iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43278

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Automating Medical Data and Using Data Science for Heart Disease Prediction.

Ritikesh Bhatt[1], Toufiq Shaikh[2], Dr. Sandeep Patil[3], Mohnish Harwani[4], Bibhu Kumar[5]
[1, 2, 3, 4, 5]*Department of Computer Engineering, International Institute of Information Technology Pune, India*

Abstract: ***Technology has aided the improvement of individual health, healthcare, biomedical research as well as public health. Therefore, healthcare institutions are seeking to develop integrated information-management environments to consolidate the inevitable application of big data to health care. There exist various entry points into the medical world where computational tools assist patient care matters; reporting results of tests, allowing direct entry of orders or patient information by clinicians, facilitating access to transcribed reports, and in some cases supporting telemedicine applications, because of disorganized and incomplete patient records pose an obstacle to patient care. The most common medium by which records of medical history are kept is paper making data management a severe impediment to productivity. However, the promise of a more efficient healthcare service is obvious through the use of automated health records management systems. Heart disease is a common disease that is overlooked by most. In this study, we discuss how a person can figure out if they need to go to a doctor for a health check-up for any heart-related issues using machine learning algorithms.***

Keywords: Data Science, Statistics, Python, Data mining, Machine learning, Analytics, Big Data, Disease Prediction, Firebase, Supervised Learning, Unsupervised Learning, ElectrocardioGram(ECG).

I. INTRODUCTION

Automated health records may be alien to quite a several health care facilities and It is observed that medical records are being kept in physical folders and then placed on cabinets. It was also noted from the record section that patients' (and staff) records are stored in many formats, and this results in disorganization of the records, leading to a higher risk of medical errors, duplicate procedures, and time loss in searching and obtaining information. Though a paper-based records system provides accountability, it has its inherent challenges e.g., difficulty of access, time-consuming to update, no data security and it is very difficult to share between different locations and maintain for a very long time without destruction. Natural disasters like a fire outbreak or damage by water could lead to the permanent loss of years of medical records and also, forgetting that the records even exist.

A fully automatic digitized health records system could tackle and improve on many of these problems also provide accountability, increase privacy levels since only authorized personnel can access a patient's record as opposed to a paper file, and increase efficiency; such that patient data could be retrieved within seconds instead of going through multiple files and provide access to generic patient data.

The proposed hospital management system handles the patient's experience from start to finish. The system receives input, stores-updates data, and outputs the required information. The primary user targets hosted on the management system are record officers (front desk), patients, doctors, pharmacists, laboratory attendants, and account personnel, all of whom have different role assignments, functional requirements, and varying interaction levels.

The patients are the primary clients of the health centre. The registration interface allows the patient to create a new record if they are first-time users. This section is used for registration where all the patients' data are collected and saved in the database. However, this would need approval from the recording officer. Also, the recording officer can create a record for patients who cannot use the self-service. The patient can also independently log in to the system to view reports.

In the next section after the patient has made appointments from the record section. The doctor selects the patient from the queue, and the doctor would be able to view the entire medical history of the patient.

We are living in the big data era where all healthcare departments generate massive amounts of data using machine learning and data science methods, we can help in utilizing these data.

To present an overview of current machine learning methods and to use it in medical research, focusing on select machine learning techniques, best practices, and deep learning, this model can be used to predict serious future heart diseases.

Therefore, the purpose of this research project is to develop a smart and secured automated healthcare record management system using data science and using different aspects of data science and learning methods to predict future health adversaries.

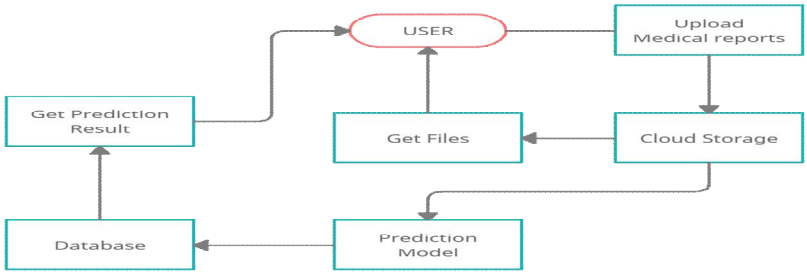


Fig1: Data Flow diagram for the model and storage

II. RELATED WORK

Several studies on heart disease have been done on heart disease prediction by Anjan Nikhil Repaka, Sai Deepak Ravikanti, and Ramya G Franklin [1] (2015) where they used different algorithms for the purpose but found Naive Bayesian the most helpful. A new convolutional neural network-based multimodal disease risk prediction algorithm using structured and unstructured data from hospitals was used by Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. The structured data that is usually taken and coupled with unstructured data, is used for disease risk prediction [2]. Extensive research efforts were made to identify those studies that applied more than one supervised machine learning algorithm to a single disease prediction. Two databases were searched for different types of search items. Thus, 48 articles were selected for the testing of which the best suited SVM algorithm by Shahadat Uddin, Arif Khan Md Ekramul Hossain, and Mohammad Ali Moni [3].

Furthermore, Manab Kumar Das and Samit Ari [8] took ECG data from the MIT-BIH arrhythmia database is used it for the performance evaluation of the proposed ECG beat classification technique. The proposed classification techniques consist of three main stages, i.e., pre-processing and QRS detection, feature extraction, and classifier. An MLP-NN is trained with the error backpropagation algorithm.

III.MAIN TITLE

A. Data Pre-processing

A patient uploads the data to the website to interpret the chances of any kind of heart disease. The data contains information like age, gender, blood pressure, height, chest pain, hypertension, smoking history, diabetes type, chest pain type, dyslipidaemia, random blood sugar, low and high-density lipoprotein, cholesterol, etc. The data must be standardized before it is fed to the machine learning algorithm. Information like gender, smoking history, chest pain type, hypertension, etc. must go through a label encoder whereas age, height, and other categories must go through a standard scaling of the data. Here's the formula for standard scaling:

$$X' = \frac{X - \mu}{\sigma}$$

B. Primary Prediction

This first part of the prediction lets the patient know whether they are prone to heart disease or not. We take several algorithms to train the data that is taken from the UCI dataset, which contains nearly 75 attributes. The algorithms take the provided variables from the user and train the machine accordingly and give out results in a probabilistic form to tell if the person is at risk of heart disease or not. Here are some of the algorithms that are tested:

Algorithm	*Accuracy*
Logistic Regression	81
Artificial Neural Network	86.04
Decision Tree	95.4

Random Forest	93.84
Naive Bayesian	90.3
Support Vector Machine	86.62
k-Nearest Neighbour	82.4

Table: Algorithms with accuracies for the first part

Decision Tree and Random Forest seem to be the most accurate, so, we use them in tandem because users don't input all 75 attributes but maybe just a few of them at a time. For that matter, we train all the models again using the attributes that the user provides in the first place and then figure out which algorithm to go with.

C. Evaluation

To evaluate performance in the experiments, firstly, TPs, FPs, TNs, and FNs were measured as true positives (instances predicted correctly as per requirement), false positives (number of instances predicted incorrectly as per requirement), true negatives (number of cases predicted correctly as per requirement), and false negatives (instances incorrectly predicted as not needed). The accuracy can be measured by:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

D. ECG Classification

We collect our ECG data in an image format. With the help of the segmentation of data, we can divide the data into smaller segments. Each beat is passed through the trained CNN model.

We use the MITDB, INCARTDB, and SVDB datasets to train the machine to be familiar with all kinds of beats. We are classifying the beats into 5 categories: Normal (n), SVEB (supraventricular ectopic beat), VEB (ventricular ectopic beat), Fusion, and unclassified beats. The most commonly used descriptors in the literature for ECG classification are based on these intervals. For singling out these beats, we separate each beat using the r-r intervals. We find the r-peak for every beat and then take 0.3 seconds before and 0.5 seconds after the r-peak. After every beat is separated, The CNN algorithm is applied.

This special case of ECG differs from the usual cases like the one used before where CNNs can be used.

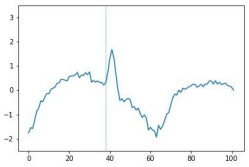


Fig: SVEB

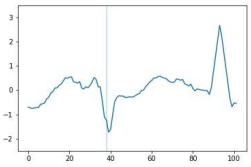


Fig: VEB

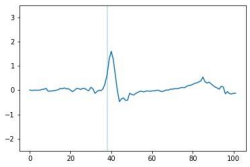


Fig: Normal beat

In this case, unlike a time series that mainly uses one-dimensional data, it is always expressed using some color channel as two-dimensional data because we are using image recognition of every beat. It provides high accuracy in the classification problem. In this case, we use a 1D CNN as the following architecture. 7 convolution layers with filter width 5 and 128 neurons + max pooling and dropout after each layer + globalAveragePooling + FCN layer with (256/128/64) neurons + dropout each layer with 5 outputs + After layer softmax.

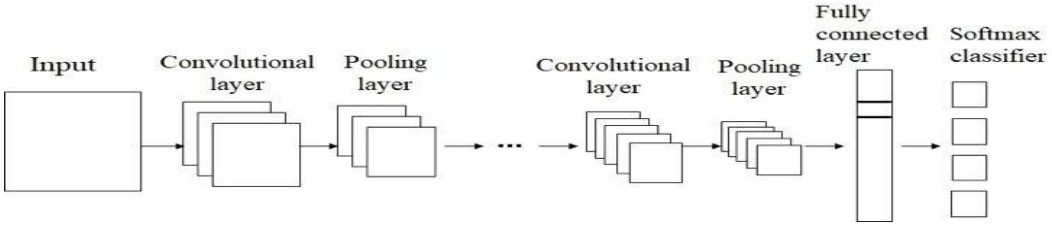


Fig: Architecture of CNN

After training and the validation process are done, in the testing process, we find the accuracy of our CNN model is around 68% for the classification of the type of beats after the testing phase of the model. The model is then exported to be used in an API for the web app and delivers the information further to the eligible personnel. This might help in figuring out what type of disease the patient might be dealing with.

IV. RESULTS AND FINDINGS

Our system puts together a whole process for a user to collect their document in one place and therefore keep track of their medical records at all times. It's convenient even for the doctors who can access the history of the patient's medical records.

When it comes to heart disease prediction, we set up a process where users can test whether they require examination or not. The first process gives you an idea based on the data that you can enter out of the 75 attributes given whether you need to seek some medical help or not. For this process, we would preferably use a Decision Tree or Random Forest based on the attributes provided by the user. If the patient does need further help, they can get an ECG done and with the help of the CNN algorithm we can figure out what kind of arrhythmia the patient might be suffering from. This will help the patient seek out the right medical help.

V. CONCLUSION AND FUTURE WORKS

The advancements in technology and wireless communications and the rise of mobile technology such as wearable sensor technology open up the opportunity for real-time healthcare system access, at the touch of a finger. In this study real-time vital signs uploading healthcare data has been proposed and further be used to predict the heart complexities. Our medical system and infrastructure have been in a very hostile situation, and it was heavily affected during this pandemic, unavailability of medical databases and records has been reported all around the

Country. The system is adaptable and can upload data at ease. Our System proposes a web-based application where a patient can store all their medical data including prescriptions, reports, and bills from day one of their medical treatment. We will make use of the latest technology to create our web application using NodeJs and hosting on Firebase (Google Cloud platform). Using data science on stored data we will analyse and predict if the person can suffer any heavy medical problems in the future. The main advantage of the application is the availability of medically issued records that can be used for future references and research purposes for both the

patients and the doctor. The future of automation of medical reports will see the adaptation of a few things like the data accumulated by the device can be further used to predict heart disease as well.

REFERENCES

[1] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin.: Design And Implementing Heart Disease Prediction Using Naives Bayesian, In: Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8 Min Chen,, Yixue Hao, Kai Hwang, Lu Wang and Lin Wang.: Disease Prediction by Machine Learning Over Big Data From Healthcare Communities, In: National Natural Science Foundation of China under Grant 61572220 and Grant 81671904, in part by the International Science and Technology Cooperation Program of Chinese Ministry of Science and Technology under Grant S2014ZR0340.

[2] Shahadat Uddin, Arif Khan Md Ekramul Hossain and Mohammad Ali Moni.: Comparing different supervised machine learning algorithms for disease prediction, In: BMC Med Inform Decis Mak 19, 281 (2019). Springer Nature

[3] Saroj Kumar Pandeya, Rekh Ram Janghelb and Vyom Vani.: Patient Specific Machine Learning Models for ECG Signal Classification, In: International Conference on Computational Intelligence and Data Science (ICCIDS 2019). ScienceDirect

[4] Sahil Dalala and Virendra P. Vishwakarma.: GA-based KELM Optimization for ECG Classification, In: International Conference on Computational Intelligence and Data Science (ICCIDS 2019). ScienceDirect

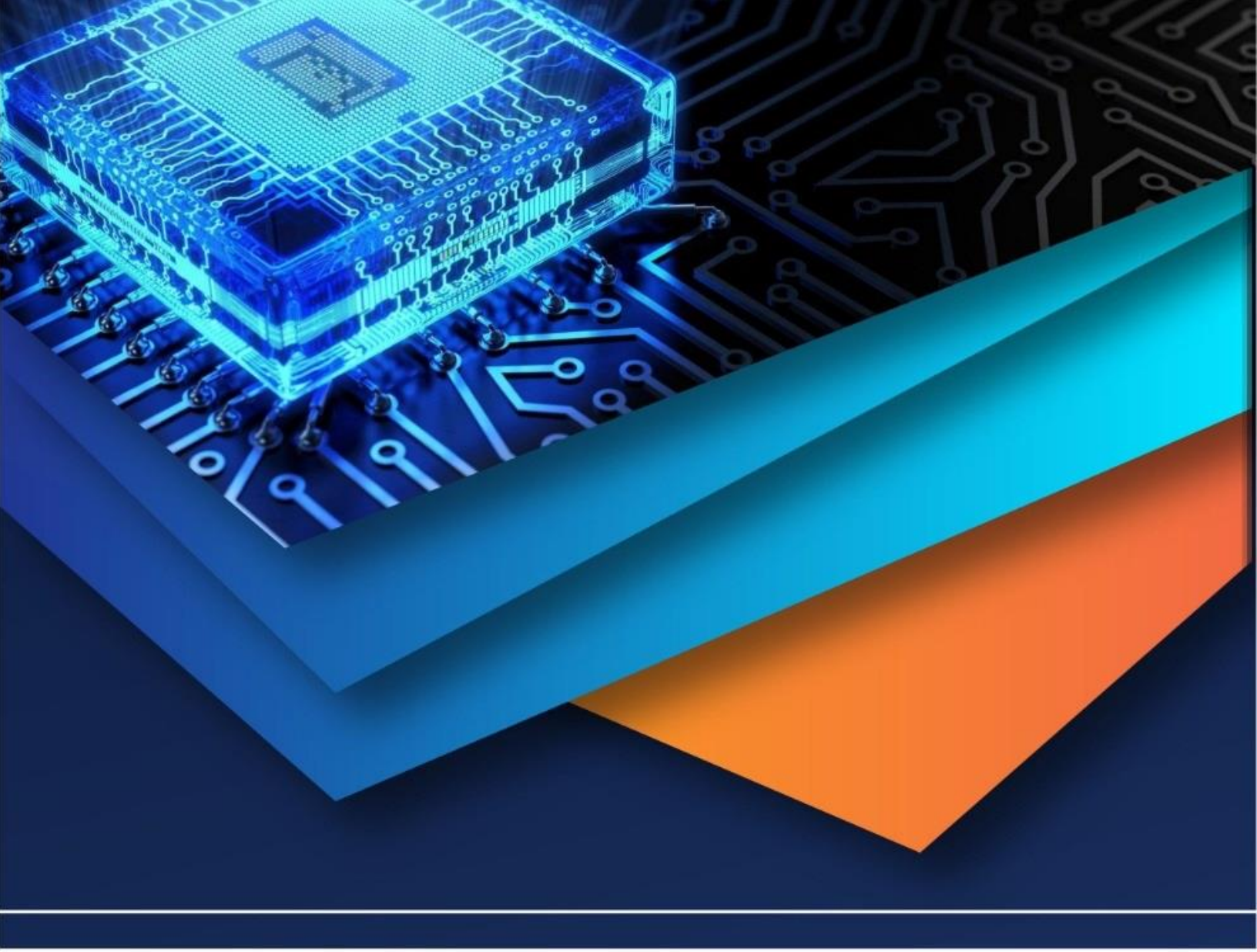
[5] Lucie Maršánová, Marina Ronzhina, Radovan Smíšek, Martin Vítek, Andrea Němcová, Lukas Smital & Marie Nováková.: ECG features and methods for automatic classification of ventricular premature and ischemic heartbeats: A comprehensive experimental study, In: Springer Nature 2017

[6] Manishaa, Sanjeev Kr. Dhulla and Krishna Kant Singh.: ECG Beat Classifiers: A Journey from ANN To DNN, In: International Conference on Computational Intelligence and Data Science (ICCIDS 2019). ScienceDirect

[7] Manab Kumar Das and Samit Ari.: ECG Beats Classification Using Mixture of Features, In: International Scholarly Research Notices Volume 2014, Article ID 178436, 12 pages

[8] Blaž Zupan, Janez Demšar, Michael W. Kattan, J. Robert Beck, I. Bratko.: Machine learning for survival analysis: a case study on recurrence of prostate cancer, In: Artificial Intelligence in Medicine 20 (2000) 59–75. Elsevier.

[9] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava.: Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, In: Digital Object Identifier 10.1109/ACCESS.2019.2923707. IEEE Access



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY