Predictive Analysis of Cardiovascular Disease

Ginjupalli Sreehasa , Bondu Sankeerthana,Pothagani Yogitha, Kadulla Bhavana,K.Venkata Prasad

Koneru Lakshmaiah Education Foundation

Abstract:- Worldwide, heart disease stands as a leading cause of fatalities. The treatment of heart disease necessitates the application of cutting-edge technologies. Within medical centers, a prevailing issue arises where many medical professionals lack the equal knowledge and expertise required for optimal patient care.Consequently, they often make independent decisions, resulting in subpar outcomes and, at times, even fatalities. To address these challenges, the prediction of cardiac illness by applying machine learning methods and techniques has emerged as a viable solution. These technologies have simplified the process of automatic diagnosis in most of all the hospitals, playing a crucial role in improving patient care. The prognosis of cardiovascular illness relies on the analysis of various health parameters in patients. Numerous algorithms, such as Naïve Bayes, KNN, ANN, are employed for heart disease prediction. Ten metal oxide semiconductor sensors are used in conjunction with a method using artificial neural networks (ANNs) to recognise scent patterns in certain individuals. Our research incorporates diverse parameters, including gender, CP and blood pressure ,age, among others.

Keywords: Heart Disease, Artificial neural networks, medical centers, Blood Pressure, Algorithms, KNN.

1. Introduction

The heart is the most sensitive organ. As the most prevalent illness in the modern world, coronary artery disease is leading to an increase in death rates. The World Health Organization claims that 3.7 persons perish from coronary artery disease world- wide. A poll indicates that because of CAD, 10 million people die. Heart disease surgery is difficult, especially in underdeveloped nations where there is a shortage of medical professionals with training, testing tools and additional resources required for a precise diagnosis and heart patient care **[1].** The World Health Organization estimates that coronary artery heart disease affected about 16 million people in the United States. It was shown that 50% of men and women over 40 and 30% of women suffer with CAD. **[2].** Early identification of cardiac conditions can stop the ignorance and the high fatality rate, many do not know how to recognize heart disease early. Health Industries are working to identify the sickness early on. In most instances, it is discovered in advanced stages of the illness or after death. The price of heart disease treatment is very high Getting a diagnosis early can be challenging **[1].** People are hesitant to receive competent care at an early stage as a result. The primary risk factors for heart disease include age, alcoholism, smoking, and obesity. **[2].** Large volumes of data are gathered by the health care sector, but they are not mined to find hidden information **[3].**

A diagnostic procedure called ECHO uses ultrasound pulses to make a virtual reflection of the heart's muscles that can be utilized to identify cardiovascular disorders (CVD). It is among the more difficult illnesses to identify requiring both advanced skills and prior knowledge. Data mining is a complex process and capturing implicit, hitherto undiscovered potentially helpful information, often known as medical knowledge employing sophisticated algorithms, data. Big data (BD) is a type of tool referred to as a large information set or record **[4]**. It is BD and DM. These two worked on projects that were comparable in that they focused on collecting a lot of data, organizing it, and producing a report on it by extracting the relevant information.

Extraction of useful insights from large data sets is made possible in large part by data mining. Instead of doing everything by hand, the customers can formulate their queries to retrieve information from databases automatically **[4].** Clustering is one such technique that involves grouping data, with hierarchical and non-hierarchical clustering as its two primary approaches. Clustering serves as a means for undirected knowledge discovery. Memory-based reasoning is employed to identify homogeneous instances from historical records and predict unknown samples. Decision-based trees are utilized for data classification and prediction. During the preprocessing stage, data imputation techniques have been employed to fill in the missing data **[12].** To determine whether cardiac disease is present or not, the multi-class variable is utilized **[9].** The authors conducted a survey based on World Health Organization data pertaining to coronary artery disease. This study explores several data mining methods that support cardiac illness early detection, thus helping to reduce its prevalence. These techniques empower the development of medical systems capable of detecting heart diseases using extensive patient datasets. The healthcare sector grapples with a vast amount of data, making it challenging to mine and analyse for heart disease detection **[2].** substantial volume of data is extracted from records of patients with heart disease, which is then examined. Big data refers to extremely large amounts of data, and data mining is the process of obtaining knowledge or information that is helpful. Although these terms differ, they share similar tasks, such as data extraction, data management, analysis, and storage. Techniques like association rules, neural networks, and genetic algorithms are employed for heart disease detection. Heart disease is predicted using Bayesian classification and association criteria. An intelligent heart disease prediction system is designed using naive Bayes, neural networks, and decision trees. Probability is the basis of the classifier. **[5].** Additionally, k-means and MAFIA algorithms are used to construct a heart attack prediction system. It highlights how big data technology may be used in the healthcare industry to improve patient care, better understand the causes of diseases, and lower the occurrence of diseases. The author used a variety of instruments and methods to analyze large amounts of data. In contrast to usual databases, large data is a grouping of semi-structured, structured, and unstructured data.. Big data poses several challenges, including worries regarding patient data security and privacy. To analyze big data, multiple tools are employed, such as MongoDB, Hadoop, and Splunk **[2].** In a study focused on predicting patient survival, various machine learning classifiers were employed. They ranked features related to significant risk factors and compared them to traditional biostatistical tests and supplied algorithms for machine learning. The initial stage of the data classification consists of root nodes, and the final stage is the attribute terminal node **[5].** Four distinct algorithms were used in the preparation and analysis of the dataset, yielding precision rates of 99.83% for the Random Forest and Decision Tree approaches, and 85.32% and 84.49% for the SVM and KNN methods, respectively.

Congestive Heart Failure (CHF) was successfully predicted by another study employing an ensemble approach. This was accomplished by using deep neural networks to fill in knowledge gaps in related domains and by assessing Heart Rate Variability (HRV). The suggested technique had an astounding 99.85% accuracy rate. The initial stage of the data classification consists of root nodes, and the final stage is the attribute terminal node **[6].** This bad performance is thought to be caused by these algorithms' incapacity to accurately identify the most important and closely related elements.

Our objective is to create a methodology that, in order to solve this problem, first finds the ideal set of features and then discovers the algorithms that work best with these features. Our analysis suggests that algorithms that exhibited strong performance benefited from a closely correlated feature set, primarily obtained through the application of Relief techniques. Conversely, the algorithms that did not perform well seemed to struggle in properly assessing the interrelated nature of the features they utilized.

2. Objective

In our study, we leveraged a pre-existing dataset and implemented five distinct techniques utilising the same dataset to forecast cardiac conditions.These algorithms consist of Artificial Neural Networks, Decision Trees, Support Vector Machines (SVM), Random Forest, and K Nearest Neighbour. Our paper aims to identify the technique that offers highest accuracy in predicting heart disease based on health parameters. Our experiments demonstrate that Naïve Bayes got 88 percentage which is the highest accuracy rate. To validate our solution, we conducted user research involving 20 individuals. By enabling an repetitious examination HDVis technology enhances the workflow of visual interpretation These findings suggest that the health care community should place greater emphasis on implementing effective legislative measures to reduce the prevalence of heart disease.

3. Literature Survey

Drawing from 1000 cases, Yanwei (2007) designed methods for data mining to predict heart disease patients' chances of survival (CHD).Since it was a problem for the medical community, the CHD prediction technique needs to be addressed as a matter of priority. The work involved extensive analysis of medical records for 1000 CHD patients for six months. Data regarding the survival rate was kept." 10-fold cross-validation evaluates and quantifies the efficacy and precision of various methods. Accuracy, sensitivity, and specificity" were the metrics employed. To calculate the three metrics, a confusion matrix was obtained. For SVM, ANN, and DT, the accuracy was 92.1%, 91.0%, and 89.6%, respectively. a comparison study using various prediction models **[4].** Pruning Classification was used by Deepika N. (PCAR) Association Rule. Apriorism algorithm is the source of the pruning Classification Association Rule. The suggested approach subtracts the minimum frequency a thing with a low frequency item sets, removes sporadic items from item sets, and then the common item set is found. By combining DT, NB, and NN, DM, Sellappan Palaniappan et al. (2008) proposed the "Intelligent Heart Disease Prediction System (IHDPS)" model.

The results showed the effectiveness of each strategy. major questions, such as "what-if" IHDPS provided answers where general DT failed to do so. Patterns with relationships to significant information were created with HD risk factors in mind. Model was created for the web, enhancing its usability. Providing both vertical and horizontal scalability, reliable. The risk factors of age, blood sugar, blood pressure detection of HD was based on history and other factors. 15 attribute information set was retrieved from database for "Cleveland" in HD.

Nidhi Bhatia et al. **[3]** suggested analysing different data-mining methods for predicting heart disease. According to the observations, neural networks with 15 properties have fared better than any other data mining methods. The analysis's key finding is that decision trees have also shown good accuracy using a genetic algorithm and selection of a feature subset .

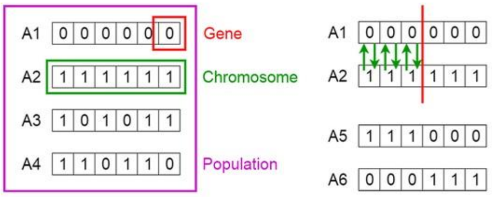
4. Methods

A detailed explanation is furnished for the development of an intelligent machine learning system using the dataset related to chronic heart ailments. To evaluate patients' degrees of severity, researchers used a hybrid technique. A modified kNN algorithm, a backpropagation neural netwok, and a genetic algorithm (GA) were used to create their technique. The GA was used for feature selection and the KNN was used for severity classification **[9].** Prior to employing ANN patterns, sensor data is scanned and retrieved **[13].** This study offers a model titled 'Data Mining-Based Approach for Heart Disease Diagnosis' as a prototype for diagnosing heart conditions.

Genetic algorithm:

A genetic algorithm (GA) is a computational method that mimics the principles of natural evolution for solving complex optimization problems. The application of genetic algorithms in our system involves their regular utilization for generating effective solutions to optimization and search challenges. Specifically, as represented in Fig-1 we employ genetic algorithms to extract attributes from a vast set of attributes, allowing us to identify and select the most relevant features for our problem-solving process **[3].**

Fig-1 Genetic algorithm



Genetic algorithms, although effective at determining optimal variable subsets for models, often entail significant computational demands. These algorithms are flexible in managing various optimization scenarios, capable of adapting to objective functions that can be either continuous or discontinuous, linear or nonlinear, stable or nonstationary, and even subject to random noise. Due to the autonomous behavior of numerous offspring within a population, this collective entity, or any subset thereof, possesses the capability to concurrently examine the search space from various directions. This characteristic renders it well-suited for algorithm parallelization during the implementation process.

K-means clustering algorithm:

The method of k-means clustering tackles the well-known clustering challenge and stands out as one of the most straightforward unsupervised learning techniques available. Its simplicity and effectiveness make it a preferred choice for data clustering tasks **[3].** This technique permits the classification of data into various clusters, offering a practical means to unveil group categories within untagged datasets autonomously, eliminating the necessity for prior training. In the initial stage of k-means clustering with K=2, the first step as shown in Fig-2(a) involves the random allocation of two centroids, represented by two arbitrarily positioned points. An important consideration is that although these points are referred to as centroids, they do not necessarily correspond to the central points within the dataset at this stage.

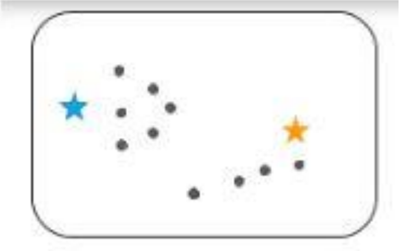


Fig-2(a) k-means clustering

Subsequently, the following stage Fig-2(b) indicates the calculation of distances between the data points and the initially assigned centroids. For each data point, its distance is computed with respect to both centroids, and it is then associated with the centroid for which it exhibits the shorter distance. This allocation of data points to the centroids is visually depicted, with data points being shown in blue and yellow, connected to their respective centroids.

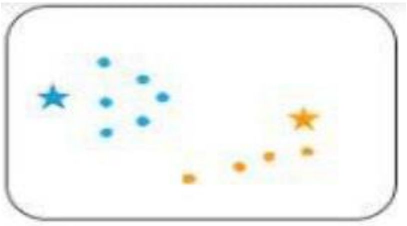


Fig-2(b) Centroids represented graphically

Following this, the subsequent step Fig-2(c) involves establishing the true centroids for these two clusters. The initial centroids, initially assigned at random, are then adjusted to the actual centroids of the respective clusters Finally, Fig-2(d) demonstrates the iteration of calculating distances and adjusting centroids persists until the final clustering configuration is achieved, at which point the repositioning of centroids concludes.

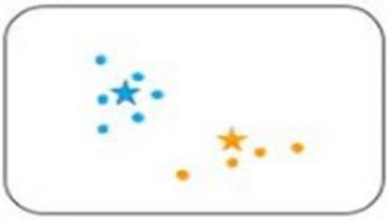


Fig-2(d) displacement of the centroids

MAFIA algorithm:

The MAFIA algorithm is utilized to discover the most frequent maximal item sets within databases, showcasing its exceptional efficiency, particularly in scenarios where the database contains extensive item sets. The algorithm's search process integrates a depth- first exploration of the itemset framework, accompanied by adept pruning techniques for optimization [3].Its efficiency shines when dealing with exceedingly lengthy item sets within transactions **.**

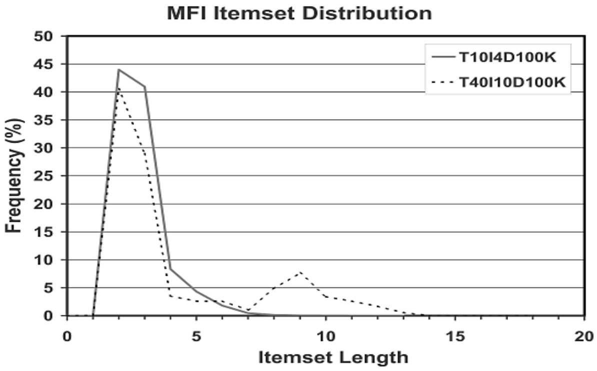


Fig-3 MAFIA algorithm

Fig-3 illustrates 'MAFIA' algorithm excels in the mining of extended item sets and surpasses alternative algorithms by a margin of three to 30 when handling dense data. We conducted three distinct types of experiments to assess how well the MAFIA algorithm performs. Initially, we assessed the impact of each pruning component within the MAFIA algorithm to illustrate its role in streamlining the search space of the itemset lattice. Subsequently, we investigated the efficiency gains achieved through compression when expediting support counting. Finally, we carried out a comparative analysis, pitting MAFIA against other contemporary algorithms across all three data types. This typically excels when applied to dense data featuring lengthy item sets, but it remains a competitive algorithm even when dealing with exceptionally sparse data.

Decision tree algorithm:

At the heart of the ensemble method lies the concept that when multiple weak learners collaborate, they can collectively form a robust learner, consequently enhancing the model's accuracy and precision **[1]** .In the realm of decision trees, one encounters a tree-like structure resembling a flow chart. In this framework, Leaf nodes hold class distributions or class information, whereas interior nodes represent attribute tests and branches indicate test results. **[3].** Fig-4 highlights this versatile tool find utility across a range of domains, serving purposes in various fields. The most well-liked and important approach for creating decision trees is called ID3 (Iterative Dichotomiser-3). The key issue is choosing the right characteristic for the tree's root and branch nodes by taking into account variables including entropy (E), information gain (IG), Gini index, gain ratio, reduction in variance, and Chi-Square statistics.**[13].**

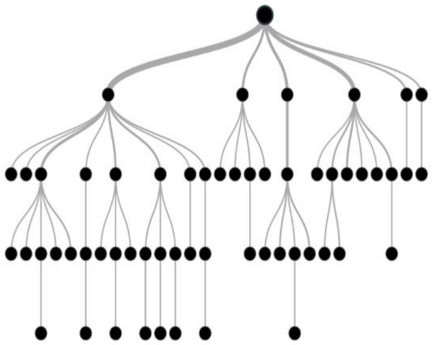


Fig-4 Decision tree

Decision trees prove valuable in addressing both classification and regression tasks, guided by their tree-like structure resembling a flowchart. Beginning at the root node and culminating in outcomes determined by the terminal leaves, decision trees visually illustrate predictions based on sequential feature-based partitions. To construct trees for training data samples D, high entropy inputs are utilized. These trees are efficiently created using a recursive, top-down divide and conquer (DAC) approach. Tree pruning is then performed on D to remove unnecessary samples [9].

$$\textbf{Entropy} = -\sum_{J=1}^{m} \boldsymbol{P_{ij}\, log_2\, p_{ij}}$$

Decision trees apply recursive partitioning, dividing each node into child nodes until a predetermined stopping condition is satisfied. This premise relies on the effective subdivision of data into smaller, more manageable subsets. Outliers can exert influence on the construction of decision trees, rendering them sensitive to such extreme values. Managing outliers effectively may necessitate the use of preprocessing techniques or robust methodologies.

Random forest algorithm:

The Random Forest Algorithm, a widely embraced supervised machine learning technique, finds immense popularity in the field **.** The Random Forest ensemble classifier is designed to construct and integrate multiple decision trees to optimize outcomes. It primarily involves tree- based learning by aggregating bootstrapped samples.

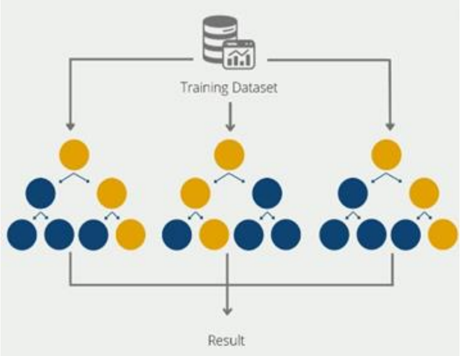


Fig-5 Random forest

By merging the outputs from several decision trees, each of which is trained on different subsets of the data, the Random Forest (RF) classifier increases prediction accuracy.The maximum number of trees in the RF typically yields the highest accuracy. This method prevents overfitting and contributes to high performance **[6].** Throughout the training process, Several decision trees are generated by Random Forest, and labels are assigned according to the majority, as illustrated in Fig-5. Unlike decision tree methods, Random Forests incorporate randomness in selecting the root node and in determining how to split the nodes **[10].**

When provided with the dataset X = {x1, x2, x3,..., xn} and the associated results Y consists of {y1, y2, y3,..., yn}, where the minimum value is set at b = 1 and the maximum value is B, the forecast for an individual sample is, denoted as x0 .To predict the outcome for x0, marked as P(x0), it is calculated by taking the average of predictions made by each of the B individual trees:

$$\boldsymbol{Pb = 1 \wedge B\big(fb(x0)\big)}$$

$$\boldsymbol{P(x0) = (1/B) * \sum b = 1^{A} B\; fb(x0)}$$

In this formulation, fb(x0) signifies the prediction generated for the sample x0 **[1].**

5. Results

We have addressed a range of approaches and technologies pertaining to the prognosis and management of cardiac disease, with a particular focus on the importance of data mining methodologies and numerous machine learning algorithms. These methods have been employed to simplify automatic diagnosis in medical centers and enhance patient care. We explored the use of diverse parameters, including gender, age, cerebral palsy (CP), blood pressure BP, among others, to predict heart disease using five distinct algorithms: Decision trees, Random Forest, k Nearest Neighbor (KNN), and Artificial Neural Networks (ANN). Our study sought to determine the best reliable method for predicting heart disease based on health parameters, and our experiments revealed that the greatest accuracy rate of 88% was attained by Naïve Bayes. To validate our approach, we conducted user research involving 20 participants, and the HDV is technology proved valuable in enhancing graphic interpretation workflows. These findings underscore the importance of prioritizing legislative measures within the medical field to reduce the incidence of heart disease. When contrasting KNN, Naive Bayes, and Decision Tree classifiers to distinguish between normal and abnormal cardiac beats, Decision Tree is found to be the most efficient method. This method shows promise in correctly identifying alignments relevant to the heart. Utilizing feature selection to reduce data complexity, machine learning techniques are used to predict and understand the signs of cardiac disease. **[7].**

6. Discussion

In this study, a survey that was carried out between 2004 and 2015 provides an insight of the many models that are accessible and the various data mining approaches used. Also stated is the accuracy these models produced. It has been noted that big data analytics have not been utilized in all methods. The combination of big data analytics and data mining will likely produce the most accurate prediction model design. Due to the ability of machine learning algorithms to find hidden patterns in data,they are crucial for precisely predicting heart illness since they forecast results and improve performance using historical data. Our capacity to anticipate and identify heart illness is enhanced by these programmes, and deep learning—which is fueled by artificial neural networks—is critical for handling complex computations on enormous volume of data .In the medical field, heart disease prediction is challenging but important. However, the death rate can be drastically decreased as soon as practically practicable if the sickness is detected in its early stages and preventative measures are implemented as soon as possible.

References

[1] PRONAB GHOSH, SAMI AZAM,MIRJAM JONKMAN, (Member, IEEE), ASIF KARIM , F. M. JAVED MEHEDI SHAMRAT, EVA IGNATIOUS , SHAHANA SHULTANA , ABHIJITH REDDY BEERAVOLU, AND FRISO DE BOER “Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques”

[2] Prerna Jain, Amandeep Kaur, “Big Data Analysis for Prediction of Coronary Artery Disease”

[3] Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M “Heart Disease Diagnosis Using Data Mining Technique”

[4] Salma Banu N.K, Suma Swamy “Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey”

[5] Priyanka N B.E., (M.Tech), Dr.Pushpa RaviKumar B.E., M.Tech., Ph.D “Usage of Data mining techniques in predicting the Heart diseases – Naïve Bayes & Decision tree”

[6] AZAM MEHMOOD QADRI, ALI RAZA, KASHIF MUNIR, AND MUBARAK S. ALMUTAIRI “Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning”

[7] TAHSEEN ULLAH, SYED IRFAN ULLAH, KHALILULLAH, MUHAMMAD ISHAQ, AHMAD KHAN, YAZEED YASIN GHADI, AND ABDULMOHSEN ALGARNI “Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection”

[8] SENTHILKUMAR MOHAN, CHANDRASEGAR THIRUMALAI, AND GAUTAM SRIVASTAVA“Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”

[9] ABDALLAH ABDELLATIF, (Member, IEEE), HAMDAN ABDELLATEF, (Member, IEEE), JEEVAN KANESAN,CHEE-ONN CHOW (Senior Member, IEEE), JOON HUANG CHUAH (Senior Member,

IEEE), AND HASSAN MUWAFAQ GHENI “An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods”

[10] Rüstem Yilmaz, Fatma Hilal Yağın “Early Detection of Coronary Heart Disease Based on Machine Learning Methods”

[11] Nabaouia Louridi, Samira Douzi and Bouabid El Ouahidi “Machine learning based identifcation of patients with a cardiovascular defect”

[12] AWAD BIN NAEEM, BISWARANJAN SENAPATI, (Senior Member, IEEE), DIPEN BHUVA, ABDELHAMID ZAIDI, ABHISHEK BHUVA, MD. SAKIUL ISLAM SUDMAN, AND AYMAN E. M. AHMED “Heart Disease Detection Using Feature Extraction and Artificial Neural Networks: A Sensor-Based Approach”

[13] SUBHASH MONDAL, (Member, IEEE), RANJAN MAITY , (Senior Member, IEEE), YACHANG OMO , SOUMADIP GHOSH , (Member, IEEE), AND AMITAVA NAG, (Senior Member, IEEE” An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches”

[14] Adedayo Ogunpola, Faisal Saeed, Shadi Basurra, Abdullah M. Albarrak, Sultan Noman Qasem“Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases”