Research Article

Heart Attack Prediction using SMLT

Aloyses. E*[1], Vidya Bhaskar. C[1], Soniya Priyatharsini. G[2], Ramesh Babu. V[2]
1. Final year, Dept of CSE, DR.M.G.R Educational and Research Institute, Tamilnadu, India.
2. Associate Professor, Department of CSE, DR.M.G.R Educational and Research Institute, Tamilnadu, India.
***Corresponding author's E-mail:** aloysiesesudoss123@gmail.com

Received: 18-01-2022; **Revised:** 25-03-2022; **Accepted:** 04-04-2022; **Published on:** 15-04-2022.

ABSTRACT

Heart is the main organ that pumps blood to the body and for proper functioning of body. Heart disease is a fatal human disease increasing globally in both developed and undeveloped countries and consequently, causes death. Normally, in this disease, the heart fails to supply a sufficient amount of blood to other parts of the body in order to accomplish their normal functionalities. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyse huge complex medical data, helping healthcare professionals to predict heart disease. The proposed method is to build a machine learning model capable of classifying the person has heart disease or not. Different algorithms are compared and the best model is used for predicting the outcome.

Keywords: Machine Learning, LVAD, Artificial intelligence, Data, Heart.

QUICK RESPONSE CODE →

DOI:
10.47583/ijpsrr.2022.v73i02.035

DOI link: http://dx.doi.org/10.47583/ijpsrr.2022.v73i02.035

INTRODUCTION

Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux. The term "data science" was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market. Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

Artificial intelligence

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals. Leading AI textbooks define the field as the study of "intelligent agents" any system that perceives its environment and takes actions that maximize its chance of achieving its goals. Some popular accounts use the term "artificial intelligence" to describe machines that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving", however this definition is rejected by major AI researchers. Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, and speech recognition and machine vision. AI programming focuses on three cognitive skills: learning, reasoning and self-correction.

Related Work

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its

ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

Review of Literature Survey

Title : A Survey on Prediction Techniques of Heart Disease using Machine Learning

Author: Mangesh Limbitote, Dnyaneshwari Mahajan, Kedar Damkondwar, Pushkar Patil

Year : 2020

Heart is one of the most important part of the body. It helps to purify and circulate blood to all parts of the body. Most number of deaths in the world are due to Heart Diseases. Some symptoms like chest pain, faster heartbeat, discomfort in breathing are recorded. This data is analysed on regular basis. In this review, an overview of the heart disease and its current procedures is firstly introduced. Furthermore, an in-depth analysis of the most relevant machine learning techniques available on the literature for heart disease prediction is briefly elaborated. Heart disease is the kind of disease which can cause the death. Each year too many peoples are dying due to heart disease. Heart disease can be occurred due to the weakening of heart muscle. Also, the heart failure can be described as the failure of heart to pump the blood. Heart disease is also called as coronary artery disease (CAD). CAD can be occurred due to insufficient blood supply to arteries.

Existing System

The Acoustic analysis has the potential to be a relatively low cost and non-invasive diagnostic tool for point-of-care assessment and remote monitoring of LVAD recipients. Prior work on acoustic analysis in LVAD recipients has focused on analysis of LVAD sounds to detect pump thrombosis. The analysis of heart sounds in LVAD recipients has been limited by the dominant LVAD sounds present within their precordial sounds. We have developed a novel signal processing pipeline to mitigate LVAD-generated sounds within precordial sounds recorded from LVAD recipients, potentially rendering automated heart sound analysis in this population feasible. Their analysis revealed that pump activity in LVAD recipients induces heart sounds with diverse acoustic signatures, such as S1 and S2 sounds with varying frequency ranges and relative amplitudes, and variabilities in S2 components based on changes in aortic valve mobility across different pump speeds.

Proposed System

The proposed method is to build a machine learning model for classification of heart disease. The process carries from data collection where the past data related to heart disease are collected. Data mining is a commonly used technique for processing enormous data in the healthcare domain. The heart disease if found before proper treatment can save lives. Machine learning is now applied and mostly used in health care where it reduces the manual effort and better model makes error less which leads in saving the life. The data analysis is done on the dataset proper variable identification done that is both the dependent variables and independent variables are found. Then proper machine learning algorithm are applied on the dataset where the pattern of data is learnt. After applying different algorithms a better algorithm is used for the prediction of outcome.

Block Diagrams

1. Architecture Diagram

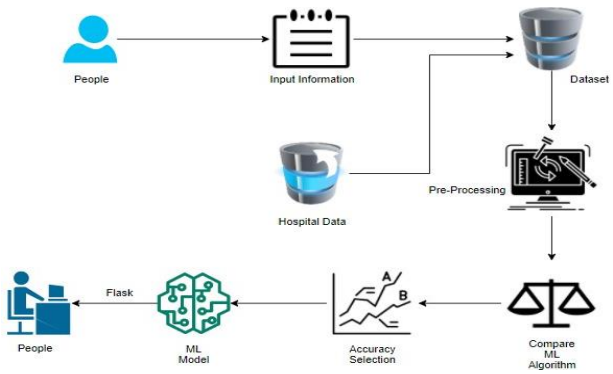

Figure 1: Architecture diagram

2. Use Case Diagram

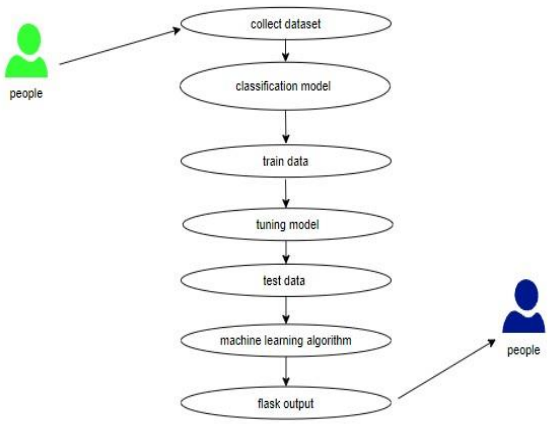

Figure 2: Use case diagram

3. Class Diagram

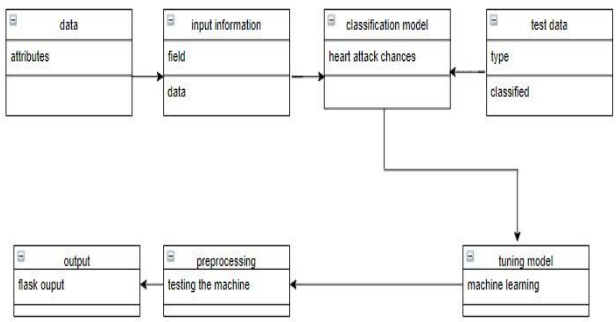

Figure 3: Class diagram

4. *Activity Diagram*

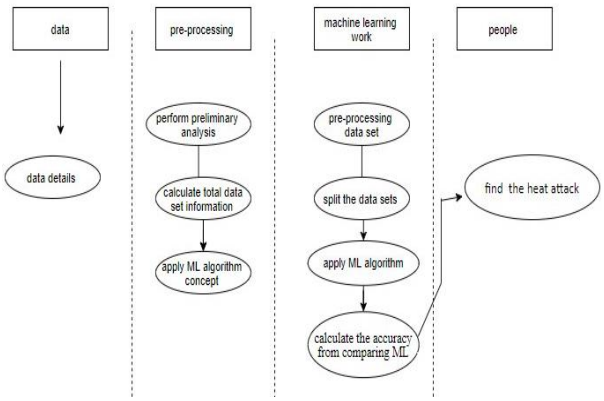

Figure 4: Activity diagram

RESULTS AND DISCUSSION

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different **data cleaning** tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, **missing values** and it able to **more quickly clean data**. It wants to **spend less time cleaning data**, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning models and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process.

The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Exploration data analysis of visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing

is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

6. a)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	2
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	1
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	2
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	1
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	1

b)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.094401	0.096920	0.273053	0.220056	0.123458	0.128171	-0.402215	0.098297	0.194234	0.159774	0.356081	0.106100	0.212322
sex	-0.094401	1.000000	0.034636	-0.062693	-0.201647	0.042140	0.039253	-0.076101	0.180022	0.097412	0.050545	0.086830	0.391046	0.297721
cp	0.096920	0.034636	1.000000	-0.043196	0.090465	-0.098537	0.074325	-0.317682	0.353160	0.167244	0.136900	0.225890	0.262659	0.417436
trestbps	0.273053	-0.062693	-0.043196	1.000000	0.173019	0.155681	0.116157	-0.039136	0.082793	0.222800	0.142472	0.085697	0.132045	0.155383
chol	0.220056	-0.201647	0.090465	0.173019	1.000000	0.025186	0.167652	-0.018739	0.078243	0.027709	-0.005755	0.126541	0.028836	0.118021
fbs	0.123458	0.042140	-0.098537	0.155681	0.025186	1.000000	0.053499	0.022494	-0.004107	-0.025538	0.044076	0.123774	0.049237	-0.016319
restecg	0.128171	0.039253	0.074325	0.116157	0.167652	0.053499	1.000000	-0.074628	0.095098	0.120034	0.160614	0.114368	0.007337	0.182091
thalach	-0.402215	-0.076101	-0.317682	-0.039136	-0.018739	0.022494	-0.074628	1.000000	-0.380719	-0.349045	-0.386847	-0.265333	-0.253397	-0.418514
exang	0.098297	0.180022	0.353160	0.082793	0.078243	-0.004107	0.095098	-0.380719	1.000000	0.274672	0.255908	0.153347	0.321449	0.419303
oldpeak	0.194234	0.097412	0.167244	0.222800	0.027709	-0.025538	0.120034	-0.349045	0.274672	1.000000	0.609712	0.255005	0.324333	0.417967
slope	0.159774	0.050545	0.136900	0.142472	-0.005755	0.044076	0.160614	-0.386847	0.255908	0.609712	1.000000	0.109498	0.283678	0.337616
ca	0.356081	0.086830	0.225890	0.085697	0.126541	0.123774	0.114368	-0.265333	0.153347	0.255005	0.109498	1.000000	0.255648	0.455336
thal	0.106100	0.391046	0.262659	0.132045	0.028836	0.049237	0.007337	-0.253397	0.321449	0.324333	0.283678	0.255648	1.000000	0.525020
target	0.212322	0.297721	0.417436	0.155383	0.118021	-0.016319	0.182091	-0.418514	0.419303	0.417967	0.337616	0.455336	0.525020	1.000000

Figure 6 (a& b): Data sets

Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below 4 different algorithms are compared:

- Logistic Regression
- Random Forest
- Decision Tree Classifier
- Naive Bayes

The K-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

Prediction result by accuracy

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

***False Positives (FP)*:** A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

***False Negatives (FN)*:** A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

***True Positives (TP)*:** A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

***True Negatives (TN)*:** A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

True Positive Rate (TPR) = TP / (TP + FN)

False Positive rate (FPR) = FP / (FP + TN)

***Accuracy*:** The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

***Precision*:** The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labelled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

***Recall*:** The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

***General Formula*:**

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Prediction of Heart Attack.

REFERENCES

1. Agrawal S et al., "Thirty-day readmissions after left ventricular assist device implantation in the United States," Circ. Heart Fail. 2018;11(3):P.e004628 doi: 10.1016/j.cardfail.2020.03.001 PMID: 32151635

2. Akhter S A et al., "Hospital readmissions after continuous-flow left ventricular assist device implantation: incidence, causes, and cost analysis," Ann. Thorac. Surg.

2015;100(3):884-9 doi: 10.1097/MAT.0000000000000684 PMID: 29095732

3. Mehra M R et al., "Healthcare resource use and cost implications in the MOMENTUM 3 long-term outcome study," Circulation. 2018;138(18):1923-1934

https://doi.org/10.1161/CIRCULATIONAHA.118.035722

4. Gordon R J et al., "Prospective, multicenter study of ventricular assist device infections," Circulation. 2013;127(6):691-702.
doi: 10.1097/MAT.0000000000000684 PMID: 29095732

5. Patel P et al., "Acoustic signatures of left ventricular assist device thrombosis," ASME J. Med. Diagn. 2019;2(2):024501.

6. Feldmann C et al., "An acoustic method for systematic ventricular assist device thrombus evaluation with a novel artificial thrombus model," J. Thorac. Dis. 2018;10(15):s1711-s1719.

doi:10.21037/jtd.2018.04.11

7. Reiss N et al., "Telemonitoring and Medical Care of Heart Failure Patients Supported by Left Ventricular Assist Devices - The Medolution Project," Stud. Health. Technol. Inform. 2017;236:267-274.

8. Feldman D et al., "The 2013 International Society for Heart and Lung Transplantation Guidelines for mechanical circulatory support: executive summary," J Heart Lung Transplant. 2013;32(2):157-87.
doi: 10.1016/j.healun.2012.09.013 PMID: 23352391

9. Sabbah H N and P. D. Stein, "Investigation of the theory and mechanism of the origin of the second heart sound," Circ. 1976,139(6):874-882.

10. Burch G E, "The Human Heart as a Pump," in Heart Perfusion, Energetics, and Ischemia, vol. 62, L. Dintenfass, D. G. Julian, and G. V. F. Seaman Eds., (NATO Advanced Science Institutes Series. Boston, MA, USA: Springer, 1983, 62: 81-95. doi:10.1007/978-1-4757-0393-1

11. Hodkiewicz M R and Norton M P, "The effect of change in flow rate on the vibration of double-suction centrifugal pumps," Proc. Inst. Mech. Eng., Part E: J. Process Mech. 2002,216(1):47-58.

12. Arvin F et al., "Frequency shifting approach towards textual transcription of heartbeat sounds," Biol. Proced. 2011;13(7):50-56. doi: 10.1186/1480-9222-13-7

13. Albraik A et al., "Diagnosis of centrifugal pump faults using vibration methods," in J. Phys. Conf. Ser., Huddersfield, UK. 2012;364:p012139

14. Perloff J K, "Auscultation - The Audible Language of the Heart," in Physical Examination of the Heart and Circulation, 4th ed. Shelton, CT, USA: People's Medical Publishing House. 2009;6:157-223.

15. Chen Y, "Valvular Heart Disease," in Clinical Guide to Cardiology, C. F. Camm and A. J. Camm Eds. Hoboken, UK: John Wiley & Sons, Incorporated. 2016;14:194-196. doi: 10.1177/1751143719835452

Source of Support: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.