

HEART DISEASE PREDICTION USING MACHINE LEARNING

Mr.VALLE HARSHA VARDHAN¹, Mr.UPPALA RAJESH KUMAR², Ms.VANUMU VARDHINI³,Ms. SABBI LEELA VARALAKSHMI⁴,Mr.A.SURAJ KUMAR⁵

1. BTECH, NADIMPALLI SATYANARAYANA RAJU INSTITUTE OF TECHNOLOGY, SONTYAM, VISAKHAPATNAM, ANDHRA PRADESH, INDIA - 531173

2. BTECH, NADIMPALLI SATYANARAYANA RAJU INSTITUTE OF TECHNOLOGY, SONTYAM, VISAKHAPATNAM, ANDHRA PRADESH, INDIA - 531173

3. BTECH, NADIMPALLI SATYANARAYANA RAJU INSTITUTE OF TECHNOLOGY, SONTYAM, VISAKHAPATNAM, ANDHRA PRADESH, INDIA - 531173

4. BTECH, NADIMPALLI SATYANARAYANA RAJU INSTITUTE OF TECHNOLOGY, SONTYAM, VISAKHAPATNAM, ANDHRA PRADESH, INDIA - 531173

5. Assistant Professor COMPUTER SCIENCE AND ENGINEERING, NADIMPALLI SATYANARAYANA RAJU INSTITUTE OF TECHNOLOGY, SONTYAM, VISAKHAPATNAM, ANDHRA PRADESH, INDIA-531173

ABSTRACT

In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions. If foreseen far in advance, such information can offer crucial intuitions to doctors, who can then modify their diagnosis and approach per patient. We are attempting to use machine learning algorithms to predict potential heart conditions in humans. In this project, we compare the performance of various classifiers, including Decision Tree, Naive Bayes, Logistic Regression, SVM, and Random Forest. We also propose an ensemble classifier that performs hybrid classification by combining the best features of both strong and weak classifiers because it can use a large number of training and validation samples. In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions. If foreseen far in advance, such information can offer crucial intuitions to doctors, who can then modify their diagnosis and approach per patient. We are attempting to use machine learning algorithms to predict potential heart conditions in humans. In this project, we compare the performance of various classifiers, including Decision Tree, Naive Bayes, Logistic Regression, SVM, and Random Forest. We also propose an ensemble classifier that performs hybrid classification by combining the best features of both strong and weak classifiers because it can use a large number of training and validation samples.

1. INTRODUCTION

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. One of the leading causes of morbidity and mortality among the global population is heart disease. One of the most crucial topics in the data analysis area is predicted cardiovascular disease. Since a few years ago, the prevalence of cardiovascular disease has been rising quickly throughout the world. Many studies have been carried out in an effort to identify the most important risk factors for heart disease and to precisely estimate the overall risk. Heart disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. Cardiovascular disease must be detected early. Aiding high-risk patients in making decisions regarding lifestyle changes will help to reduce the difficulties.

Making choices and predictions from the vast amounts of data generated by the healthcare sector is made easier with the help of machine learning. By evaluating patient data that uses a machine-learning algorithm to categorise whether a patient has heart disease or not, this study hopes to predict future cases of heart disease. Machine learning methods can be extremely helpful in this situation. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, despite the fact that heart disease can manifest itself in various ways. By gathering information from numerous sources, organising it into categories that make sense, and then performing analysis to get out the desired information based on statistics, we may conclude that this technique is quite adaptable.

The main difficulty with heart disease is detecting it. There are tools that can forecast heart disease, but they are either expensive or ineffective at calculating the likelihood of heart disease in a human. The mortality rate and total consequences can be reduced by early identification of heart disorders. Since it takes more intelligence, time, and knowledge, it is not always possible to accurately monitor patients every day, and a doctor cannot consult with a patient for a whole 24 hours. As there is a lot of data available nowadays, we can use a variety of machine learning methods to search for hidden patterns. The underlying patterns may be utilised in medical data for health diagnosis.

2. LITERATURE SURVEY AND RELATED WORK

In recent years, several experiments and researches have been conducted in the area of medical science and machine learning, resulting in the publication of important publications.

- [1] Purushottam, et al. Hill climbing and decision tree algorithms were proposed in a study by. are used in the System for Effective Heart Disease Prediction. The outcomes of algorithms like SVM and KNN are based on split conditions that can be vertical or horizontal depending on the dependent variables. Yet, a decision tree is a structure that resembles a tree with a root node, leaves, and branches, and it is based on the decisions made in each tree. The value of the attributes in the dataset is also explained by the decision tree. Also, they used the Cleveland data set. Using some techniques, the data set is divided into 70% training and 30% testing. The accuracy of this method is 91%. Naive Bayes, the second algorithm, is used for categorization. Since it can handle complex, nonlinear, dependent data, the heart disease dataset—which is similarly complex, dependent, and nonlinear in nature—is seen to be a good fit. This programme gives 87% accuracy
- [2] Sonam Nikhar et al proposed paper “ In their study, "Prediction of Heart Disease Using Machine Learning Algorithms," the Naive Bayes and decision tree classifiers, which are utilised specifically in the prediction of Heart Disease, are explained in detail.
- [3] Decision Trees had higher accuracy than Bayesian classifiers, according to certain studies that considered the use of predictive data mining approach on the same dataset.
- [4] The multi-layer perceptron neural network approach is used for dataset training and testing in the study "Prediction of Heart Disease Using Machine Learning" that Aditi Gavhane et al. proposed. There will be one input layer, one output layer, and perhaps more hidden layers in this algorithm between the two input and output layers. Each input node is connected to the output layer by hidden layers. Weights chosen at random are assigned to this link. The second input is referred to as bias, and it is given weight based on the needs of the connection between the nodes.
- [5] Avinash Golande et al, proposed Several data mining techniques are utilised in "Heart Disease Prediction Using Efficient Machine Learning Approaches," which helps doctors distinguish between different types of heart disease. K-nearest Neighbor, Decision Tree, and Naive Bayes are common techniques. Packing calculation, Part thickness, consecutive negligible streamlining, neural systems, straight Kernel selfarranging guidance, and SVM are other novel characterization-based procedures that are used (Bolster Vector Machine).
- [6] Lakshmana Rao et al, proposed "Machine Learning Methods for Heart Disease Prediction," where the causes of heart disease are discussed in more detail. Hence, it is challenging to distinguish heart illness. Several neural networks and data mining techniques are used to predict the severity of heart disease among patients.

- [7] Abhay Kishore et al proposed "Heart Attack Prediction Using Deep Learning" uses a Recurrent Neural System to forecast the likelihood of heart-related infections in the patient in addition to a heart attack prediction system. To provide the most accurate model with the fewest errors, this model employs deep learning and data mining. This study serves as a reliable benchmark for other heart attack prediction programmes.

3. Implementation Study

Heart disease is even being emphasised as a silent killer that causes a person to pass away without showing any outward signs. Growing concern about the illness and its effects is a result of the disease's nature. Thus, efforts to foresee the potential occurrence of this fatal disease in the past continue.

a group of datasets

For the foundation of our heart disease prediction system, we first gather a dataset. We divided the dataset into training and testing data after it was collected. The learning of the prediction model takes place on the training dataset, and the evaluation of the prediction model occurs on the testing dataset. 30% of the data are utilised for testing in this project, while 70% are used for training. The information gathered for this project

is UCI Heart Disease. The dataset has 76 properties, of which the system uses 14 for its operation.

Selection of attributes

The choice of acceptable attributes for the prediction system is included in attribute or feature selection. This is done to make the system more effective. For the prediction, a number of patient characteristics are used, including gender, the nature of the patient's chest discomfort, fasting blood pressure, serum cholesterol, and exang.

Pre-processing of Data

The pre-processing of data is a critical stage in the development of a machine learning model. Data that isn't initially clean or in the model's required format can lead to inaccurate results. Pre-processing involves transforming data into the format we need. It is used to handle the dataset's noise, duplication, and missing values. Activities like importing datasets, partitioning datasets, attribute scaling, etc. are all part of data pre-processing. Preprocessing the data is necessary to increase the model's accuracy.

Balancing of Data

Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling

(a) Under Sampling:

In Under Sampling, dataset balance is done by the reduction of the size of the ample class. This process is considered when the amount of data is adequate.

(b) Over Sampling:

In Over Sampling, dataset balance is done by increasing the size of the scarce samples. This process is considered when the amount of data is inadequate.

Prediction of Disease

For classification, a variety of machine learning algorithms are employed, including SVM, Naive Bayes, Decision Trees, Random Trees, Logistic Regression, Ada-boost, and Xg-boost. Algorithms are compared, and the one that predicts heart disease with the best degree of accuracy is chosen.

4. PROPOSED WORK AND ALGORITHM

The collection of data and selection of the most crucial attributes is the first step in the system's operation. The relevant data is then preprocessed into the format needed. After that, the data is split into training and testing data. The algorithms are used, and the training data is used to train the model. By testing the system with test data, the correctness of the system is determined. The modules listed below are used to implement this system.

1. Collection of Dataset
2. Selection of attributes
3. Data Pre-Processing
4. Balancing of Data

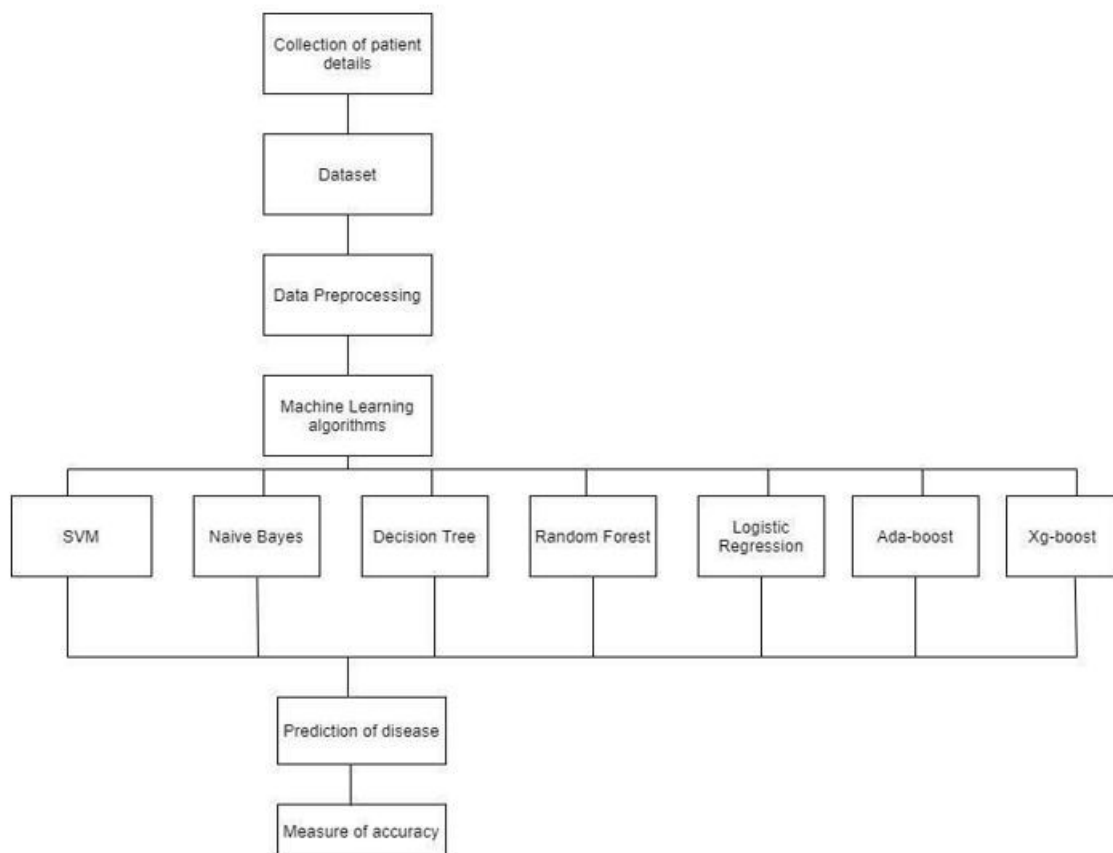


Fig: SYSTEM ARCHITECTURE

SUPPORT VECTOR MACHINE (SVM):

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues.

The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors are the phrase for these extreme circumstances, and as a result, Support Vector Machine is the name of the algorithm.

Strong yet adaptable supervised machine learning methods called support vector machines (SVMs) are employed in for regression and classification. Nonetheless, they are typically employed in classification issues. SVMs were initially introduced in the 1960s, but around 1990 they underwent further development. SVMs are implemented in a different way than other machine learning algorithms. They have recently gained a lot of popularity because to their capacity to manage numerous continuous and categorical variables.

NAIVE BAYES ALGORITHM:

The Naive Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set.

One of the most straightforward and efficient classification algorithms is the Naive Bayes Classifier, which aids in the development of quick machine learning models capable of making accurate predictions.

Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur. Spam filtration, Sentimental analysis, and article classification are some examples of Naive Bayes algorithms that are often used.

It is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence. A Naive Bayes classifier, to put it simply, believes that the existence of a specific feature in a Class and the existence of any other feature are independent.

DECISION TREE ALGORITHM:

Although it may be used to solve classification and regression problems, Decision Tree is a Supervised learning technique that is typically used for classification problems. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree.

Whereas Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches. The given dataset's features are used to execute the test or make the decisions. It is a graphical representation of all potential responses to a decision or difficulty based on the circumstances. It is known as a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure resembling a tree. The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to construct a tree.

5. METHODOLOGIES

METHODOLOGY includes the usage of the following concepts.

Python

Python is an interpreted, high-level, general-purpose programming language that was developed by Guido Van Rossum and initially released in 1991. With its noticeable usage of large White space, Python places a strong emphasis on code readability. Its language constructs and object-oriented methodology are designed to aid programmers in creating clean, comprehensible code for both little and big projects. Python has garbage collection and dynamic typing. Procedural, object-oriented, and functional programming are just a few of the programming paradigms it supports.

Sklearn

The most effective and reliable Python machine learning library is called Sklearn (Skit-Learn). With a consistent Python interface, it offers a variety of effective methods for statistical modelling and machine learning, including dimensionality reduction, clustering, and classification. This library is based on NumPy, SciPy, and Matplotlib and was written primarily in Python.

Numpy

A library for the Python programming language called NumPy adds support for big, multi-dimensional arrays and matrices as well as a tonne of high-level mathematical operations that can be performed on these arrays. Jim originally developed Numeric, the predecessor of NumPy, with assistance from a number of other developers. Travis developed NumPy in 2005 by heavily altering Numeric to incorporate capabilities of the rival Numarray. Several people have contributed to the open source programme NumPy.

Librosa

It offers the components required to build music information retrieval systems. Librosa uses a variety of signal processing techniques to extract features from audio sources and to assist display them.

Matplotlib

For the Python programming language and its NumPy numerical mathematics extension, Matplotlib is a graphing library. For applications employing all-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK, it offers an object-oriented API for embedding plots. It is not recommended to use the procedural "pylab" interface, which is based on a statemachine (similar to OpenGL) and was created to closely resemble the MATLAB interface.

Seaborn

A matplotlib-based Python data visualisation library is called Seaborn. It offers a sophisticated user interface for creating visually appealing and educational statistical visuals. Python's Seaborn package is mostly used to create statistical visuals. On top of matplotlib, Seaborn is a data visualisation framework that is tightly connected with Python's pandas data structures. The core of Seaborn is visualisation, which aids in data exploration and comprehension.

SciPy

SciPy includes modules for common tasks in science and engineering like optimisation, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, and ODE solvers. Moreover, SciPy is a family of conferences for users and creators of these tools, including SciPy.in, EuroSciPy, and SciPy in the United States (in India). Enthought founded the SciPy conference in the United States, and it continues to support several conferences abroad in addition to hosting the SciPy website. A scientific computation package called SciPy is built on top of NumPy. It offers more helpful functions for signal processing, statistics, and optimisation.

6. RESULTS AND DISCUSSION SCREENSHOTS

INPUT

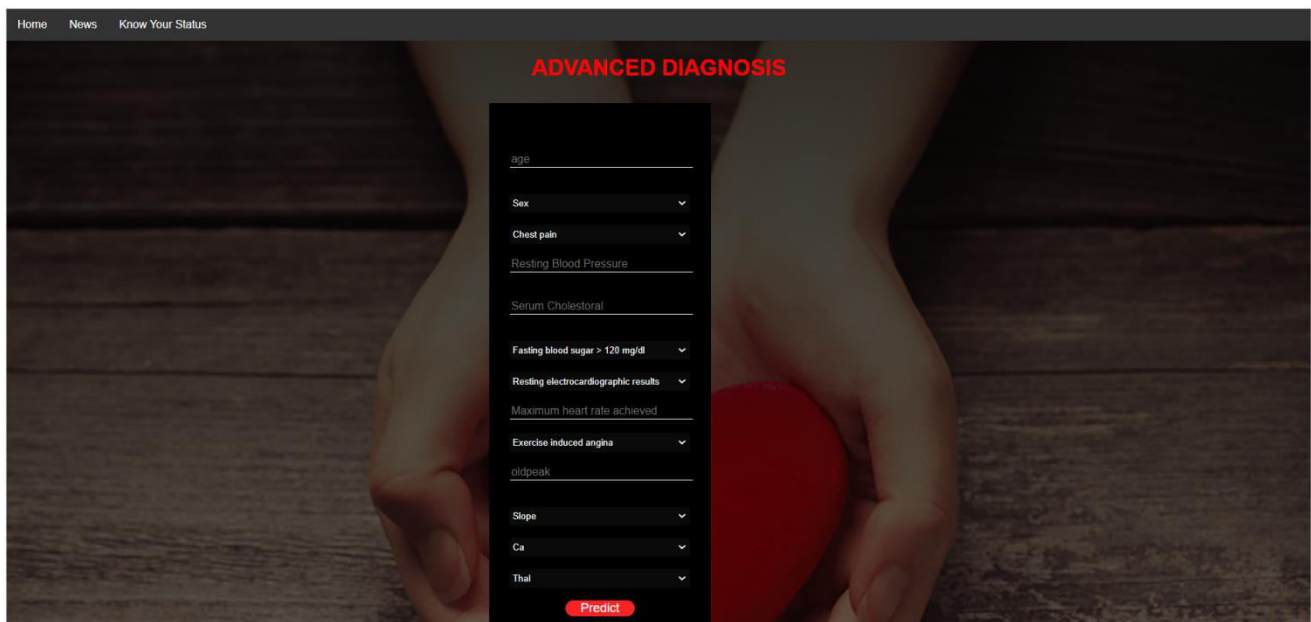


Fig - Interface for Heart disease prediction

60

♂ Male

atypical angina

89

120

No

showing probable or definite left ventricular hypertrophy by ECG

150

No

1

Downsloping

1

Reversible defect

Predict

OUTPUT

Home Know Your Status

ADVANCED DIAGNOSIS REPORT

| Features | User Data |
|--|--|
| Age | 60 |
| Sex | Male |
| Chest pain | Atypical angina |
| Resting blood pressure | 89 |
| Resting cholesterol | 120 |
| Fasting blood sugar greater than 120mg/dl | No |
| Resting electrocardiographic results | showing probable or definite left ventricular hypertrophy by ECG |
| Exercise induced angina | No |
| ST depression induced by exercise relative to rest | 150 |
| The shape of the peak exercise ST segment | downsloping |
| Number of major vessels (0-3) colored by fluoroscopy | 1 |
| Thal | reversible defect |
| Result | You are detected with heart problems. You need to consult a doctor immediately |

*This result is does not have standard medical approval. So for final result please approach a doctor.
For more accurate result use Advanced Diagnosis

| ADVANCED DIAGNOSIS REPORT | |
|---|--|
| Features | User Data |
| Age: | 60 |
| Sex: | Male |
| Chest pain: | Atypical angina |
| Resting blood pressure: | 89 |
| Serum cholestural: | 120 |
| Fasting blood sugar greater than 120mg/dl: | No |
| Resting electrocardiographic results: | showing probable or definite left ventricular hypertrophy by Estes |
| Exercise induced angina: | No |
| ST depression induced by exercise relative to rest: | 150 |
| The slope of the peak exercise ST segment: | downsloping |
| Number of major vessels (0-3) colored by fluoroscopy: | 1 |
| That: | reversible defect |
| Result | You are detected with heart problems. You need to consult a doctor immediately |
| *This result is does not have standard medical approval. So for final result please approach a doctor. For more accurate result use Advanced diagnosis | |

Fig - Report for Heart disease prediction

7. CONCLUSION AND FUTURE WORK

Application of promising technology, such machine learning, to the first prediction of heart problems would have a significant social impact because heart diseases are a leading cause of death in India and around the world. Early detection of cardiac disease can help high-risk patients make decisions about lifestyle modifications that will lessen problems, which can be a significant advancement in the field of medicine. Each year, more people are diagnosed with cardiac illnesses. This calls for an early diagnosis and course of action. The medical community as well as patients may benefit greatly from the use of appropriate technology support in this area. The seven machine learning algorithms employed in this study to gauge performance are SVM,

applied to the dataset along with Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting.

The dataset, which includes 76 features, contains the expected characteristics that contribute to heart disease in individuals, and 14 significant characteristics are chosen from them to help assess the system. If all the features are taken into account, the creator receives a less efficient system. Attribute selection is carried out to improve efficiency. In this case, n characteristics must be chosen in order to evaluate the model that provides greater accuracy. Several dataset features have virtually equal correlations, so they are eliminated. When all of the dataset's qualities are taken into consideration, the effectiveness drops significantly.

A prediction model is created after comparing the accuracy of each of the seven machine learning techniques. So, the objective is to employ a variety of evaluation metrics, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the disease. The extreme gradient boosting classifier has the highest accuracy (81%), when all seven are compared.

8. REFERENCES

- [1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. *International Journal of Computer Applications*, 17(8), 43-8
- [2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, 47(10), 44-8.
- [3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine*, 10(2), 334-43.
- [4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naive Bayes algorithm. *International Journal of Computer Science and Information Technologies*, 6(1), 637-9.
- [5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In *International Conference on Information Society (i-Society 2014)* (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9
- [6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. *BMJ open*, 4(5), e005025.
- [7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. *Arteriosclerosis, thrombosis, and vascular biology*, 33(9), 2267-72.
- [8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In *2013 International Mutli- Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)* (pp. 40- 6). IEEE.
- [9] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register. *BMJ*, 315(7101), 159-64.
- [10] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women. *International journal of epidemiology*, 18(2), 361-7.

- [11]Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557- 60). IEEE.
- [12]Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).
- [13]Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.
- [14]Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device -kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.
- [15]Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). Validation of heart failure events in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone. Current controlled trials in cardiovascular medicine
- [16]Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.
- [17]A. Aldallal and A. A. A. Al-Moosa, "Using Data Mining Techniques to Predict Diabetes and Heart Diseases", 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), pp. 150-154, 2018, September.
- [18]Takci H (2018). Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.
- [19]Ankita Dewan and Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)
- [20]Aditya Methaila, Prince Kansal, Himanshu Arya and Pankaj Kumar, "Early heart disease prediction using data mining techniques", Computer Science & Information Technology Journal, pp. 53-59, 2014.

Web Links

- [1] <https://www.geeksforgeeks.org/ml-heart-disease-prediction-using-logistic-regression/>
- [2] <https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction>
- [3] <https://github.com/g-shreekant/Heart-Disease-Prediction-using-Machine-Learning>