DETECTION OF CARDIOVASCULAR DISEASE USING A.I & M.L

A Project report submitted in partial fulfillment of the requirements of the award of the degree of

Bachelor of Technology

in

Computer Engineering (Artificial Intelligence)

by

Divyansh Johari, PCE21CA020
Vidhan Solanki, PCE21CA004
Hardik Sharma, PCE21CA022
Aaditya Shukla, PCE21CA003

under the guidance of

Dr. Kamlesh Gautam, Assistant Professor (Computer science engineering)

(Session 2022-23)

Department of Advance Computing

Poornima College of Engineering

ISI-6, RIICO Institutional Area, Sitapura, Jaipur – 302022

December 2022

Department Certificate

This is to certify that Mr. Divyansh Johari, registration no. PCE21CA020, of the Department of Advance Computing, has submitted this project report entitled "DETECTION OF CARDIOVASCULAR DISEASE USING A.I & M.L" under the supervision of Dr. Kamlesh Gautam, working as Assistant Professor in the department of Advance Computing as per the requirements of the Bachelor of Technology program of Poornima College of Engineering, Jaipur.

Dr. MIthilesh Arya
Dy. Head of Department,
Dept. of Advance Computing

Ms. Archika Jain
Coordinator-Project

CANDIDATE'S DECLARATION

I hereby declare that the work which is being presented in this project report entitled "DETECTION OF CARDIOVASCULAR DISEASE USING A.I & M.L" in the partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Engineering(Artificial Intelligence), submitted in the Department of Advance Computing, Poornima College of Engineering, Jaipur, is an authentic record of my own work done during the period from July 2022 to Dec 2022 under the supervision and guidance of Dr. Kamlesh Gautam.

I have not submitted the matter embodied in this project report for the award of any other degree.

Signature	Signature
Name of Candidate: Divyansh Johari Registration no.: PCE21CA020	Name of Candidate: Vidhan Solanki Registration no.: PCE21CA0
Signature	Signature
Name of Candidate: Hardik Sharma Registration no.: PCE21CA0	Name of Candidate: Aaditya Shukla Registration no.: PCE21CA003

Dated: -
Place: Jaipur

SUPERVISOR'S CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Signature)
Dated: (Dr.Kamlesh Gautam)
Place: Jaipur (Assistant Professor)
Poornima College of Engineering, Sitapura, Jaipur

ACKNOWLEDGEMENT

I would like to convey my profound sense of reverence and admiration to my supervisor Dr.Kamlesh Gautam, Assistant Professor, **Department of Computer Engineering, Poornima College of Engineering,** for his intense concern, attention, priceless direction, guidance, and encouragement throughout this research work.

I am grateful to **Dr. Mahesh Bundele**, Director of Poornima College of Engineering for his helping attitude with a keen interest in completing this dissertation in time.

I extend my heartiest gratitude to all the teachers, who extended their cooperation to steer the topic toward its successful completion. I am also thankful to the non-teaching staff of the department to support in the preparation of this dissertation work.

My special heartfelt gratitude goes to **Dr. Mithilesh Arya, Dy. Head, Department of Advance Computing, Ms. Archika Jain, Project Coordinator, Department of Advance Computing, Poornima College of Engineering**, for unvarying support, guidance, and motivation during the course of this research.

I would like to express my deep sense of gratitude towards the management of Poornima College of Engineering including **Dr. S. M. Seth**, Chairman Emeritus, Poornima Group, and former Director NIH, Roorkee, **Shri Shashikant Singhi**, Chairman, Poornima Group, **Mr. M. K. M. Shah**, Director Admin & Finance, Poornima Group, and **Ar. Rahul Singhi**, Director Poornima Group for the establishment of the institute and for providing facilities for my studies.

I would like to take the opportunity of expressing my thanks to all faculty members of the Department, for their kind support, technical guidance, and inspiration throughout the course.

I am deeply thankful to my parents and all other family members for their blessings and inspiration. Last but not least I would like to give special thanks to God who enabled me to complete my dissertation on time.

Divyansh Johari , Department of Advance Computing, <PCE21CA020 >
Vidhan Solanki, Department of Advance Computing, <PCE21CY004 >
Hardik Sharma, Department of Advance Computing, <PCE21CA022 >
Aaditya Shukla, Department of Advance Computing, <PCE21CA003>

Table of content

Abstract

In the 21st century according to stats, the risk of cardiovascular disease has become more common and the death rate caused by it is increasing way more. around 17.9 million lives are taken by CVDs each year. There are various reasons causing it but most importantly it is caused by not identifying it, hence it becomes a very important task for the human race to identify the CVDs and deal with the proper treatment so that the death dance caused by CVDs can be decreased and risk of it at an early age too. This work mainly aims to review and analyze various methods and approaches to detect the presence or absence of CVDs using AI and ML with accurate predictions. An artificial intelligence system for detecting heart disease from phonocardiogram (PCG) signals has been developed utilizing Artificial Neural Networks (ANN) algorithms and also by driving various A.I algorithm on the given electrocardiogram (ECG) data of the patients we can predict the absence or presence of CVDs.

Heart disease, another name for cardiovascular illness, is a serious global issue that has a big impact on people. According to a recent study, cardiac disorders were responsible for millions of deaths worldwide, or 31% of all fatalities. Medical research has shown that some risk factors increase a person's likelihood of developing heart disease (CVD). According to, some of these factors an unhealthy diet, nicotine use, depression, stress, excessive alcohol use, physical inactivity, inherited obesity, and age are the common causes of CVD. The World Health Organization has published several papers showing an increase in CVD-related deaths, which are primarily attributable to inadequate preventative actions despite rising risk factors.

For successful prognosis of cardiovascular diseases (CVDs), an early and quick diagnosis is essential. Heart disease and strokes are the predominant causes and account for more than 80% of CVD deaths, whilst one-third of these deaths occurs prematurely in people under 70 years of age. For CVD diagnosis, patients need to show an elevated level of biomarkers in the blood sample associated with severe pain in the chest, and diagnostic electrocardiogram (ECG). The majority of CVD patients making CVD diagnosis difficult for physicians show a surprisingly normal ECG pattern. Artificial intelligence techniques can radically improve and optimize CVD outcomes. AI has the potential to provide novel tools and techniques to collect and interpret data and make faster and more accurate decisions reducing hospitalization cost, thereby increasing the quality of life. AI has also improved medical knowledge by unlocking clinically relevant information from the voluminous and complex data received from various resources. This paper reviews various AI-ML techniques, which can effectively be used for early and accurate detection of CVD, thereby improving cardiac care

CHAPTER 1

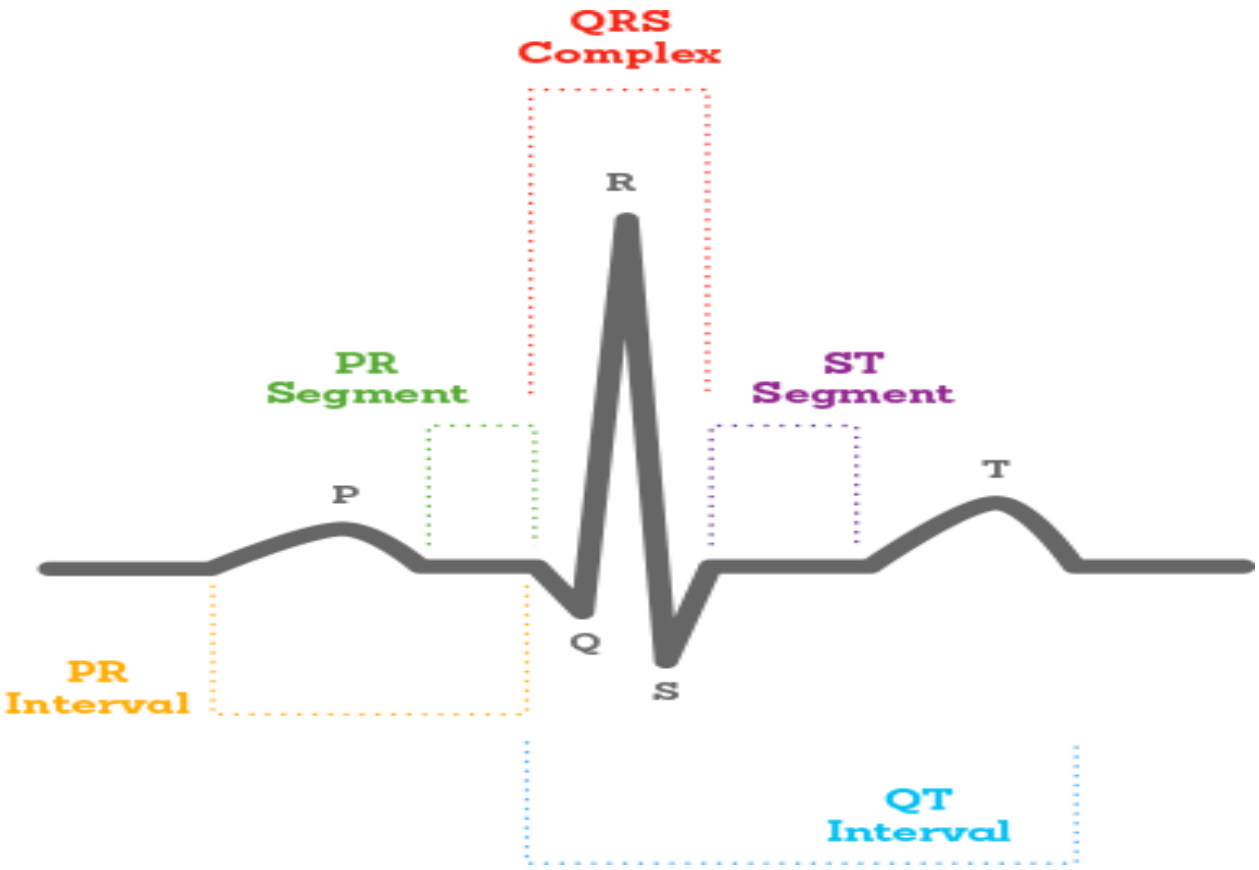
INTRODUCTION

The heart is one of many organs in the human body that provides blood supply through a function akin to a pump. A healthy heart is fundamental and necessary for human well-being. The leading cause of death in the modern period is cardiovascular disease (CVD), generally known as heart disease. The classification of associated disorders is a challenging undertaking that involves several biological markers and risk factors due to the highly complicated mechanism of the heart. Professionals in related fields employ cardiac physiological signals like the electrocardiogram (ECG) and phonocardiogram (PCG) to monitor or detect cardiovascular-related disorders. Now due to the heart's complex structure and the death rate by the CVDs have grasped the attention of various researchers and scientists to find and perform various approaches, techniques, and methods that are required in order to detect the CVDs presence or absence of a patient. hence some scientists came up with the output of using AI and various Machine-learning techniques

Cardiovascular diseases (CVDs) are mostly brought on by the buildup of plaque in the arterial walls, which results in atherosclerosis, a dangerous condition that causes artery narrowing. A heart attack or stroke can result from the narrowing of arteries, which restricts blood flow and makes it harder for blood to circulate freely. One of the main causes of death worldwide and a significant barrier to sustained welfare and growth for people is cardiovascular disease (CVD). According to a study, low- to middle-income developing countries account for the bulk of CVD cases.

Early disease identification is necessary since there are more people with cardiovascular disease. Cardiovascular disease can be identified by biochemical testing on patient-provided samples of blood, urine, or tissue. Basic biochemical risk factors for heart disease diagnosis, such as blood pressure, smoking, glucose, cholesterol, low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and physical inactivity, are additional signs. Additionally, blood tests using biochemical methods are performed to identify cardiovascular disease. These tests identify blood lipids, including LDL, HDL, and triglycerides as well as blood sugar and glycosylated hemoglobin, which is used to identify diabetes.

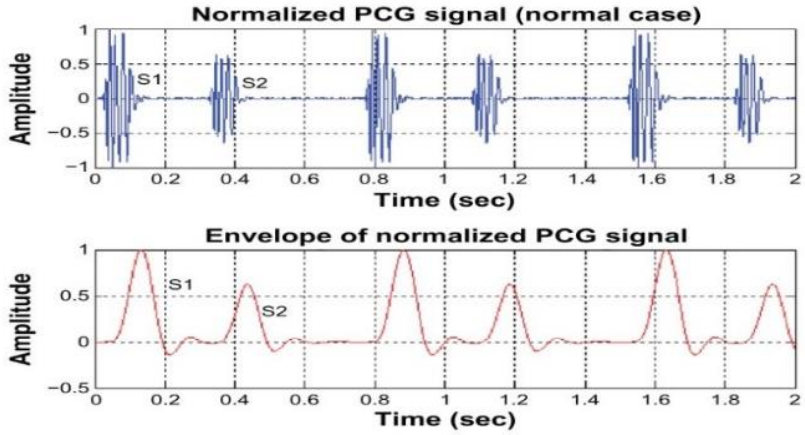
In this paper we've reviewed about 5 algorithm that can be used to detect CVDs and this could be done with the help of data retrieved from ECGs and PCGs So, now to simplify the approach let us first understand more about ECGs and PCG and grasp the main understanding of the Conduction System of heart
[I] Mainly the SA (Sino Atrial) Node, sometimes known as "The Pacemaker," and the AV (Atrio-ventricular) Node, which induce the lower heart to contract while the SA Node contracts the upper heart chambers, create the majority of the electrical signals that are recorded by an ECG. Numerous electrodes attached to the device and to the patient's body make up the ECG. Each sensor detects a shift in the electrical charge underneath the skin to identify impulses. The cells near the heart receive an impulse that travels swiftly. The typical waveform of an ECG was displayed in Figure 1. Therefore, by decoding this impulse, we may discover any abnormal heart behavior that may exist and then apply our algorithm to it to calculate the likelihood of developing a CVD.



The P wave represents atrial depolarization, or the contraction of the atria, followed by the QRS complex, a rotated V-shaped wave that represents ventricular depolarization and atrial repolarization, and finally the T wave, which represents ventricular repolarization, or the relaxation process of the ventricles, may be easily identified in the diagram.

We can determine the patient's heart condition by interpreting this ECG graph by calculating the P waves, PR interval, QRS complex, hearts rhythm, and heart rate (ECG should be 6-second strip) by using the 6-second method. Normally, a healthy heart has about 60-100 BPM caused by the SA node and by 40-60 BPM AV node.

[II] PCG, also known as phonocardiography, is used for the same reason, and this graph shows us the sound wave of the cardiac cycle, which is the heart's contraction and relaxation, which produces the murmurs sound, also known as the Lub-Dub sound of the heart by laypeople.



These recorded sound tracks aid medical professionals by providing knowledge about valve function and the effectiveness of blood pumping beneath the body. The S1 and S2 beats of the typical heartbeat are separated by the PCG method. On the basis of these recorded soundtracks, we can identify the related heart diseases. A normal, healthy heart produces the S1 sound, which represents the closure of valves in the upper and lower chambers, while the S2 sound, which represents the closure of valves with a frequency over 100Hz, occurs. So, this is the fundamental related field answer for identifying cardiac irregularities in order to estimate the likelihood of developing cardiovascular disease (CVD).

A description of the machine learning methods used in this research follows Five well-known classification models—Decision Tree, K-Nearest Neighbor, Logistic Regression, Naive Bayes, and Support Vector Machine—have been developed, and their prediction accuracy has been compared.

Phonocardiogram

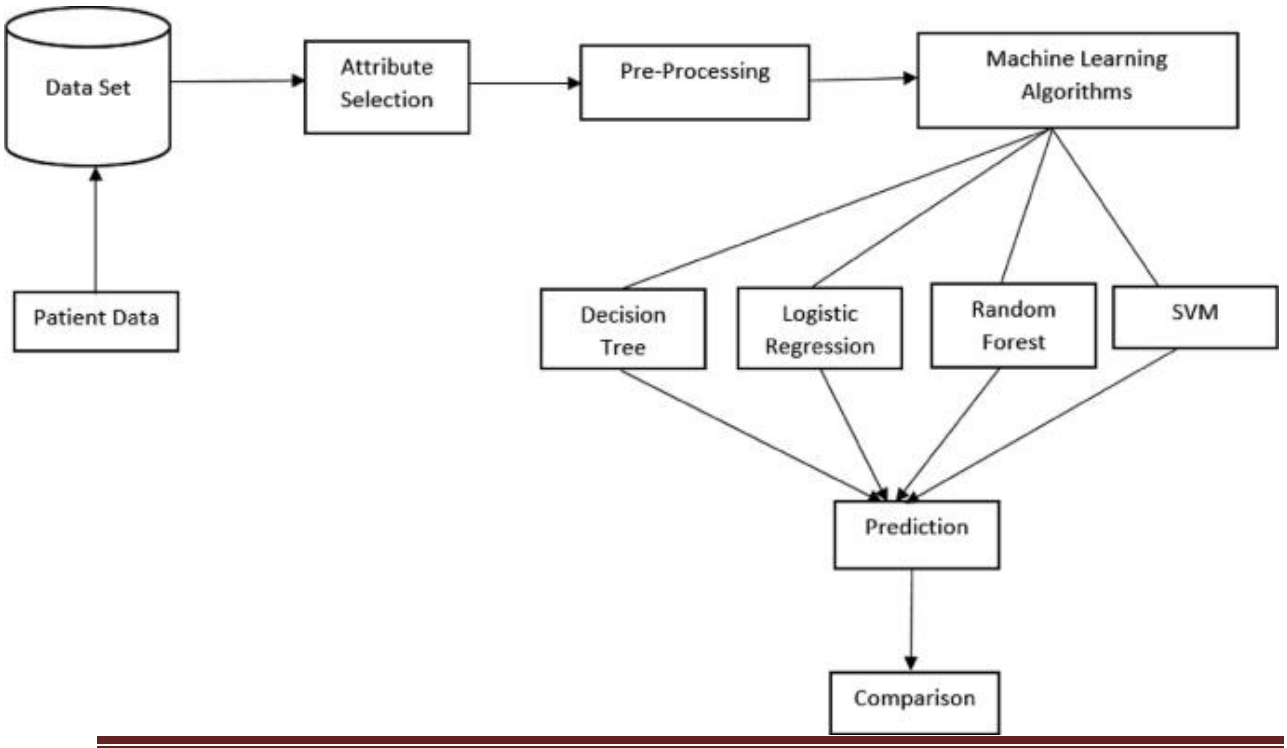
Electrocardiogram

CHAPTER 2

PROBLEM STATEMENT & OBJECTIVE

Detection of Cardiovascular Disease using A.I & M.L Algorithms

The research that is suggested in this study focuses mostly on different machine learning techniques used to forecast cardiac disease. The main organ of the human body is the heart. In essence, it controls the flow of blood throughout our body. Any heart irregularity can exacerbate pain in other body areas. Heart disease refers to any condition that impairs the heart's regular operation. In today's modern society, heart disease is one of the main causes of most fatalities. Heart disease can be brought on by living a sedentary lifestyle, smoking, drinking alcohol, and eating a lot of fat, which can raise blood pressure. The World Health Organization estimates that more than 10 million people worldwide pass away each year as a result of heart disease. A healthy lifestyle and earliest detection are only ways to prevent the heart related diseases.
The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences.
This paper presents performance analysis of various ML techniques such as Naive Bayes, Decision Tree, Logistic Regression and Random Forest for predicting heart disease at an early stage. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyses the given data. After analysis of data ML techniques help in heart disease prediction and early diagnosis.



CHAPTER 3
LITERATURE REVIEW

Literature Review

Paper 01: **Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study**

Paper (26 July 2021)

By: W. M. Jinjri, P. Keikhosrokiani and N. L. Abdullah, "Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study," *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 132-138, Doi: 10.1109/ICIT52682.2021.9491677.

Summary-
Heart disease, also known as cardiovascular disease, is a serious condition that has a big impact on a lot of people's lives. The early detection of heart disease is crucial for reducing the disease's power. The most popular techniques for classification and detection still include machine learning. This study attempts to develop and create a model that best categorizes cardiovascular disease and accurately predicts individuals' propensity for the condition using machine learning techniques. Therefore, to classify data related to cardiovascular illness, this research examines the five most potent machine learning platforms. Support vector machine (SVM), K-nearest neighbor (K-NN), logistic regression (LR), decision tree (DT), and naive bayes (NB) are the five classifiers that have been proposed for the categorization of cardiovascular disease (CVD).
The dataset was taken from the open Kaggle repository in order to validate the work. Applying a variety of performance variables allows for the analysis, evaluation, and comparison of the algorithms' performances. Results show that the approaches of logistic regression and support vector machines (SVM) are the most effective for detecting cardiovascular disease.
Applications for data mining are frequently utilized in the medical field to identify disorders and provide patients with a heart disease diagnosis based on their medical records. The most effective machine learning methods for categorizing cardiovascular illness have been established using patient data, and we have explored these methods in this work. Based on evaluation metrics including precision, recall, f1-score, accuracy, and training time, the various classification algorithms SVM, KNN, DT, LR, and NB have been contrasted. Support vector machine (SVM) and logistic regression (LR) techniques are the most effective for detecting cardiovascular disease, according to our proposed work. The performance of these fundamental categorization methods will be improved in the future by the creation of a meta-model that will be applied to the prediction of cardiovascular disease in human subjects.

Paper 02**: A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data**

Paper (2022)

By: Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M. Afshar Alam, Safdar Tanweer, Guojun Wang, A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data, Medical Engineering & Physics, Volume 105, 2022, 103825, ISSN 1350-4533,

Summary-

The prevalence of cardiovascular illnesses is significantly increasing worldwide. Making cardiovascular prediction as accurate as possible is crucially important. In order to identify cardiovascular disease (CVD) with the highest level of precision and accuracy, a forecast based on machine learning approaches can be useful. Effective disease prediction contributes to early diagnosis, which lowers the death rate. Effective CVD identification and prediction are necessary given a person's medical history and the causes of heart disease. Data analytics is useful for making predictions based on a ton of data, and it helps medical facilities forecast the course of diseases. A sizable amount of patient-related data is regularly kept. Future illness emergence can be predicted using the knowledge acquired.

This study thoroughly assesses the chosen papers and identifies gaps in the literature, allowing researchers to develop and apply their findings in clinical sectors, mainly on datasets pertaining to heart disease. The results of this study will help physicians anticipate potential heart dangers and take preventive action.

Since it extracts more accurate and efficient data from large datasets, machine learning (ML) is frequently recommended for heart disease prediction. It serves as the fundamental building block of machine learning, which facilitates the management of enormous amounts of data, has a quick processing time, and produces predictions in the early phases of development. Applications of machine learning (ML) reduce preventable hospital deaths, enhance illness prevention, early diagnosis, and health policy. Similar research has been done by many researchers, especially figuring out which ML techniques could diagnose cardiac illness.

Paper 03: **Artificial Intelligence Algorithm for Heart Disease Diagnosis using**

Phonocardiogram Signals

Paper (2021)

By: Ibrahim Abdel-Motaleb and Rohit Akula Department of Electrical Engineering, Northern Illinois University, DeKalb, IL 60115 ibrahim@niu.edu

SUMMARY:
To detect cardiac disease from Phonocardiogram (PCG) signals, an artificial intelligence system has been built using Artificial Neural Networks (ANN) algorithms. The neural network receives four new signal features as input: activity, complexity, mobility, and the spectral peaks from the power spectral density plots. In this study, the accuracy of the neural networks was evaluated using 94 PCG signals for three cardiac disorders. The features are given to the neural networks after the signals have been filtered and the feature properties have been retrieved. The Radial Basis Function (RBF) network and the Back Propagation Network (BPN) approaches are used for classification. The accuracy of both structures is assessed using a calculation known as receiver operating characteristic (ROC). According to the findings, RBF had a 98% accuracy rate in predicting the disease compared to BPN's 90.8%. The created artificial intelligence system has been demonstrated to be an effective method for PCG signals-based automatic diagnosis of cardiac disorders.
In this study, 94 human participants' sick cardiac sound signals are preprocessed. These people have three diseases: coarctation of the aorta (disease 2), mitral stenosis (disease 1), and 32 have mitral regurgitation (disease 1). (Disease-3). 94 signals were used in total, of which 66 were used for training, 5 for validation, and 23 for testing. Using the Wavelet transform, the prepossessing was carried out. The PCG signals' four separate feature characteristics are extracted. Both the radial basis functions network algorithm and the conventional back-propagation network algorithm get these properties as inputs. The 66 samples were used to train the two networks. 23 samples from the three distinct diseases were used to test the networks.

Comparison Table

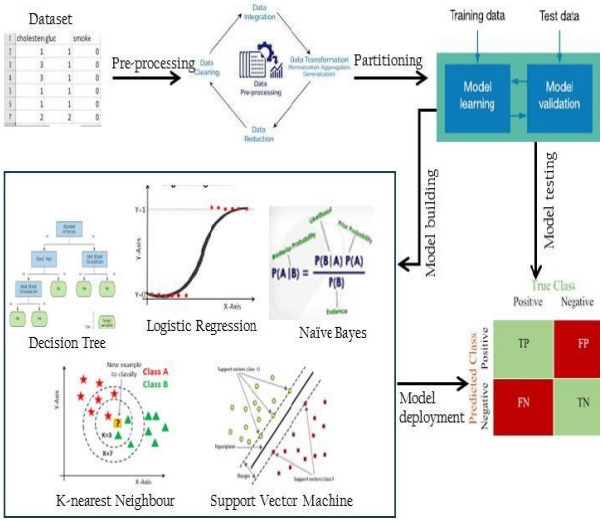
S.No	Paper title	Author's Name	Year	Approach used	Finding	S/w and H/w Required
1	**Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study**	W. M. Jinjri, P. Keikhosrokiani and N. L. Abdullah	2021	**Machine Learning Algorithms-** Decision Tree, K-Nearest Neighbor, Logistic Regression, Naive Bayes, and Support Vector Machine	Logistic Regression and svm are the most accurate and widely iused models for ml detection of heart related diseases	
2	A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data	Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M. Afshar Alam, Safdar Tanweer, Guojun Wang	2022	machine learning approaches using medical big data		
3	**Artificial Intelligence Algorithm for Heart Disease Diagnosis using Phonocardiogram Signals**	Ibrahim Abdel-Motaleb and Rohit Akula	2021	Machine algorithms applied on the Data of Phonocardiogram signals	Detection of heart disease using Pcg is highly accurate and is based on Radial Basis Function and Back propogation Network	

CHAPTER 3
PROPOSED APPROACH

In medical field , classification is one of the most important , critical and popular decision making tools. There have been numerous computational intelligence methods that support the medical healthcare domain [15] –[17] .Some of the work done by other researchers of same field area.

Some of them are stated here : A.S. Ebenezer , S.J. priya and their group selected ten different algorithms to classify coronary artery disease risk assessment. ANN , decision tree (DT), naïve bayes(NB),SVM, random forest , CHAID , etc. From there research they found NB and SVM is good at predicting .
A group of H. mansoor , R. Segal[22] , they compare different algorithms like Random forest and logistic regression (LR) ,they used these techniques for predicting the risk of heart diseases. And result of there research is that LR is doing slightly well then other one.

The dataset for categorization is first collected as part of the approach for this study. One of the key jobs in data mining is classification, whose goal is to group documents into one or more classes or categories as a result, this work creates a useful technique for the dataset used to categorize cardiovascular disease. In order to achieve this, it is necessary to make the assumption that there are only two classes: the positive class with unknown results and the negative class without surprising findings.
Anaconda Jupyter Notebook, a Python 3 application, is used to implement the algorithms. The datasets are split into training and test sets after pre-processing. The majority of researchers opt for a 70:30 split (70 percent for training and 30 percent for testing), as more training data results in more optimal and accurate outcomes. As a consequence, the 70:30 partitioning ratio is employed.



3.1. Dataset- The dataset used to evaluate and contrast the methods used in this work was obtained from the Kaggle web repository. The dataset, which consists of 77,000 patient clinical trial records gathered by hospitals for cardiovascular illnesses, has three input components: examination (outcomes of medical investigation), subjective (realistic information), and objective (realistic information) (data obtained from a patient). Eleven attributes total, including one target variable with the label "(Absence or Presence) for diagnosis," four objective features, four examination features, three subjective features, and four other attributes, make up the dataset. Table 1 provides summary of the CVDs dataset collected for the study.

Table 1 Description of the Dataset

	Attributes	Input Features	Data Type/ Description
1	Age	Objective features	Int / days
2	Height		Int / centimeters
3	Weight		Float/ kilograms
4	Gender		Categorical code 1: male, 2: female
5	Systolic blood pressure	Examination features	Int/
6	Diastolic blood pressure		Int/
7	Cholesterol		1: normal, 2: above normal, 3: well above normal
8	Glucose		1: normal, 2: above normal, 3: well above normal
9	Smoking	Subjective features	Binary
10	Alcohol		Binary
11	Physical activity		Binary
12	Cardiovascular	Target	Presence or absence of CVD / target variable.

3.2.Classification Methods-

A description of the machine learning methods used in this research follows Five well-known classification models—Decision Tree, K-Nearest Neighbor, Logistic Regression, Naive Bayes, and Support Vector Machine—have been developed, and their prediction accuracy has been compared. Numerous research contrasted data mining techniques using various parameter settings. The majority of these earlier research concluded that these techniques outperformed their statistical equivalents because they were less restricted by presumptions and produced better categorization outcomes. A few of these techniques are covered in brief.

3.2.1.Linear Regression - Logistic regression (LR). is one of the most often used machine learning methods for analyzing multivariate regression issues in the medical industry**.** Using a continuous independent variable that aids in both the diagnosis and prediction of illnesses, LR is used to anticipate the outcome of a dependent variable. It is a method for discriminating between categories that use the input vector to extract important statistical data points from the model or forecast data trends. In the LR, the dependent variable is a binary variable that only accepts data that is coded as 0 (yes, success, etc.) or 1. (no, failure, etc.). Calculating the log chances of an event is the basic goal of an LR analysis. As shown mathematically, LR calculates multiple linear regression functions as follows:

$$log\frac{p(y=1)}{1-(p=1)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k;$$

3.2.2.Support Vector Machine: SVM was first introduced in 1992 by Boser, Guyon, and Vapnik as a binary classification method. Since then, a number of classification and regression problems have been added to its list of uses. Without adding prior knowledge, its generalization performance makes it a good classifier even with a lot of data. By choosing the best hyperplane that maximises the margin between classes, data with separable classes can be categorised. SVM is modelled as a collection of vector spaces with limited dimensions, where each dimension corresponds to a unique property of an item. Issues in high-dimensional space have been successfully addressed using this method. Due to its processing efficiency on huge datasets, SVM has demonstrated great performance for illness prediction in the medical field in recent years. The main goal is to reduce generalization mistakes and develop it as a supervised learning system for regression and classification applications.
The SVM is mathematically expressed as:

$$\text{If } Y_i = +1; wx_i + b \geq 1 \qquad (2)$$

$$\text{If } Y_i = -1; wx_i + b \leq 1 \qquad (3)$$

$$\text{For all i: } y_i(w_i + b) \geq 1 \qquad (4)$$

The above equations shows "w" which means weight and "x" as a vector point. Therefore, the data in (3) and (4) must be constantly larger than zero and below zero, respectively, in order to distinguish the data in (2). SVM selects the hyperplane with the greatest distance between it and all other potential hyperplanes.

3.2.3. The K-NearestNeighbour Method - The (K-NN) technique classifies cases based on how closely they resemble one another. When a case is new at a certain location, its distance from each model case—which is calculated as the nearest neighbor and is the most comparable to the approach that suggests the case—indicates the case. In this manner, the case is added to the output of neighbors that are closest to it. A new input class label is predicted by the K-NN algorithm, which bases its prediction on how similar the new input is to samples of its input from the training set. The K-NN classification output is poor if the new input is identical to the training set's samples.
The distance between two points x(xo, x′) is calculated using the following mathematical formula, where p:x*x=R is a function that gives distance.

$$p(x, x^i) = \left|x - x^i\right| = \sqrt{\sum_{i=1}^{d}(x_i - x_i')^2}$$

3.2.4.Decision Tree : A supervised learning strategy called a DT is applied mostly to classification-related problems by evaluating the entropy and gain of information, DT classifies the data using decision rules derived from the training data. A classification tree structure where each node represents a property The root node will be the primary node, followed by the child nodes. The leaf nodes then represent the decision's result. It works well with both categorical and continuous characteristics. With DT, the population is split into two or more groups based on important predictors. The entropy for each feature is computed as the first step in DT. The dataset was then divided depending on the variables and predictors, either with high data gain or low entropy. The two phases are followed by the remaining qualities, as mentioned.

$\text{Entrophy(E)} = \sum_{k=1}^{l} -_{qk} \log_{2qk}$

"qk" is the proportion of the count of the kth class procedures to the total count of models, whereas "l" referred to a response variable module count.

3.3. Evaluation Method - Specific measures, including as the f1 score, precision, recall, and accuracy, are used as the foundation. We also keep track of how long each algorithm takes to train. The classification algorithm's output is displayed in the confusion matrix, and this information forms the basis for further parameter calculations. As a result, the confusion metrics may examine a model's correctness and determine whether a classification algorithm frequently labels items with the incorrect labels by comparing anticipated values with actual values. The confusion matrix's values and its depiction in Table 2 are quickly described by the parameters below.

True positive (TP) circumstances occur when the datapoint's actual class and projected class are both 1.

- False positive (FP) scenarios take place when a data point's real class is 0, despite the fact that the predicted class is 0.
- False negative (FN) scenarios occur when a data point's real class is 0 and its anticipated class is 1.
- True negative (TN) circumstances occur when a data point's real class is 1 and its anticipated class is 0, respectively.

Table II- Confusion Matrix

		Actual value	
		Classified as absence	***Classified as presence***
Predicted value	*Absence*	TP	FN
	Presence	FP	TN

3.3.1.F1 score

When an f1-score achieves its highest value at 1 and its poorest score at 0, it is a function that is understood as a weight of recall average and precision. The f1-score formula is as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.3.2.Recall

Recall measures how much pertinent data is recovered from any machine learning system. The capacity to locate all connected events in the data is the main focus. The recall is represented by the equation below:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

3.3.3.Precision

Being precise means being exact and correct. Precision conveys the sense of incidents that were accurately expected. It measures the proportion of genuine positives among all positives and quantifies forecasts that fall into the positive category as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

3.3.4.Accuracy

An important metric for describing an algorithm's performance is accuracy. It establishes the threshold at which an algorithm can accurately forecast both positive and negative instances and is quantified by the following formula:

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

3.4.Result of the algorithm –

Five artificial intelligence techniques were used in this work to classify the cardiovascular disease dataset: decision trees (DT), logistic regressions (LR), K-nearest neighbors (KNN), and support vector machines (SVM). The dataset consists of 70,000 samples and 11 attributes. The classes can be examined both with and without the disease.

35021 of the data samples are marked as missing, whereas 334979 are marked as suffering from cardiovascular illness. Training and testing sets are split up into the data in the ratio of 70:30.

The confusion matrix of the DT, LR, KNN, and SVM prediction results is shown in Tables 1 through 5.

		Predicted Value		
		Absence	Presence	Actual value
Actual value	Absence	6562 (31.35%)	3899 (18.46%)	10461
	Presence	3775 (18.03%)	6764 (32.16%)	10539
	Total Predicted	10337	10663	21000

Table 3 – Confusion Matrix for Decision Tree

		Predicted Value		
		Absence	Presence	Actual value
Actual value	Absence	8243 (39.25%)	2296 (10.93%)	10539
	Presence	4031 (19.20%)	6430 (30.62%)	10461
	Total Predicted	12274	8726	21000

Table 4 – Confusion Matrix for KNN

		Predicted Value		
		Absence	Presence	Actual value
Actual value	Absence	5363 (38.31%)	1625 (11.61%)	6988
	Presence	2244 (16.03%)	4768 (34.06%)	7012
	Total Predicted	7607	6393	14000

Table 5 – Confusion Matrix for Logistic Regression

		Predicted Value		
		Absence	Presence	Actual value
Actual value	Absence	5671 (40.51%)	1317 (9.41%)	6988
	Presence	2509 (17.92%)	4503 (32.16%)	7012
	Total Predicted	8180	5820	14000

Table 6 – Confusion Matrix for SVM

Hence the best-performing algorithm according to the Confusion Matrix, F1 scores and precision is examined and selected for identifying the CVD in an individual.

4.Deep Learning Methods applied to ECG

The development of thorough, human-like interpretation abilities is one of the main goals of using AI for ECG interpretation. Since the creation of the digital ECG more than 60 years ago, attempts have been made to provide a quick, accurate, and thorough computer-generated interpretation of the ECG. The problem seems to be solved since ECG interpretation is a fairly constrained

application of pattern recognition to a small dataset. Early digital ECG interpretation software was capable of quickly identifying fiducial sites, taking precise measurements, and defining common measurable anomalies. Modern technology has Advance past these rule-based methods to find patterns in vast amounts of labeled ECG data.
AI-driven algorithms have been developed by a number of organizations, and some of these algorithms are already in use in a few limited clinical scenarios. Research has employed a large number of single-lead ECG datasets to build CNNs, which have then been applied to 12-lead ECGs. One team, for example, employed a CNN to categories 2 million tagged single-lead ECG recordings from the Clinical Outcomes in Digital Electro cardiology study into six different categories of anomalies on the 12-lead ECG. The technique was shown to be viable in this work, but its wider application or external validation in other 12-lead ECG datasets is still to come. Another team discovered that CNN's could make some diagnosis more reliably than practising cardiologists in comparable studies utilising single-lead ECGs. It remains to be seen, nevertheless, if this method will result in 12-lead ECG interpretation software that is therapeutically relevant. A CNN was created for the multilabel identification of 21 different heart rhythms based on the 12-lead ECG in an assessment that was published in 2020 utilizing a training and validation dataset of >80,000 ECGs from >70,000 patients. The cardiologists' committee's consensus labels served as the reference standard. The best network performed much better than a single cardiologist interpretation in a test dataset of 828 ECGs, matching the gold standard labels in 80% of the ECGs. The model's sensitivity, specificity, and mean area under the curve (AUC) receiver operating characteristic scores were 98%, 87%, and 99%, respectively.

The team of researchers developed a thorough ECG-interpretation infrastructure using their own dataset of >8 million ECGs done for clinical purposes (all of which have been labeled by experienced ECG readers and are connected to the relevant electronic health record). They showed that a CNN has good diagnostic performance and can recognize 66 distinct codes or diagnosis labels. Recently, Researchers created a unique technique that translates ECG characteristics into ECG codes and text strings using a transformer network and a CNN to extract ECG features. By providing information in a same manner and using comparable language, this approach produces a model output that is more like that of a human ECG reader. It also makes sense of related codes, preventing the display of opposing or mutually conflicting interpretations that a human reader would not present. This approach will be especially important as our reliance on ECG data collected by cutting-edge, consumer-facing apps that are greatly scalable increases. For instance, single-lead ECG traces collected from mobile, smartwatch-enabled recordings have been subjected to AI-ECG algorithms for the diagnosis of AF. The democratization of ECG technology will cause the volume of signals that need to be interpreted to increase quickly, maybe faster than the rate at which human ECG readers can handle them. These autonomous, consumer- or patient-facing models are projected to be essential for telehealth technology. They could also make it possible to build essential lab spaces with the capacity to store and process massive quantities of data.

The signal quality produced with these devices can vary, as seen in the wristwatch research cited above, and AI-ECG may be less able than human expert over-readers to classify the heart rhythm utilizing inferior tracings. Similar to this, another study discovered that a deep neural network built using ECG recordings from smartwatches performed well for passive detection of atrial fibrillation (AF) in comparison to AF diagnosed from 12-lead ECGs, but that performance was noticeably less reliable when referencing a self-reported history of persistent AF.

Nevertheless, despite significant advancements achieved, a full, human-like ECG interpretation package is still a long way off. Even in its most Advance form, the package doesn't have the precision required for execution without human supervision. The interpretation of an ECG generated by a machine also has the potential to affect human over readers and, if erroneous, can be a cause of bias or systemic mistake. This worry is especially important if the algorithms were developed in populations that are different from the populations where they are used.
This weakness highlights the necessity for a diversified derivation sample, demanding external validation studies, gradual implementation, and continuous model performance and effectiveness evaluations (ideally including a diverse patient population and diverse means of data collection that reflect real-world practices).

5.Deep Learning Methods applied to Phonocardiogram

The first fundamental analytic technique used to assess the heart's functioning status is cardiac auscultation. If the heart sound from the Phonocardiogram (PCG) test indicates any abnormalities, an electrocardiography (ECG) test is required. The ability to hear and see the heart auscultation on the screen increases confidence in the precision of the first diagnosis. However, the diagnosis could not be as precise as expected due to noise and human error. If an artificial intelligence computer is utilized to generate potential diagnoses utilizing some distinguishing characteristics of the heart sound waves, diagnosis accuracy can be significantly improved. This method should lower death rates and healthcare expenses.
Numerous articles have put forth various methods for deriving characteristics from heart sounds and categorizing them using neural networks. Mohamed and Raafat created a mathematical model in the late 80s to use a limited set of parameters to characterize cardiac murmurs and noises. The smallest difference between the characteristics of the observed pattern and the reference patterns was used to classify the data in this case. Features were extracted using a fourth-order linear prediction of the cardiac cycle frames.
Patil and Kumaraswamy proposed an intelligent heart attack prediction system based on data mining and artificial neural networks. By applying the K-means clustering algorithm to the supplied data, this technique computes the parameters crucial to the heart attack. The Maximal Frequent Itemset Algorithm is used to extract these common patterns from the data (MAFIA). Following that, the designs are chosen based on the calculated significant weightage. Although the aforementioned study claimed that this technology may detect heart attacks using the MAFIA algorithm, the prediction accuracy for the work was not stated. Additionally, rather than using feature characteristics of the heart sound signal, this method makes use of features that correspond to the subject's behavioral patterns, such as drinking and smoking.
The GAL (Grow and Learn) algorithm is a revolutionary technique for segmenting heart sounds utilizing homomorphic filtering and feature extraction from wavelet coefficients. This method's accuracy was estimated to be 90.9%. In their study on the analysis of heart sounds for symptom identification, Reed et al. used wavelet decomposition to segment and alter the heart sounds. By eliminating levels with the shortest scales, the altered vectors were condensed into lower vector sizes. A three-layer neural network was used to classify each vector, and it provided 100% accuracy for all heart sounds. The drawback of this method is the requirement for using several hidden layer neurons—up to 50 layers.
Arrhythmia categorization based on heart rate variability (HRV) has been documented. This strategy is built on the General Discriminant Analysis (GDA) and Multi Layer Perceptron (MLP) techniques. The outcomes showed that this approach produced 100% accuracy for the data the

authors obtained from the MIT-BIH database. But instead of PCG signals, this approach employs HRV signals based on the ECG. It should be highlighted that getting an ECG signal does not qualify as a regular test for primary care doctors since it necessitates laboratory setups, which takes time and is less efficient financially.

CHAPTER 4

CONCLUSION & FUTURE SCOPE

Here we used we different techniques algorithms and datsets related to medical field in different ways like ECG(Electrocardiography) , number of heartbeats per second, and many other methods using different algorithms like SVM , DNN, CNN, DT . we found that all are working quite well but DNN is the best in all them to find the problems in less time with more accuracy.

Many of AI application are widely used in field of medical healthcare to detect diseases and diagnose heart disease patient by using their medical data. In future we intend to enhance the performance of these classification techniques by making different meta models that will used to predict the diseases between people at risk of heart diseases.

Cardiovascular Diseases are a threat that needs to be reduced. Traditional methods of identifying CVD need a lot of human interaction and are expensive hence there is a need for AI-augmented devices which can accurately predict the abnormalities of the heart. Applications for data mining are frequently utilized in the medical field to identify disorders and provide patients with a heart disease diagnosis based on their medical records. The most effective machine learning methods for categorizing cardiovascular illness have been established using patient data, and we have explored these methods in this work. The numerous classification algorithms SVM, KNN, DT, LR, and NB have been compared based on assessment measures such as precision, recall, f1-score, accuracy, and training time.

By developing a meta-model that will be used to forecast cardiovascular illness in those at risk for heart disease, we want to enhance the efficacy of these core cat -egorization algorithms in the future.

Further, we discuss the advancement of AI and ML on the ECG and PCG dev ices in the literature review and discuss the contribution of various researchers in the advancement of automated detection of CVD.

References

1. P. Keikhosrokiani, Perspectives in the development of mobile medical information systems: Life cycle, management, methodological approach and application. 2019. 2. WHO, "cardiovascular diseases (CVDs): key facts," World Health Organization, 2017.
3. P. Keikhosrokiani, N. Mustafa, and N. Zakaria, "Success factors in developing iHeart as a patient-centric healthcare system: A multigroup analysis," Telematics and Informatics, vol. 35, no. 4, 2018, Doi: 10.1016/j.tele.2017.11.006.
4. P. Keikhosrokiani, "Chapter 5 - Success factors of mobile medical information system (mMIS)," P. B. T.-P. in the D. of M. M. I. S. Keikhosrokiani, Ed. Academic Press, 2020, pp. 75–99.
5. P. Keikhosrokiani, N. Mustafa, N. Zakaria, and R. Abdullah, "Assessment of a medical information system: the mediating role of use and user satisfaction on the success of human interaction with the mobile healthcare system (iHeart)," Cognition, Technology & Work, vol. 22, no. 2, pp. 281–305, 2020, Doi: 10.1007/s10111-019-00565-4.
6. A. D'Souza, "Heart disease prediction using data mining techniques," International Journal of Research in Engineering and Science (IJRES) ISSN (Online), pp. 2320–9364, 2015.
7. J. Patel, D. Teja Upadhyay, and S. Patel, "heart disease prediction using machine learning and data mining technique," heart disease, vol. 7, no. 1, pp. 129–137, 2015.
8. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques," IETE Journal of Research, pp. 1–20, 2020.
9. S. S. Tripathy. "System for diagnosing valvular heart disease using heart sounds", Master's Thesis, India, 2005
10. A. S. A. Mohamed and H. M. Raafat. "Recognition of heart sounds and murmurs for cardiac diagnosis". In Proceedings of 9th International Conference on Pattern Recognition, 1988.
11. W. M. Jinjri, P. Keikhosrokiani and N. L. Abdullah, "Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study," *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 132-138, Doi: 10.1109/ICIT52682.2021.9491677.
12. I. Abdel-Motaleb and R. Akula, "Artificial intelligence algorithm for heart disease diagnosis using Phonocardiogram signals," *2012 IEEE International Conference on Electro/Information Technology*, 2012, pp. 1-6, Doi: 10.1109/EIT.2012.6220714.
13. Zaibunnisa L. H. Malik, Momin Fatema, Nikam Pooja, Gawandar Ankita, 2021, Heart Disease Prediction using Artificial Intelligence, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NREST – 2021 (Volume 09 – Issue 04),
14. Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M. Afshar Alam, Safdar Tanweer, Guojun Wang, A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data, Medical Engineering & Physics, Volume 105, 2022, 103825, ISSN 1350-4533,
15. Safial Islam Ayon, Md. Milon Islam & Md. Rahat Hossain (2022) Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques, IETE Journal of Research, 68:4, 2488-2507,
16. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol. 2021 Jul;18(7):465-478. Doi: 10.1038/s41569-020-00503-2. Epub 2021 Feb 1. PMID: 33526938; PMCID: PMC7848866.

17. Siontis, K.C., Noseworthy, P.A., Attia, Z.I. *et al*. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* **18**, 465–478 (2021).
18. Zaibunnisa L. H. Malik, Momin Fatema, Nikam Pooja, Gawandar Ankita, 2021, Heart Disease Prediction using Artificial Intelligence, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NREST – 2021 (Volume 09 – Issue 04),
19. P. Keikhosrokiani, "Chapter 6 - Emotional-persuasive and habit change assessment of mobile medical information Systems (mMIS)," P. B. T.-P. in the D. of M. M. I. S. Keikhosrokiani, Ed. Academic Press, 2020, pp. 101–109
20. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management by Konstantinos C. Siontis, Peter A. Noseworthy, Zachi I. Attia and Paul A. Friedman
21.Yatendra Mohan Sharma, Pawan Kumar Saini, Neelam Sharma "Effective Decision Support Scheme Using Hybrid Supervised Machine Learning Procedure"2nd International Conf. on Information Management and Machine Intelligence, Springer Singapore, pp.-569-575.

Detection of Cardiovascular Disease using A.I & M.L - A Review

Kamlesh Gautam
Assistant Professor, Department of Advance Computing, Poornima College of Engineering, Jaipur

Divyansh Johari, Vidhan Solanki, Hardik Sharma, Aditya Shukla
Student Researchers, Department of Advance Computing, Poornima College of Engineering, Jaipur

Abstract-In the 21st century according to stats, the risk of cardiovascular disease has become more common and the death rate caused by it is increasing way more. around 17.9 million lives are taken by CVDs each year. There are various reasons causing it but most importantly it is caused by not identifying it, hence it becomes a very important task for the human race to identify the CVDs and deal with the proper treatment so that the death dance caused by CVDs can be decreased and risk of it at an early age too. This work mainly aims to review and analyze various methods and approaches to detect the presence or absence of CVDs using AI and ML with accurate predictions. An artificial intelligence system for detecting heart disease from phonocardiogram (PCG) signals has been developed utilizing Artificial Neural Networks (ANN) algorithms and also by driving various A.I algorithm on the given electrocardiogram (ECG) data of the patients we can predict the absence or presence of CVDs.

Keywords— Cardiovascular disease, Artificial Intelligence, machine learning, classification, comparative analysis, PCG, ECG, Deep Learning.

I. INTRODUCTION-heart disease, another name for cardiovascular illness, is a serious global issue that has a big impact on people. According to a recent study, cardiac disorders were responsible for millions of deaths worldwide, or 31% of all fatalities. Medical research has shown that some risk factors increase a person's likelihood of developing heart disease (CVD). According to, some of these factors an unhealthy diet, nicotine use, depression, stress, excessive alcohol use, physical inactivity, inherited obesity, and age are the common causes of CVD. The World Health Organization has published several papers showing an increase in CVD-related deaths, which are primarily attributable to inadequate preventative actions despite rising risk factors.

The heart is one of many organs in the human body that provides blood supply through a function akin to a pump. A healthy heart is fundamental and necessary for human well-being. The leading cause of death in the modern period is cardiovascular disease (CVD), generally known as heart disease. The classification of associated disorders is a challenging undertaking that involves several biological markers and risk factors due to the highly complicated mechanism of the heart. Professionals in related fields employ cardiac physiological signals like the electrocardiogram(ECG) and phonocardiogram (PCG) to monitor or detect cardiovascular-related disorders. Now due to the heart's complex structure and the death rate by the CVDs have grasped the attention of various researchers and scientists to find and perform various approaches, techniques, and methods that are required in order to detect the CVDs presence or absence of a patient. hence some scientists came up with the output of using AI and various Machine-learning techniques

There are a number of publications that proposes different techniques to apply to the patient's data to find the collective research

whether it is detecting Heart health by extracting its features through its sound that is by using a Phonocardiogram and classifying them using a Neural Network. the implementation of Deep-learning methods on Electrocardiogram data in order to predict CVD can also be applied.

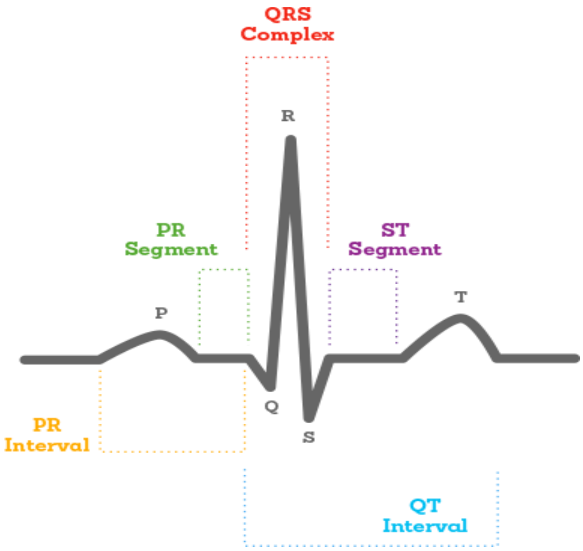
Mohamed and Raafat created a mathematical model in the late 1980s to use a limited set of parameters to characterize heart murmurs and sounds.

For the classification of cardiovascular disease, five alternative methods have been presented: support vector machine (SVM), K-nearest neighbor (K-NN), logistic regression (LR), decision tree (DT), and naive Bayes (NB). These methods were used to classify the various patient data in order to determine the presence of CVDs.

2. UNDERSTANDINGS & WORKINGS

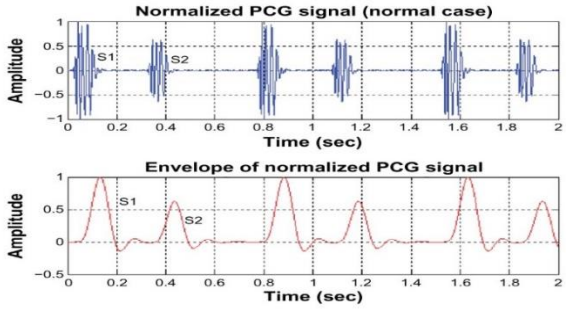
So, now to simplify the approach let us first understand more about ECGs and PCG and grasp the main understanding of the Conduction System of heart

[I] Mainly ECG is a device that records the electrical signals produced by heart mainly by SA (Sino Atrial) Node {The Pacemaker} and AV(Atrio-ventricular) Node where SA node causes the upper heart chambers to contract while AV node to cause the Lower heart contract The ECG is made up of multiple electrodes that are wired to the instrument and to the patient's body. To detect impulses, each sensor detects a change in the electrical charge beneath the skin. An impulse travel quickly and is transmitted to the heart's surrounding cells. Figure 1 showed the electrocardiogram's typical waveform. So, by decoding this impulse we can find the abnormal behavior of the heart if it exists and then we can run our algorithm on it in order to predict the probability of CVDs risk.



According to the diagram you can easily spot three main areas of the ECG graph where the first one is knowns as the P wave depicting atrial depolarization i.e., contraction of atria, then comes the rotated V-shaped wave named QRS complex that depicts ventricular depolarization and atrial repolarization following it is T wave depicting ventricular repolarization i.e., the relaxation process of ventricles. Normally a healthy heart has about 60-100 BPM caused by the SA node and by 40-60 BPM AV node so we check the patient's heart condition by interpreting this ECG graph by calculating the P waves, PR interval, QRS complex, hearts rhythm, heart rate (ECG should be 6-second strip) by applying 6-second method.

[II] Now for the same purpose PCG as referred to as phonocardiography where this graph gives us the sound wave of the cardiac cycle which is the contraction and relaxation of the heart that causes the murmurs sound which layman often refers to as the Lub-Dub sound of the heart.

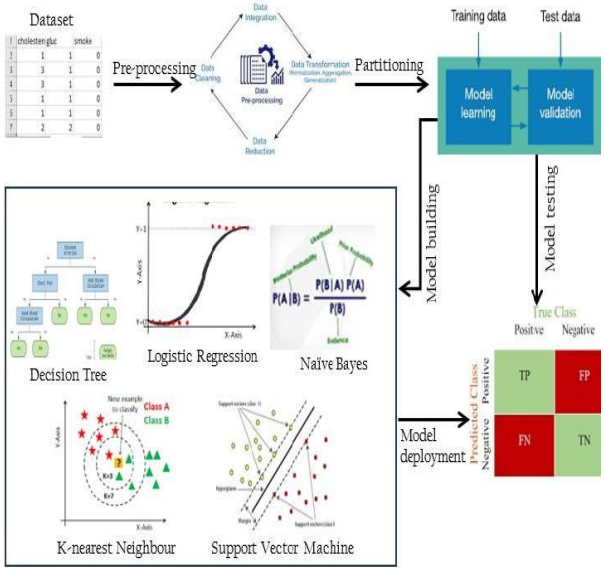


These recorded sound tracks provide information regarding valve action and the efficiency of blood pumping under the body, assisting medical experts. Typically, the PCG approach divides the normal heartbeat into the S1 and S2 beats. We can determine the corresponding heart illnesses based on these recorded soundtracks. S1 sound depicts the closure of valves in the upper and lower chambers of the heart usually the frequency remains between 30-100 Hz range for a normal healthy heart while the S2 sound depicts the closure of the valves which has a frequency above 100Hz. So, this is the basic related field solution for finding the abnormalities in the heart in order to predict the probabilities of the risk of CVDs.

3.CVDs classification using machine learning algorithms (METHODOLOGY)- The dataset for categorization is first collected as part of the approach for this study. One of the key jobs in data mining is classification, whose goal is to group documents into one or more classes or categories as a result, this work creates a useful technique for the dataset used to categorize cardiovascular disease. In order to achieve this, it is necessary to make the assumption that there are only two classes: the positive class with unknown results and the negative class without surprising findings.

Anaconda Jupyter Notebook, a Python 3 application, is used to implement the algorithms. The datasets are split into training and test sets after pre-processing. The majority of researchers opt for a 70:30 split (70 percent for training and 30 percent for testing), as more training data results in more optimal and accurate outcomes. As a consequence, the 70:30 partitioning ratio is employed.

3.1. Dataset- The dataset used to evaluate and contrast the methods used in this work was obtained from the Kaggle web repository. The dataset, which consists of 77,000 patient clinical trial records gathered



by hospitals for cardiovascular illnesses, has three input components: examination (outcomes of medical investigation), subjective (realistic information), and objective (realistic information) (data obtained from a patient). Eleven attributes total, including one target variable with the label "(Absence or Presence) for diagnosis," four objective features, four examination features, three subjective features, and four other attributes, make up the dataset. Table 1 provides summary of the CVDs dataset collected for the study.

Table 1 Description of the Dataset

3	Attributes	Input Features	Data Type/ Description
1	Age		Int / days
2	Height		Int / centimeters

3	Weight	Objective features	Float/ kilograms
4	Gender		Categorical code 1: male, 2: female
5	Systolic blood pressure	Examination features	Int/
6	Diastolic blood pressure		Int/
7	Cholesterol		1: normal, 2: above normal, 3: well above normal
8	Glucose		1: normal, 2: above normal, 3: well above normal
9	Smoking	Subjective features	Binary
10	Alcohol		Binary
11	Physical activity		Binary
12	Cardiovascular	Target	Presence or absence of CVD / target variable.

3.2.Classification Methods-

A description of the machine learning methods used in this research follows Five well-known classification models—Decision Tree, K-Nearest Neighbor, Logistic Regression, Naive Bayes, and Support Vector Machine—have been developed, and their prediction accuracy has been compared.. Numerous research contrasted data mining techniques using various parameter settings. The majority of these earlier research concluded that these techniques outperformed their statistical equivalents because they were less restricted by presumptions and produced better categorization outcomes. A few of these techniques are covered in brief.

3.2.1.Linear Regression - Logistic regression (LR). is one of the most often used machine learning methods for analyzing multivariate regression issues in the medical industry**.** Using a continuous independent variable that aids in both the diagnosis and prediction of illnesses, LR is used to anticipate the outcome of a dependent variable. It is a method for discriminating between categories that use the input vector to extract important statistical data points from the model or forecast data trends. In the LR, the dependent variable is a binary variable that only accepts data that is coded as 0 (yes, success, etc.) or 1. (no, failure, etc.). Calculating the log chances of an event is the basic goal of an LR analysis. As shown mathematically, LR calculates multiple linear regression functions as follows:

$$log\frac{p(y=1)}{1-(p=1)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k;$$

3.2.2.Support Vector Machine: SVM was first introduced in 1992 by Boser, Guyon, and Vapnik as a binary classification method. Since then, a number of classification and regression problems have been added to its list of uses. Without adding prior knowledge, its generalization performance makes it a good classifier even with a lot of data. By choosing the best hyperplane that maximises the margin between classes, data with separable classes can be categorised. SVM is modelled as a collection of vector spaces with limited dimensions, where each dimension corresponds to a unique property of an item. Issues in high-dimensional space have been successfully addressed using this method. Due to its processing efficiency on huge datasets, SVM has demonstrated great performance for illness prediction in the medical field in recent years. The main goal is to reduce generalization mistakes and develop it as a supervised learning system for regression and classification applications.

The SVM is mathematically expressed as:

$$\text{If } Y_i = +1; wx_i + b \geq 1 \qquad (2)$$

$$\text{If } Y_i = -1; wx_i + b \leq 1 \qquad (3)$$

$$\text{For all i; } y_i(w_i + b) \geq 1 \qquad (4)$$

The above equations shows "w" which means weight and "x" as a vector point. Therefore, the data in (3) and (4) must be constantly larger than zero and below zero, respectively, in order to distinguish the data in (2). SVM selects the hyperplane with the greatest distance between it and all other

potential hyperplanes.

3.2.3. The K-NearestNeighbour Method - The (K-NN) technique classifies cases based on how closely they resemble one another. When a case is new at a certain location, its distance from each model case—which is calculated as the nearest neighbor and is the most comparable to the approach that suggests the case—indicates the case. In this manner, the case is added to the output of neighbors that are closest to it. A new input class label is predicted by the K-NN algorithm, which bases its prediction on how similar the new input is to samples of its input from the training set. The K-NN classification output is poor if the new input is identical to the training set's samples.

The distance between two points x(xo, x′) is calculated using the following mathematical formula, where p:x*x=R is a function that gives distance.

$$p(x, x^i) = \left|x - x^i\right| = \sqrt{\sum_{i=1}^{d}(x_i - x_i')^2}$$

3.2.4.Decision Tree : A supervised learning strategy called a DT is applied mostly to classification-related problems by evaluating the entropy and gain of information, DT classifies the data using decision rules derived from the training data. A classification tree structure where each node represents a property The root node will be the primary node, followed by the child nodes. The leaf nodes then represent the decision's result. It works well with both categorical and continuous characteristics. With DT, the population is split into two or more groups based on important predictors. The entropy for each feature is computed as the first step in DT. The dataset was then divided depending on the variables and predictors, either with high data gain or low entropy. The two phases are followed by the remaining qualities, as mentioned.

$$\text{Entrophy}(E) = \sum_{k=1}^{l} -_{qk} \log_{2qk}$$

"qk" is the proportion of the count of the kth class procedures to the total count of models, whereas "l" referred to a response variable module count.

3.3. Evaluation Method - Specific measures, including as the f1 score, precision, recall, and accuracy, are used as the foundation. We also keep track of how long each algorithm takes to train. The classification algorithm's output is displayed in the confusion matrix, and this information forms the basis for further parameter calculations. As a result, the confusion metrics may examine a model's correctness and determine whether a classification algorithm frequently labels items with the incorrect labels by comparing anticipated values with actual values. The confusion matrix's values and its depiction in Table 2 are quickly described by the parameters below.

True positive (TP) circumstances occur when the datapoint's actual class and projected class are both 1.

- False positive (FP) scenarios take place when a data point's real class is 0, despite the fact that the predicted class is 0.
- False negative (FN) scenarios occur when a data point's real class is 0 and its anticipated class is 1.
- True negative (TN) circumstances occur when a data point's real class is 1 and its anticipated class is 0, respectively.

Table II- Confusion Matrix

		Actual value	
		Classified as absence	***Classified as presence***
	Absence	TP	FN
	Presence	FP	TN

3.3.1.F1 score

When an f1-score achieves its highest value at 1 and its poorest score at 0, it is a function that is understood as a weight of recall average and precision. The f1-score formula is as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.3.2.Recall

Recall measures how much pertinent data is recovered from any machine learning system. The capacity to locate all connected events in the data is the main focus. The recall is represented by the equation below:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

3.3.3.Precision

Being precise means being exact and correct. Precision conveys the sense of incidents that were accurately expected. It measures the proportion of genuine positives among all positives and quantifies forecasts that fall into the positive category as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

3.3.4.Accuracy

An important metric for describing an algorithm's performance is accuracy. It establishes the threshold at which an algorithm can accurately forecast both positive and negative instances and is quantified by the following formula:

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

3.4.Result of the algorithm –

Five artificial intelligence techniques were used in this work to classify the cardiovascular disease dataset: decision trees (DT), logistic regressions (LR), K-nearest neighbors (KNN), and support vector machines (SVM). The dataset consists of 70,000 samples and 11 attributes. The classes can be examined both with and without the disease.

35021 of the data samples are marked as missing, whereas 334979 are marked as suffering from cardiovascular illness. Training and testing sets are split up into the data in the ratio of 70:30.

The confusion matrix of the DT, LR, KNN, and SVM prediction results is shown in Tables 1 through 5.

		Predicted Value		
		Absence	Presence	Actual value
Actual value	Absence	6562 (31.35%)	3899 (18.46%)	10461
	Presence	3775 (18.03%)	6764 (32.16%)	10539
	Total Predicted	10337	10663	21000

Table 3 – Confusion Matrix for Decision Tree

		Predicted Value		
		Absence	Presence	Actual value
	Absence	8243 (39.25%)	2296 (10.93%)	10539
			6430 (30.62%)	10461
			8726	21000

Table 4 – Confusion Matrix for KNN

		Predicted Value		
		Absence	Presence	Actual value
Actual value	Absence	5363 (38.31%)	1625 (11.61%)	6988
	Presence	2244 (16.03%)	4768 (34.06%)	7012
	Total Predicted	7607	6393	14000

Table 5 – Confusion Matrix for Logistic Regression

		Predicted Value		
				Actual value
Actual value	Absence	5671 (40.51%)	1317 (9.41%)	6988
	Presence	2509 (17.92%)	4503 (32.16%)	7012
	Total Predicted	8180	5820	14000

Table 6 – Confusion Matrix for SVM

Hence the best-performing algorithm according to the Confusion Matrix, F1 scores and precision is examined and selected for identifying the CVD in an individual.

4.Deep Learning Methods applied to ECG

The development of thorough, human-like interpretation abilities is one of the main goals of using AI for ECG interpretation. Since the creation of the digital ECG more than 60 years ago, attempts have been made to provide a quick, accurate, and thorough computer-generated interpretation of the ECG. The problem seems to be solved since ECG interpretation is a fairly constrained

application of pattern recognition to a small dataset. Early digital ECG interpretation software was capable of quickly identifying fiducial sites, taking precise measurements, and defining common measurable anomalies. Modern technology has advanced past these rule-based methods to find patterns in vast amounts of labeled ECG data.

AI-driven algorithms have been developed by a number of organizations, and some of these algorithms are already in use in a few limited clinical scenarios. Research has employed a large number of single-lead ECG datasets to build CNNs, which have then been applied to 12-lead ECGs. One team, for example, employed a CNN to categories 2 million tagged single-lead ECG recordings from the Clinical Outcomes in Digital Electro cardiology study into six different categories of anomalies on the 12-lead ECG. The technique was shown to be viable in this work, but its wider application or external validation in other 12-lead ECG datasets is still to come. Another team discovered that CNN's could make some diagnosis more reliably than practising cardiologists in comparable studies utilising single-lead ECGs. It remains to be seen, nevertheless, if this method will result in 12-lead ECG interpretation software that is therapeutically relevant. A CNN was created for the multilabel identification of 21 different heart rhythms based on the 12-lead ECG in an assessment that was published in 2020 utilizing a training and validation dataset of >80,000 ECGs from >70,000 patients. The cardiologists' committee's consensus labels served as the reference standard. The best network performed much better than a single cardiologist interpretation in a test dataset of 828 ECGs, matching the gold standard labels in 80% of the ECGs. The model's sensitivity, specificity, and mean area under the curve (AUC) receiver operating characteristic scores were 98%, 87%, and 99%, respectively.

The team of researchers developed a thorough ECG-interpretation infrastructure using their own dataset of >8 million ECGs done for clinical purposes (all of which have been labeled by experienced ECG readers and are connected to the relevant electronic health record). They showed that a CNN has good diagnostic performance and can recognize 66 distinct codes or diagnosis labels. Recently, Researchers created a unique technique that translates ECG characteristics into ECG codes and text strings using a transformer network and a CNN to extract ECG features. By providing information in a same manner and using comparable language, this approach produces a model output that is more like that of a human ECG reader. It also makes sense of related codes, preventing the display of opposing or mutually conflicting interpretations that a human reader would not present. This approach will be especially important as our reliance on ECG data collected by cutting-edge, consumer-facing apps that are greatly scalable increases. For instance, single-lead ECG traces collected from mobile, smartwatch-enabled recordings have been subjected to AI-ECG algorithms for the diagnosis of AF. The democratization of ECG technology will cause the volume of signals that need to be interpreted to increase quickly, maybe faster than the rate at which human ECG readers can handle them. These autonomous, consumer- or patient-facing models are projected to be essential for telehealth technology. They could also make it possible to build essential lab spaces with the capacity to store and process massive quantities of data.

The signal quality produced with these devices can vary, as seen in the wristwatch research cited above, and AI-ECG may be less able than human expert over-readers to

classify the heart rhythm utilizing inferior tracings. Similar to this, another study discovered that a deep neural network built using ECG recordings from smartwatches performed well for passive detection of atrial fibrillation (AF) in comparison to AF diagnosed from 12-lead ECGs, but that performance was noticeably less reliable when referencing a self-reported history of persistent AF.

Nevertheless, despite significant advancements achieved, a full, human-like ECG interpretation package is still a long way off. Even in its most advanced form, the package doesn't have the precision required for execution without human supervision. The interpretation of an ECG generated by a machine also has the potential to affect human over readers and, if erroneous, can be a cause of bias or systemic mistake. This worry is especially important if the algorithms were developed in populations that are different from the populations where they are used.

This weakness highlights the necessity for a diversified derivation sample, demanding external validation studies, gradual implementation, and continuous model performance and effectiveness evaluations (ideally including a diverse patient population and diverse means of data collection that reflect real-world practices).

5.Deep Learning Methods applied to Phonocardiogram

The first fundamental analytic technique used to assess the heart's functioning status is cardiac auscultation. If the heart sound from the Phonocardiogram (PCG) test indicates any abnormalities, an electrocardiography (ECG) test is required. The ability to hear and see the heart auscultation on the screen increases confidence in the precision of the first diagnosis. However, the diagnosis could not be as precise as expected due to noise and human error. If an artificial intelligence computer is utilized to generate potential diagnoses utilizing some distinguishing characteristics of the heart sound waves, diagnosis accuracy can be significantly improved. This method should lower death rates and healthcare expenses.

Numerous articles have put forth various methods for deriving characteristics from heart sounds and categorizing them using neural networks. Mohamed and Raafat created a mathematical model in the late 80s to use a limited set of parameters to characterize cardiac murmurs and noises. The smallest difference between the characteristics of the observed pattern and the reference patterns was used to classify the data in this case. Features were extracted using a fourth-order linear prediction of the cardiac cycle frames.

Patil and Kumaraswamy proposed an intelligent heart attack prediction system based on data mining and artificial neural networks. By applying the K-means clustering algorithm to the supplied data, this technique computes the parameters crucial to the heart attack. The Maximal Frequent Itemset Algorithm is used to extract these common patterns from the data (MAFIA). Following that, the designs are chosen based on the calculated significant weightage. Although the aforementioned study claimed that this technology may detect heart attacks using the MAFIA algorithm, the prediction accuracy for the work was not stated. Additionally, rather than using feature characteristics of the heart sound signal, this method makes use of features that correspond to the subject's behavioral patterns, such as drinking and smoking.

The GAL (Grow and Learn) algorithm is a revolutionary technique for segmenting heart sounds utilizing homomorphic filtering and feature extraction from wavelet coefficients. This method's accuracy was estimated to be 90.9%. In their study on the

analysis of heart sounds for symptom identification, Reed et al. used wavelet decomposition to segment and alter the heart sounds. By eliminating levels with the shortest scales, the altered vectors were condensed into lower vector sizes. A three-layer neural network was used to classify each vector, and it provided 100% accuracy for all heart sounds. The drawback of this method is the requirement for using several hidden layer neurons—up to 50 layers.

Arrhythmia categorization based on heart rate variability (HRV) has been documented. This strategy is built on the General Discriminant Analysis (GDA) and Multi Layer Perceptron (MLP) techniques. The outcomes showed that this approach produced 100% accuracy for the data the authors obtained from the MIT-BIH database. But instead of PCG signals, this approach employs HRV signals based on the ECG. It should be highlighted that getting an ECG signal does not qualify as a regular test for primary care doctors since it necessitates laboratory setups, which takes time and is less efficient financially.

6.Conclusion – Cardiovascular Diseases are a threat that needs to be reduced. Traditional methods of identifying CVD need a lot of human interaction and are expensive hence there is a need for AI-augmented devices which can accurately predict the abnormalities of the heart. Applications for data mining are frequently utilised in the medical field to identify disorders and provide patients with a heart disease diagnosis based on their medical records. The most effective machine learning methods for categorizing cardiovascular illness have been established using patient data, and we have explored these methods in this work. The numerous classification algorithms SVM, KNN, DT, LR, and NB have been compared based on assessment measures such as precision, recall, f1-score, accuracy, and training time.

By developing a meta-model that will be used to forecast cardiovascular illness in those at risk for heart disease, we want to enhance the efficacy of these core categorization algorithms in the future.

Further, we discuss the advancement of AI and ML on the ECG and PCG devices in the literature review and discuss the contribution of various researchers in the advancement of automated detection of CVD.

7. References –

1.P. Keikhosrokiani, Perspectives in the development of mobile medical information systems: Life cycle, management, methodological approach and application. 2019. 2. WHO, “Cardiovascular diseases (CVDs): key facts,” World Health Organization, 2017.

3. P. Keikhosrokiani, N. Mustafa, and N. Zakaria, “Success factors in developing iHeart as a patient-centric healthcare system: A multigroup analysis,” Telematics and Informatics, vol. 35, no. 4, 2018, Doi: 10.1016/j.tele.2017.11.006.

4. P. Keikhosrokiani, “Chapter 5 - Success factors of mobile medical information system (mMIS),” P. B. T.-P. in the D. of M. M. I. S. Keikhosrokiani, Ed. Academic Press, 2020, pp. 75–99.

5. P. Keikhosrokiani, N. Mustafa, N. Zakaria, and R. Abdullah, “Assessment of a medical information system: the mediating role of use and user satisfaction on the success of human interaction with the mobile healthcare system (iHeart),” Cognition, Technology & Work, vol. 22, no. 2, pp. 281–305, 2020, doi: 10.1007/s10111-019-00565-4.

6. A. D’Souza, “Heart disease prediction using data mining techniques,” International Journal of Research in Engineering and Science (IJRES) ISSN (Online), pp. 2320–9364, 2015.

7. J. Patel, D. TejalUpadhyay, and S. Patel,

"Heart disease prediction using machine learning and data mining technique," Heart Disease, vol. 7, no. 1, pp. 129–137, 2015.

8. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques," IETE Journal of Research, pp. 1–20, 2020.

9. S. S. Tripathy. "System for diagnosing valvular heart disease using heart sounds", Master's Thesis, India, 2005

10. A. S. A. Mohamed and H. M. Raafat. "Recognition of heart sounds and murmurs for cardiac diagnosis". In Proceedings of 9th International Conference on Pattern Recognition, 1988.

11. W. M. Jinjri, P. Keikhosrokiani and N. L. Abdullah, "Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study," *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 132-138, doi: 10.1109/ICIT52682.2021.9491677.

12. I. Abdel-Motaleb and R. Akula, "Artificial intelligence algorithm for heart disease diagnosis using Phonocardiogram signals," *2012 IEEE International Conference on Electro/Information Technology*, 2012, pp. 1-6, doi: 10.1109/EIT.2012.6220714.

13. Zaibunnisa L. H. Malik, Momin Fatema, Nikam Pooja, Gawandar Ankita, 2021, Heart Disease Prediction using Artificial Intelligence, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NREST – 2021 (Volume 09 – Issue 04),

14. Javed Azmi, Muhammad Arif, Md Tabrez Nafis, M. Afshar Alam, Safdar Tanweer, Guojun Wang, A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data, Medical Engineering & Physics, Volume 105, 2022, 103825, ISSN 1350-4533,

15. Safial Islam Ayon, Md. Milon Islam & Md. Rahat Hossain (2022) Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques, IETE Journal of Research, 68:4, 2488-2507,

16. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol. 2021 Jul;18(7):465-478. doi: 10.1038/s41569-020-00503-2. Epub 2021 Feb 1. PMID: 33526938; PMCID: PMC7848866.

17. Siontis, K.C., Noseworthy, P.A., Attia, Z.I. *et al.* Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* **18**, 465–478 (2021).

18. Zaibunnisa L. H. Malik, Momin Fatema, Nikam Pooja, Gawandar Ankita, 2021, Heart Disease Prediction using Artificial Intelligence, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NREST – 2021 (Volume 09 – Issue 04),

19. P. Keikhosrokiani, "Chapter 6 - Emotional-persuasive and habitchange assessment of mobile medical information Systems (mMIS)," P. B. T.-P. in the D. of M. M. I. S. Keikhosrokiani, Ed. Academic Press, 2020, pp. 101–109

20. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management by Konstantinos C. Siontis, Peter A. Noseworthy, Zachi I. Attia and Paul A. Friedman

21.Yatendra Mohan Sharma, Pawan Kumar Saini, Neelam Sharma "Effective Decision Support Scheme Using Hybrid Supervised Machine Learning Procedure"2nd International Conf. on Information Management and Machine Intelligence, Springer Singapore, pp.-569-575.