Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766209

PREDICTION OF HEART DISEASE USING MACHINE LEARNING

Dr.V.Rameshbabu[1], Dr.F.Antonyxavior, Dr.D.Usha[1], D.Rajesh kumar[2],P.Poovarasan[2], Ravi kumar[2]

[1]Professor, Department of CSE, Dr.MGR Educational and Research Institute,Chennai
[2]Final Year B.Tech-CSE, Dr.MGR Educational and Research Institute,Chennai
drvramesh25@gmail.com ,[1]antony.cse@drmgrdu.ac.in,usha.cse@drmgrdu.ac.in
[2]rajeshkavitha71002@gmail.com,poovarasanp604@gmail.com

ABSTRACT:
The main organ responsible for transporting blood throughout the body is the heart. Heart disease is an illness that kills people and is spreading globally in both developed and underdeveloped nations. In this disease, the heart typically struggles to deliver enough blood to other body components so that they can perform their usual activities. It links a number of heart disease risk factors with the requirement for time to obtain precise, trustworthy, and sane methods to make an early diagnosis and achieve quick disease care. Data mining is a popular method for processing vast amounts of data in the healthcare industry.The suggested approach is creating a machine learning model that can predict whether or not a person will experience a heart attack.

I. INTRODUCTION

1.1 Data Science

Data science is an interdisciplinary field that uses scientific methods, procedures, algorithms, and systems to extract information and insights from both organised and unstructured data. It then applies knowledge and practical insights from data across many application domains.

The term "data science" was coined in 1974 by Peter Naur as a new moniker for computer science. The International Federation of Classification Societies was the first conference to specifically focus on data science in 1996. But there was still disagreement over the term. The term "data science" was initially introduced in 2008 by D.J. Patil and Jeff Hammerbacher, the innovative heads of data and analytics operations at LinkedIn and Facebook. It has fast become one of the most well-liked and in-demand jobs.

Data Scientist:

The questions that need to be answered and the locations of the relevant data are examined by data scientists. They are able to mine, clean, and present data and possess analytical and business acumen. Large amounts of unstructured data are sourced, managed, and analysed by businesses using data scientists.

Required Skills for a Data Scientist:

- **Programming**: Python, SQL, Scala, Java, R, MATLAB.
- **Machine Learning**: Natural Language Processing, Classification, Clustering.

- **Data Visualization**: Tableau, SAS, D3.js, Python, Java, R libraries.
- **Big data platforms**: MongoDB, Oracle, Microsoft Azure, Cloudera.

1.2 ARTIFICIAL INTELLIGENCE

Emulation of human intellect, or artificial intelligence (AI), is the process through which machines are made to act and think like people. The term can also be applied to any computer that exhibits traits of the human intellect, such as learning and problem-solving.

In contrast to natural intelligence, artificial intelligence (AI) is intelligence that is displayed by machines. The study of "intelligent agents," or any system that senses its environment and takes actions to maximise its chance of attaining its goals, is how leading AI textbooks characterise the field. However, this definition is rejected by leading AI researchers. Some popular accounts use the term "artificial intelligence" to denote robots that simulate "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving". Artificial intelligence is the ability of machines, particularly computer systems, to mimic human intellectual functions. Some examples of particular applications of AI include expert systems, natural language processing, speech recognition, and machine vision.

The many subfields of AI research are centred on particular goals and the application of particular methods. Some of the traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, sensing, and the capacity to move and manipulate objects. General intelligence, or the ability to solve any problem, is one of the long-term goals of the field. AI researchers use a variety of techniques to tackle these problems, including formal logic, artificial neural networks, various search and mathematical optimization techniques, and methods from statistics, probability, and economics. AI also has an impact on many other fields, including computer science, psychology, linguistics, philosophy, and many others.

The idea that human intellect "can be so thoroughly characterised that a machine may be constructed to imitate it" served as the foundation for the study. This sparks philosophical discussions on the nature of the mind and the morality of developing intelligent artificial entities.From antiquity, myth, fiction, and philosophy have all addressed these concerns. With its great potential and power, science fiction and futurology have also proposed that AI may endanger humanity's existence.

The term "AI" is frequently used to refer to just one component of AI, such as machine learning. AI needs a base of specialised hardware and software to build and train machine learning algorithms. There isn't a single programming language that is solely related to AI, but a few are, such Python, R, and Java.

Learning processes:

Gathering data and creating the rules that will allow the data to be translated into knowledge are the main goals of this branch of AI programming. The rules, sometimes

referred to as algorithms, provide computing devices with comprehensive instructions on how to carry out a certain action

Reasoning processes: The optimum algorithm for a given result is what this area of AI programming is all about.

Self-correction processes: Algorithms are always being improved as part of AI programming to ensure that they produce the most accurate results. AI is important because, in some situations, it can execute tasks better than people can and because it may give businesses previously unattainable insights into their processes. Particularly when it comes to repetitive, detail-oriented tasks like evaluating a large number of legal documents to ensure that important fields are filled in accurately, AI systems usually complete work quickly and with few errors.

Artificial intelligence (AI) technologies like deep learning and artificial neural networks are rapidly developing, mostly because AI can process enormous volumes of data far more quickly and correctly than a human can.

Natural Language Processing (NLP):

Natural language processing (NLP) has made it possible for computers to read and understand human language. With the aid of a sufficiently powerful natural language processing system, natural language user interfaces and knowledge acquisition straight from human-written sources, such newswire documents, may be feasible. Information retrieval, text mining, question answering, translation, and information retrieval are a few straightforward uses for natural language processing. Many contemporary approaches make advantage of the word co-occurrence frequencies to construct syntactic representations of text. A search for "dog" might only turn up papers that contain the word "dog" in its literal form, missing ones that have the term "poodle." Search algorithms for finding keywords are widespread, scalable, and stupid. By analysing the frequency of words like "accident," "lexical affinity" approaches can ascertain the tone of a document.Modern statistical NLP systems can combine several of these strategies with others, and they frequently result in acceptable accuracy at the page or paragraph level. The ultimate goal of "narrative" NLP, which goes beyond semantic NLP, is to accurately represent common sense reasoning.

II. LITERATURE SURVEY

z.briqech et al.,(2016)The experiment results show that Health-Radio can achieve a median MI detection accuracy of 81.2 percent when the users are stationary, which is comparable to ECG-based MI detection. Even when the users are not stationary, Health-Radio can still achieve a median detection accuracy of 66.5 percent.

x.xu et al.,(2019)To extract breathing pattern in ESD signals, BreathListener eliminates interference from driving environments in ESD signals utilizing background subtraction andEnsemble Empirical Mode Decomposition (EEMD). After that, theextracted breathing pattern is transformed into Hilbert spectrum,and we further design a deep learning architecture

based on Generative Adversarial Network (GAN) to generate fine-grained breathing waveform from the Hilbert spectrum of extracted breathing patterns in ESD signals.

M.O.Mendez et al.,(2010) 24 full
Polysomnograpy recording from healthy
sleeper were used for the two sets of 12 each; training and test sets. The classificationperformance for accuracy=0.793 and sensitivity=0.702
P.Melillo,et al.,(2011) This study investigates the variations of Heart Rate Variability (HRV) due to a real-life stressor and proposes a classifier based on nonlinear features of HRV for automatic stress detection.

W. Wang,et al.,(2015)CARM uses this correlation as the profiling mechanism and recognizes a given activity by matching it to the best-fitprofile. We implemented CARM using commercial WiFidevicesand evaluated it in several different environments. Our results showthat CARM achieves an average accuracy of greater than 96%.
M. Zhao et al.,(2016) We describe the design and implementation of EQ-Radio, and demonstrate through a user study that its emotion recognition accuracy is on par with stateof-the-art emotion recognition systems that require a person to be hooked to an ECG monitor
U. R. Acharya et al.,(2016)Our proposed method can be used as an automated diagnostic tool for (i) the detection of different (10 types of) MI by using 12 lead ECG signal, and also (ii) to locate the MI by analyzing only one lead without the need to analyze other leads. Thus, our proposed algorithm and computerized system software (incorporated into the ECG equipment) can aid the physicians and clinicians in accurate and faster location of MIs, and thereby providing adequate time available for the requisite treatment decision.

III. Materials and methods

3.1 Aim:

One of the key elements in our healthcare system is heart attack. There are many patients who are there in the world right now. It is challenging to locate the heart attack. So, our project can quickly identify the attack.

3.2 Objectives:The objective is to create a machine learning model for heart attack prediction, which might eventually take the place of updateable supervised machine learning classification models by foretelling outcomes with the highest degree of accuracy by contrasting supervised algorithms.

3.3 Scope of the Project

The project's objective in this case is to see if integrating a patient's attack with a computer-based prediction could lower errors and enhance prediction accuracy. This approach has promise since techniques for data modelling and analysis, such as data mining, have the ability to create a knowledge-rich environment that can help to considerably enhance the accuracy of heart attack prediction.

IV. MODULES

Data Pre-processing:

The error rate of the machine learning (ML) model is obtained using validation procedures, and is thought to be as close to the actual error rate of the dataset as possible. You might not require the validation approaches if the volume of the data is sufficient to be representative of the population. Yet, working with data samples that could not be a realistic reflection of the population of a given dataset in real-world circumstances.Finding duplicate values, missing values, and information about the data type—whether a float variable or an integer—are all necessary.the subset of data used to assess a model's fit to a training dataset while adjusting model hyper parameters. As skill from the validation dataset is incorporated into the model setup, the evaluation gets increasingly skewed. A given model is evaluated using the validation set, although this is done frequently. This information is used by machine learning developers to adjust the model hyper parameters. A time-consuming to-do list might result from the collection, analysis, and process of dealing with data content, quality, and structure. Understanding your data and its characteristics can assist you choose the method to utilise to construct your model during the data identification phase.

Several of these sources contain merely careless errors. Sometimes there may be a more significant cause for missing data. It's critical from a statistical perspective to comprehend these various missing data kinds. The kind of missing data will affect how it is handled in terms of filling in the blanks, identifying missing values, basic imputation, and a thorough statistical methodology. Before writing any code, it's crucial to comprehend where the missing data is coming from. These are a few typical explanations for missing data:

- The user neglected to complete a field.
- In manually transferring data from an older database, data was lost.
- A programming error occurred.

Why Individuals declined to enter information in a field because they had preconceived notions about how the results would be used or interpreted.

Data Validation/ Cleaning/Preparing Process

Loading the specified dataset while importing the library packages. To evaluate the missing values, duplicate values, and variable identification by data shape and type. A validation dataset is a sample of data withheld from model training that is used to measure model competence when fine-tuning models and techniques that you may employ to maximise the utilisation of validation and test datasets when assessing your models.To evaluate the uni-variate, bi-variate, and multi-variate processes, data cleaning and preparation steps include renaming the provided dataset, deleting columns, etc. Depending on the dataset, different procedures and methods will be used to clean the data. To make data more valuable for analytics and decision-making, data cleaning's main objective is to find and remove mistakes and abnormalities.

MODULE DIAGRAM

Exploration data analysis of visualization

In applied statistics and machine learning, data visualisation is a crucial ability. In fact, the main focus of statistics is on numerical estimates and descriptions of data. An essential set of tools for obtaining a qualitative understanding is provided by data visualisation. This can be useful for discovering trends, corrupt data, outliers, and much more when exploring and getting to know a dataset.Data visualisations can be utilised to convey and illustrate critical relationships in plots and charts that are more visceral and engaging to stakeholders than measurements of association or importance with a little subject knowledge. It will suggest a deeper look at some of the books suggested at the conclusion because data visualisation and exploratory data analysis are entire fields in themselves.

Data may not always make sense unless it is presented visually, such as through charts and graphs. Both applied statistics and applied machine learning value fast visualisation of data samples and other objects. It will show you the many plot types that are available for use when visualising data in Python and how to utilise them to comprehend your own data.
Pre-processing refers to the modifications done to our data before we give it to the algorithm. Data Preprocessing is a technique for turning incomplete data sets into complete ones. In other words, if data are gathered from various sources, they are gathered in an unprocessed way that prevents analysis. For the machine learning method used to apply the model to produce better results, the data must be correctly organised. A number of machine learning models have particular data requirements, such as the Random Forest algorithm's inability to accept null values. In order to use the random forest method, null values from the initial raw data collection must be kept under control.The data collection should also be organised so that numerous Machine Learning and Deep Learning algorithms can be applied to a single dataset.

MODULE DIAGRAM

Comparing Algorithm with prediction in the form of best accuracy result

It is crucial to systematically compare the performance of various machine learning algorithms, and it will become clear that scikit-learn in Python may be used to build a test harness for this purpose. You can apply this test harness as a model for your own machine learning issues and include additional and various algorithms to contrast. There will be variations in the performance attributes of each model.You may estimate each model's potential accuracy on unobserved data by using resampling techniques like cross validation. It must be able to select one or two of the best models from the group of models you have developed using these estimates. It is a good idea to visualise new datasets using a variety of ways in order to view the data from many angles. The choice of models follows the same logic. In order to select the one or two that will be used for finalisation, you should examine the estimated accuracy of your machine learning algorithms in a variety of methods.Using various visualisation techniques to display the average accuracy, variance, and other characteristics of the distribution of model accuracies is one way to achieve this.

Each algorithm is tested using the K-fold cross validation technique, which is crucially configured with the same random seed to guarantee that the splits to the training data are carried out consistently and that each algorithm is evaluated in the same manner. Prior to the comparison algorithm, installing Scikit-Learn libraries and creating a machine learning model. Preprocessing, a linear model with the logistic regression method, cross-validation using the KFold method, an ensemble with the random forest method, and a tree with a decision tree classifier must all be completed in this library package. Separating the train set and test set is also a good idea. to compare accuracy when forecasting an outcome.

Prediction result by accuracy:
A value is also predicted using a linear equation and independent predictors by the logistic regression algorithm. The predicted value can be anything between -infinity and infinity. The output of the algorithm must classify variables. The logistic regression model forecasts outcomes with a better level of precision by comparing the best accuracy.

False Positives (FP):
A potential defaulter who will pay is expected. when the anticipated class is true but the actual class is false. For instance, if the predicted class informs you that the passenger would survive but the actual class reports that the passenger did not survive.
False Negatives (FN): One who is likely to default is a payer. when the projected class is no but the actual class is yes. For instance, if the passenger's actual class value shows that they survived and the forecast class suggests that they would pass away.
True Positives (TP): Anonpayer who is considered a defaulter. These successfully predicted positive values imply that both the actual class value and the expected class value are true. For instance, if both the projected class and the actual class value suggest that this passenger survived, then you know the same thing.
True Negatives (TN):
One who is likely to default is a payer. These are the accurately predicted negative values, indicating that neither the actual class value nor the projected class value are positive. For instance, suppose the projected class and the actual class both report that the passenger did not survive.

True Positive Rate(TPR) = TP / (TP + FN)
False Positive rate(FPR) = FP / (FP + TN)

Accuracy:
The percentage of total forecasts that are accurate, or in other words, how frequently the model predicts defaulters and non-defaulters with accuracy.
Accuracy calculation:
Accuracy = (TP + TN) / (TP + TN + FP + FN)
The easiest performance metric to understand is accuracy, which is just the proportion of properly predicted observations to all observations. One would believe that if our model is accurate, it is the best. Indeed, accuracy is an excellent indicator, but only when the values of the false positive and false negative rates are nearly equal in the datasets.

Precision: The proportion is positive predictions that are correct approximately.
Precision = TP / (TP + FP)
In terms of positive observations, precision is the proportion of accurately anticipated observations to all predicted positive observations. The question that this measure answer is of all passengers that labelled as surviving, how many actually survived? High precision and low false positive rate are related. Our precision was 0.788, which is quite good.
Recall:
The percentage of projected positive observed values that came true.
Recall = TP / (TP + FN)
Recall(Sensitivity) -Recall is defined as the proportion of accurately predicted positive observations to all of the actual class observations, yes.The weighted average of Precision and Recall is the F1 Score. Hence, both false positives and false negatives are considered while calculating this score. Although F1 is generally more beneficial than accuracy, especially if you have an uneven class distribution, it is not intuitively as simple to understand as accuracy. When false positives and false negatives cost about the same, accuracy performs best. It is preferable to include both Precision and Recall if the costs of false positives and false negatives are significantly different.
General Formula:
F- Measure = 2TP / (2TP + FP + FN)
F1-Score Formula:
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

ALGORITHM AND TECHNIQUES

Algorithm Explanation: Classification is a supervised learning strategy used in machine learning and statistics, where a computer software learns from the data input provided to it and then applies this learning to classify fresh observations. This data set may be multi-class or it may just be bi-class (for example, indicating whether the individual is male or female or if the message is spam or not). Speech recognition, handwriting recognition, biometric identity, document classification, etc. are a few instances of classification issues. Algorithms are taught by supervised learning from tagged data. After gaining an understanding of the data, the algorithm decides which label fresh data should receive based on patterns and associations with unlabeled new data.

Logistic Regression

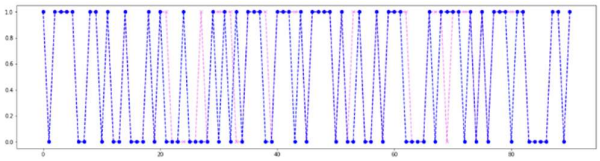
It is a statistical approach used to examine a set of data when the outcome is influenced by one or more independent variables. To gauge the outcome, a dichotomous variable is employed (in which there are only two possible outcomes). The goal of logistic regression is to find the model that best explains the relationship between a set of independent (predictor or explanatory) features and an interesting dichotomous characteristic (dependent variable = response or outcome variable). Logistic regression, a machine learning classification algorithm, is used to estimate the likelihood of a categorical dependent variable.

In other words, P(Y=1) is predicted by the logistic regression model as a function of X.

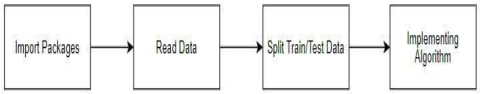
```
accuracy = accuracy_score(y_test,predicted)
print('Accuracy of Logistic Regression',accuracy*100)
```

```
Accuracy of Logistic Regression 84.61538461538461
```

Fig-1:Accuracy of logistic regression:-



MODULE DIAGRAM



Random Forest Classifier:Random forests, also known as random decision forests, build a lot of decision trees during the training phase as an ensemble learning technique for classification, regression, and other tasks, and they produce a class that represents the mean of the classes (for classification) or mean prediction (for regression) of the individual trees. Random decision forests are used to remedy decision trees' propensity to overfit their training set. Random forest is a technique to supervised machine learning built on ensemble learning. You can combine various algorithm types or employ the same strategy more than once in ensemble learning to get a prediction model that is more accurate. The random forest approach combines multiple algorithms of the same type, or various decision trees, into a forest of trees, hence the name "Random Forest". The random forest method can be used for both classification and regression applications.

The basic steps in applying the random forest algorithm are as follows:
Choose N records at random from the dataset.On the basis of these N records, construct a decision tree

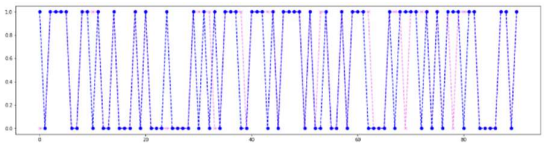
Repeat steps 1 and 2 until you have the amount of trees you want in your algorithm.

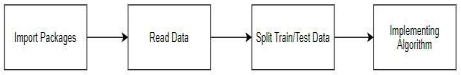
Each tree in the forest predicts a value for Y in the case of a regression problem for a new record (output). By averaging all of the values projected by every tree in the forest, the ultimate value may be determined. Alternatively, if there is a categorization issue, each tree in the forest can forecast which category the new data belongs to. The category that receives the majority of the votes is finally given the new record.

Fig-2 Accuracy of KNeoghbors classifier:

```
accuracy = accuracy_score(y_test,predicted)
print('Accuracy of KNeighbors Classifier',accuracy*100)
```

```
Accuracy of KNeighbors Classifier 83.51648351648352
```





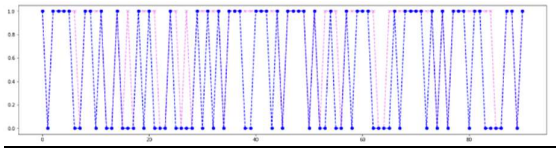
MODULE DIAGRAM

MLP Classifier: Multi-layer Perceptron classifier, or MLP Classifier, is connected to a neural network by the name itself. MLP Classifier uses an underlying Neural Network to execute the task of classification, unlike other classification methods like Support Vectors or Naive Bayes Classifier. MLP is a deep learning method because neurons are arranged in layers. MLP is frequently employed. for supervised learning-related tasks, computational neuroscience research, and parallel distributed computing. MLPs work well for classification prediction issues where inputs have a class or label assigned to them. They work well for issues involving regression prediction in which a real-valued quantity is forecasted from a collection of inputs. Because MLP is fully integrated, its drawbacks include having too many parameters.

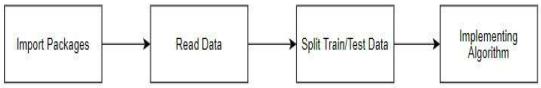
Fig-3 Accuracy of decision tree:

```
accuracy = accuracy_score(y_test,predicted)
print('Accuracy of Decision Tree',accuracy*100)
```

```
Accuracy of Decision Tree 75.82417582417582
```



MODULE DIAGRAM

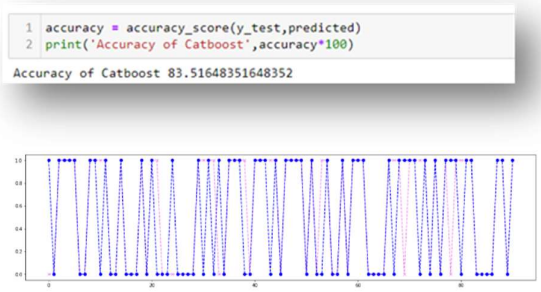


CatBoost Classifier algorithm:A collection of features (xi, xi) must be taken into account during training in order to select the model (yy) that will most efficiently solve the problem (regression, classification, or multiclassification) for any input object. This model is discovered using a training dataset, which is a set of objects with well-known feature values and label values. Just the validation dataset, which has data in the same format as the training dataset, is utilised to evaluate the training's quality and accuracy (it is not used for training). CatBoost is

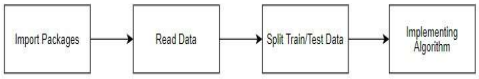
built on gradient-boosted decision trees. During training, a number of decision trees are built in succession. Each following tree is built with less loss than the ones that came before it.

The initial settings determine the number of trees. Use the overfitting detector to avoid overfitting. The growth of trees stops when it is activated.

Fig-4 accuracy of catboost:

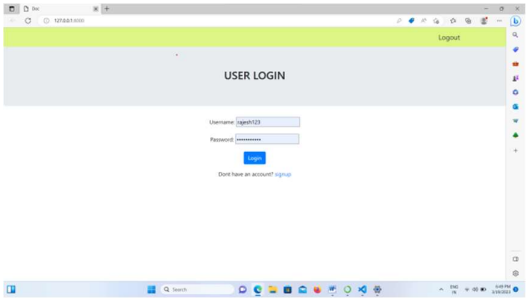


MODULE DIAGRAM



Result and discussion :-

Fig-4:



Output will display the user login then we will go for the patient report and enter the all wanted details of patient then we will get the final output predicted

V. Conclusion

Data preparation and processing, missing value analysis, exploratory analysis, and model construction and evaluation came first in the analytical process. It will be discovered who has the highest accuracy score on the public exam set. This application can assist in determining heart attack prediction.

REFERENCES:

[1] Z. Briqech and A.-R.Sebak, "Millimeter-wave imaging system using a 60 GHz dual-polarized AFTSA-SC probe,"in Proc. IEEE Nat. Radio Sci. Conf., 2016, pp. 325–332

[2] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, "BreathListener: Fine-grained breathing monitoring in driving environments utiliz-ing acoustic signals," in Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Serv., 2019, pp. 54–66.

[3] M. O. Mendez, M. Matteucci, V. Castronovo, L. Ferini-Strambi, S. Cerutti, and A. Bianchi, "Sleep staging from heart rate variabil-ity: Time-varying spectral features and hidden Markov models," Int. J. Biomed. Eng. Technol., vol. 3, pp. 246–263, 2010.

[4] P. Melillo, M. Bracale, and L. Pecchia, "Nonlinear heart rate vari- ability features for real-life stress detection. Case study: Students under stress due to university examination," Biomed. Eng. Online, vol. 10, 2011, Art.no. 9

[5] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in Proc. Int. Conf. Mobile Comput.Netw., 2015, pp. 65–76

[6] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in Proc. Annu. Int. Conf. Mobile Comput.Netw., 2016, pp. 95–108.

[7] D. Mozaffarian et al., "Heart disease and stroke statistics-2016 update a report from the american heart association," Circulation,vol. 133, pp. e38–e48, 2016.

[8] U. R. Acharya et al., "Automated detection and localization of myo-cardial infarction using electrocardiogram: A comparative study of different leads," Knowl.-Based Syst., vol. 99, pp. 146–156, 2016.

[9] D. Bhatnagar, V. Kumar, A. Kumar, and I. Kaur, "Graphenequan-tum dots FRET based sensor for early detection of heart attack in human," Biosensors Bioelectron., vol. 79, pp. 495–499, 2016.

[10] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," Inf. Sci., vol. 415, pp. 190–198, 2017.