates the pressure in the arteries when the heart muscle contracts and diastolic indicates that the pressure in the arteries when the heart muscle is in resting state. The level of lipids or fats increased in the blood is causing the heart disease problem. The lipids are in the arteries hence the arteries become narrow and blood flow also becomes slow [6].

The Neural Network (NN) provides the minimized error of the heart disease prediction [7]. The patient records are classified in all above-mentioned techniques and predicted continuously. The patient activity is monitored continuously if there are any changes occur, and then various risk level of disease is informed to the patient and doctor. The doctors are able to predict heart diseases at an earlier stage because of a machine learning algorithm and with the help of computer technology. This paper provides a description about, KNN, Logistic Regression, Random Forest, Disease Prediction data mining technique used to predict heart diseases.

II. LITERATURE REVIEW:

Sr. No.	Title	Techniques	Dataset size or attribute	Result	Remarks
1	Intelligent Heart Disease Prediction System Using Data Mining Techniques [2].	Decision Tree, Naive Bayes, Neural Network	909 records with 15 attributes (Sex, Chest pain, Fasting blood sugar, Restecg, Exang, slope, Ca, Thal,old peak, age) **Tool:** Data Mining Extension (DMX), SSL-style.	They proposed IHDPS(Intelligent Heart Disease prediction System) using Decision Tree 80.4%, Naive Bayes - 86.12%, NN- 85%	It is used for only categorical data. Size of data is quite small. It can be enhanced or expanded using contagious data.
2	Classification using Convolutional Neural Network for Heart and Diabetics Datasets [4].	CNN algorithm	768 samples by 8 features **Tool:** Weka	Classification of heart and diabetics dataset using CNN with 80% accuracy.	Multiple convolution layers can be increased.
3	Heart Disease Prediction Using Data Mining Techniques [6].	KNN, Decision Tree, Naïve Bayes	909 records with 13 attributes (Sex, Pain kind, Abstinence blood glucose, Restack resting electrographic results, Exang exercise induced angina, ST, Ca,Thal, Trest, old peak, Age) **Tool:** Tomcat	-	Can be applied on larger real time data.
4	Human Heart Disease Prediction System using Data Mining Techniques [3].	KNN, ID3	87 attributes (Age, Gender, Blood pressure, Pulse rate and Cholesterol)	Detecting heart disease risk rate using KNN and ID3 with 80% accuracy	Accuracy is not more than 50 % because of general attribute. By reducing number of attribute accuracy can be increased.
5	Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees [1].	Decision Tree	1500 consecutive CHD subjects with 17 risk factor for 3 different criteria (Age, Sex, FH, SMBEF, HDL, LDL,TG, GLU,TC, SBP)	Predicting CHD (Coronary Heart Disease) with 82 % accuracy.	Help to reduce CHD morbidity and possibly, mortality
6	Deep Learning Approach for Prediction of Heart Disease Using Data mining Classification Algorithm Deep Belief Network [7].	Deep Belief Network.	300 records with 16 Attribute (Age, gender, blood pressure, cholesterol, Heredity, Blood sugar, PQ, ST, QT,QRS, R, Heart beat rate, BMI, smoking habit, Alcohol intake, Mental Stress) **Tool:** Matlab 8.1	Provide difference between CNN and DBN algorithm with 82% and 92% accuracy respectively.	Developed the system by using DBN which provides high accuracy than CNN.
7	Heart Attack Prediction Using Deep Learning [5].	RNN technique	303 records with 13 attributes (Age, Sex, CP, Trestbps, Chol, Fbs, Thalach, Old peak ST, slope, Ca, Thal, Class, Exang Continuous maximum heart rate achieved).	Provides 80% accuracy for heart attack data.	Heart Attack system using RNN algorithm. Another algorithm can be used for increasing accuracy.

8	Effective Heart Disease Prediction using Frequent Feature Selection Method [8].	Frequent Feature selection method	1000 records with 8 Attributes. (Age, Sex, Smoking, overweight, alcohol intake, exercise, cholesterol, Blood pressure, blood sugar, heart rate) **Tool:** Weka 3.6.0	Find significant pattern for heart disease prediction using MAFIA algorithm.	Prediction accuracy can be increased by using various classification techniques.
9	Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems [9].	ANN, Decision Tree, Naïve Bayes	900 records with 13 attributes	-	Small number of attributes. Continuous more research required to generate Knowledge rich Health Care environment

Table 1. Literature Review

III. PROPOSED SYSTEM

In the proposed system decision making from the electronic health record is going to be done by using Machine Learning algorithm for predicting Heart Disease. First, we are going to identify or study on heart disease patients data and then decide which factors are common in same heart disease patient. Here we are using CSV file which is containing 703 records with 14 attributes. In this system the patient's data is updated by the particular doctor who is responsible for the diagnosis of the patient and the patient can share his/her records with the other doctors in emergency cases, also the disease analysis from the EHR are going to be done.

A. Architecture

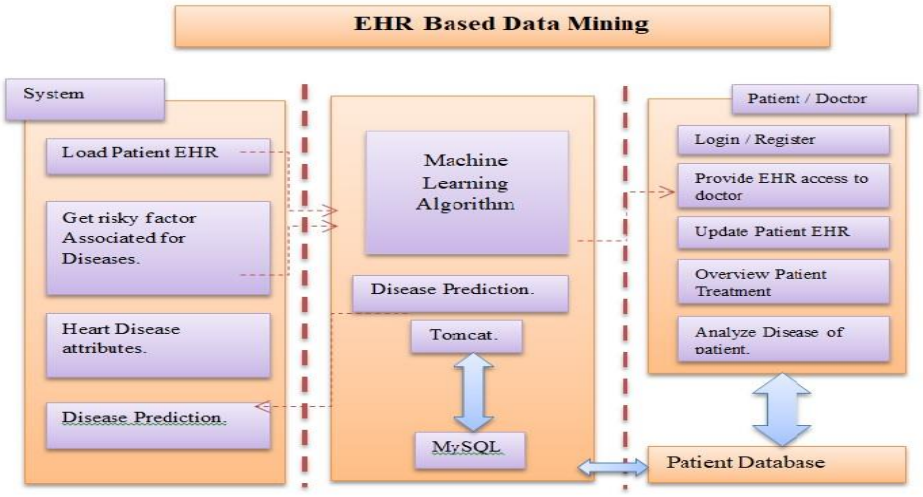

Fig. 1 Architecture of Proposed System.

EHR system is the next step in the continued progress of healthcare that can strengthen the relationship between patients and clinicians. The data and the timeliness and availability of it will enable providers to make better decisions and provide better solution and care.

For example, the EHR system can improve patient care by:

- Reducing the incidence of medical error by improving the accuracy and clarity of medical data.
- Making the health information's available, reducing duplication of tests, reducing delays in treatment, and patients well informed to make better decisions.

- Reducing medical error by improving the accuracy and clarity of medical data.

Working flow of the system is as follows:

1. Load Patient Data
2. Extract the Risky Factor associated with Heart Disease
3. Predicting Heart Disease
4. Apply machine learning algorithm.
5. Share result and data.

IV. ADVANTAGES

1. Helps in analyzing patient details and predict.
2. A patient can log on to his own records and see the trend of the lab results over the last year, which can help to motivate him to take his medications and keep up with the lifestyle changes that have improved the numbers.
3. Helps to predict disease based on basic parameters.

V. APPLICATIONS

1. Smart disease level detection system.
2. Disease Prediction System.

VI. EXPERIMENTAL RESULT

At first, we have connected three fundamental information mining order procedures, for example, KNN, calculated relapse, and Random Forest. These calculations are executed in python on 703 records of coronary illness dataset with 14 traits, for example age, sex, cp (Chest Pain), trestbps (Resting blood pressure), chol (Cholesterol), fbs (Fasting Blood Pressure), restecg (Resting electrocardiographic results), thalach (Maximum Heart rate), exang (Exercise-induced angina), oldpeak (ST depression induced by exercise relative to rest), slope (The slope of the peak exercise ST statement), ca (No. of measured vessels colored by flourosopy), thal , diagnosis (Result ranging from 0 to 4). Execution of actualized calculations is looked at on premise of their precision.

A. KNN (K-nearest neighbor algorithm):

K nearest neighbors is a straightforward calculation that stores every accessible case and orders new cases dependent on a likeness measure i.e. distance function. We have utilized 80% information for training and 20% information for testing and determined accuracy, precision, recall, f1-score and confusion matrix as as appeared table 2.

B. Logistic Regression:

Logistic regression is the most commonly used algorithm after linear regression. It is generally used for classification of data by using a sigmoid function. Same as KNN accuracy, precision, recall, f1-score and confusion matrix are determined.

C. Random Forest:

Random Forest is a supervised learning algorithm it creates the forest and makes it somehow random. This algorithm can be used for classification and regression.

Sr. No.	Algorithm	Accuracy	Precision	Recall	F1-Score
1	KNN	64.52	0.57	0.65	0.60
2	Logistic Regression	67.74	0.65	0.68	0.66
3	Random Forest	67.74	0.57	0.68	0.61

Table 2. Experimental Results of algorithm.

A. Confusion Matrix:

1. KNN (K-Nearest Neighbor):

CM	0	1	2	3	4
0	17	0	1	0	0
1	2	2	0	1	0
2	0	1	1	0	0
3	1	1	2	0	0
4	0	1	1	0	0

2. Logistic Regression:

CM	0	1	2	3	4
0	18	0	0	0	0
1	2	2	0	1	0
2	0	1	0	1	0
3	0	0	1	1	2
4	0	1	1	0	0

3. Random Forest:

CM	0	1	2	3	4
0	18	0	0	0	0
1	2	2	0	1	0
2	0	1	1	0	0
3	2	1	1	0	0
4	1	0	0	1	0

VII. CONCLUSION

In this paper, we have examined different existing methods for predicting heart diseases with the help of data mining. Many Access control methods and security permits EHRs system to protect sensitive patient information. The improvement and use of big data analysis methods on EHRs may help make a persistently learning EHR ecosystem. This survey is to briefly introduce the EHR system and its important factors like privacy and security with its benefits. From this study, we got the knowledge about how to apply data mining technique to predict heart disease. We applied KNN, Logistic regression and Random forest for comparison of their accuracy. For increasing accuracy of an existing system, we can use deep learning algorithm for prediction of Heart Diseases on a large dataset.

REFERENCES

[1] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE

[2] Sellappan Palaniappan, Rafiah Awangr, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, 2008 IEEE.

[3] "Theresa Prince. R and J. Thomas", Human Heart Disease Prediction System using Data Mining Techniques. IEEE 2016.

[4] “Tharani.S1, Dr. C. Yamini2” Classification using Convolutional Neural Network for Heart and Diabetics Datasets, IJARCCE

[5] Abhay Kishore1, Ajay Kumar2, Karan Singh3, Maninder Punia4, Yogita Hambir5 “Heart Attack Prediction Using Deep Learning”, IRJET

[6] “Abhishek Rairikar, Vedant Kulkarni, Vikas Sabale, Harshavardhan Kale” HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES, I2C2

[7] “Dr. T. Karthikeyan, V.A.Kanimozhi” Deep Learning Approach for Prediction of Heart Disease Using Data mining Classification Algorithm Deep Belief Network, IJARSET

[8] “S.Saravanakumar1, S.Rinesh2”, Effective Heart Disease Prediction using Frequent Feature Selection Method, IJIRCCE

[9] “ Eman AbuKhousa1, Piers Campbell2”, Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems, IIT

[10] Sunita Soni, Jyoti Soni,Ujma Ansari,Dipesh Sharma, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, International J journal of Computer Application (IJCA, 0975 – 8887) Volume 17 – No.8, March 2011.

[11] Casey Bennett, and Thomas W. Doub, “Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice” 6th International Conference on Data Mining 2010.

[12] Jingshu Liu, Zachariah Zhang, Narges Razavian. “Deep EHR: Chronic Disease Prediction Using Medical Notes” arXiv:1808.04928v1 [cs.LG] 15 Aug 2018.

[13] Monika Gandhi and Dr. Shailendra Narayan Singh “Predictions in Heart Disease Using Techniques of Data Mining,” International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE) 2015 IEEE.