[B.Sc. Engg. Thesis]

Prediction of Arrhythmia after Acute Myocardial Infarction Using Machine Learning and Statistical Techniques

Salauddin Tapu

Ummay Umama Gronthy

Electronics and Communication Engineering Discipline

Science, Engineering and Technology School

Khulna University, Khulna 9208,

Bangladesh

February 2023

Prediction of Arrhythmia after Acute Myocardial Infarction Using Machine Learning and Statistical Techniques

This thesis is submitted to the Electronics and Communication Engineering Discipline in partial fulfillment of the requirements for the degree of Bachelor of Science in Electronics and Communication Engineering, abbreviated as, B.Sc. Engg. (ECE).

By

Salauddin Tapu

Student ID: 180923

Ummay Umama Gronthy

Student ID: 180939

Electronics and Communication Engineering Discipline

Science, Engineering and Technology School

Khulna University, Khulna 9208,

Bangladesh

February 2023

Recommendation

This thesis is reported and presented as a requirement for the degree of Bachelor of Science in Electronics and Communication Engineering, abbreviated as B.Sc. Engg. (ECE), awarded by Electronics and Communication Engineering (ECE) Discipline, Khulna University. Authors declare that the work solely performed by them.

Approved By

Dr. Uzzal Biswas **(Supervisor)**
Associate Professor
Electronics and Communication Engineering Discipline
Khulna University, Khulna-9208, Bangladesh

Dr. Abdullah-Al Nahid **(External Member)**
Professor
Electronics and Communication Engineering Discipline
Khulna University, Khulna-9208, Bangladesh

Dr. Md. Abdul Alim **(Member & Head)**
Professor
Electronics and Communication Engineering Discipline
Khulna University, Khulna-9208, Bangladesh

Declaration by Authors

We, the undersigned, Salauddin Tapu and Ummay Umama Gronthy hereby declare that we are the sole authors of this thesis titled **"Prediction of Arrhythmia after Acute Myocardial Infarction Using Machine Learning and Statistical Techniques"** under the sincere guidance of our supervisor **Dr. Uzzal Biswas**. To the best of our knowledge this thesis contains no material previously published by any other person except where due references has been made. This thesis contains no material which has been accepted or published as part of the requirements of any other academic degree or non-degree program, in English or in any other language. This is a true copy of the thesis, including final revisions. If our work is called into doubt due to unethical methods, we shall accept full responsibility.

Salauddin Tapu
Student ID: 180923
Electronics and Communication Engineering Discipline
Khulna University, Khulna

Ummay Umama Gronthy
Student ID: 180939
Electronics and Communication Engineering Discipline
Khulna University, Khulna

Acknowledgement

First and foremost, we would like to express our profound gratitude to the Almighty for allowing us to complete our thesis.

Then, we would like to express our deepest gratitude to Dr. Uzzal Biswas, Associate Professor of Electronics and Communication Engineering Discipline, Khulna University, for his significant assistance in our undergraduate thesis. With that, we would like to convey our heartfelt appreciation for his constant support, patience, inspiration, passion, and deep expertise. His door was always open anytime we had a problem or a query concerning our thesis. He always directed us on the correct path when he believed we needed it.

We could not have undertaken this journey without our respected external member, Dr. Abdullah-Al Nahid, Professor of Electronics and Communication Engineering Discipline, Khulna University, Khulna. This thesis would not have been completed without his support and committed engagement in every step of the process.

We are grateful and fortunate to have consistent encouragement and assistance from the faculty members of the Electronics and Communication Engineering Discipline, Khulna University, Khulna, which assisted us in finishing our thesis work. Furthermore, we would like to express our heartfelt gratitude to all laboratory personnel for their prompt assistance.

We would like to thank our parents; whose love and guidance are with us in whatever we pursue. They are our ultimate role models. We would also like to express our gratitude to our loved ones and friends who have always motivated us.

Abstract

Arrhythmia is the condition when our heart beats irregularly. This life-threatening condition can lead to stroke, heart failure, and even death in critical cases. Sometimes it also causes unnecessary hospitalization. Every year millions of people die due to this disease. According to our existing health care system various clinical tests are needed to detect arrhythmia. This clinical diagnosis process is very time consuming and expensive. Patient's condition gets worsen due to this long diagnosis process. But early detection and on time prediction of arrhythmia can reduce the risk of morbidity, hospitalization and mortality. Many state-of-the-art technologies are now being used to predict arrhythmias, including statistical analysis, machine learning (ML) based approaches, and deep learning (DL) based approaches. So, in this thesis we have introduced efficient approaches for predicting arrhythmia. We have studied on acute myocardial infarction (AMI) dataset and used two different approaches (statistical analysis and ML) for predicting arrhythmia within a reasonable time. We have used both statistical analysis and ML to find out the influential features responsible for causing arrhythmia. In this work Chi-square test, gamma coefficient and crosstab analysis have been used as statistical techniques. We have used Chi-square test to determine the association between arrhythmia and other features. Gamma coefficient has been used to find their association strength and finally, we have done the crosstab analysis to analyze the categorical relationship between arrhythmia and other features. From statistical analysis, we have found RA (up and down), BBB, P-R and RA (up and down) as influential features, which are responsible for the occurrence of arrhythmia. In the ML analysis approach, we have used an optimized random forest (RF) classifier to predict arrhythmia. Along with the statistical analysis, we have also used a meta-heuristic algorithm, CSA to extract features. The model has shown an accuracy of 69.71% with all the features and 68.89% with extracted features. We have extracted BBB, RA (right and left), heart beats and DD-P as important predictors for arrhythmia. In this work, we have also analyzed the time latency of the model. In addition, we have compared our work with state-of-the-art works and validated the significance of the extracted features for predicting arrhythmia.

Contents

List of Figures

List of Tables

List of Abbreviations

AF	Atrial Fibrillation
AI	Artificial Intelligence
AMI	Acute Myocardial Infarction
ANN	Artificial Neural Network
AUC	Area Under Curve
BBB	Bundle Branch Block
CART	Classification and Regression Tree
CK-MB	Creatine Kinase Isoenzyme
Cr	Creatinine
CRP	C-reactive Protein
CSA	Cuckoo Search Algorithm
CV	Cross Validation
CWT	Continuous Wavelet Transform
DBP	Diastolic Blood Pressure
DCT	Discrete Cosine Transform
DD-P	D Dimer
DETR	Detection Transformer
DL	Deep Learning
DM	Data Mining
DOST	Discrete Orthogonal Stockwell Transform
dt	E Deceleration Time
DT	Decision Tree
DWT	Discrete Wavelet Transform
E/A	Mitral Valve Peak Velocity Early Diastolic Filling (E Wave) to Peak Velocity of Late Diastolic Filling (A Wave) Ratio
ECC	Error-correction Code
ECG	Electrocardiogram
FN	False Negative
FP	False Positive
FS	Feature Selection
GA	Genetic Algorithms
GR	Gain Ratio
HDL	High-density Lipoprotein
HF	Heart Failure
HILB	Hilbert Transform
HPO	Hyper Parameter Optimization
HRV	Heart Rate Variability
ICA	Independent Component Analysis
IOT	Internet of Things

IVST	Interventricular Septum Thickness
KL	Kullback-Leibler
KNN	k-Nearest Neighbor
LA	Left Atrium Diameter
LDL	Low-density Lipoprotein
LR	Logistic Regression
LSTM	Long Short-term Memory
LVEDD	Left Ventricular End-diastolic Diameter
LVEF	Left Ventricular Ejection Fraction
LVPWT	Left Ventricular Posterior Wall Thickness
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naïve Bayes
NN	Neural Network
NYHA	New York Heart Association
OAA	One Against-all
OAO	One Against-one
PA	Pulmonary Artery
PCA	Principal Component Analysis
PCI	Percutaneous Coronary Intervention
P-R	PR Interval
Prior CHD	Prior Coronary Heart Disease
Prior CI	Prior Cerebral Infarction
Prior HF	Prior Heart Failure
Prior MI	Prior Myocardial Infarction
Pro-BNP	Pro-B-type Natriuretic Peptide
Q-Tc	QTc Interval
RA	Right Atrium
RA (right and left)	Right Atrium Right and Left Diameter
RA (up and down)	Right Atrium Up and Down Diameter
RF	Random Forest
ROC	Receiver Operating Characteristic
SA Node	Sinoatrial Node
SBP	Systolic blood pressure
SpEn	Sample Entropy
SU	Symmetric Uncertainty
SVM	Support Vector Machine
TC	Total Cholesterol
TCSC	Threshold Crossing Sample Count
TG	Triglyceride

TN	True Negative
TNI	Troponin I
TP	True Positive
UCI	University of California Irvine
UGLU	Urine Glucose
Vao	Peak Aortic Velocity
VF	Ventricular Fibrillation
Vpa	Pulmonary Peak Flow Rate
VT	Ventricular Tachycardia
VWM	Ventricular Wall Motion
VWMA	Ventricular Wall Motion Abnormal
WoS	Web of Science

Chapter 1
Introduction

1.1 Background

Heart is a vital organ of human body and over the lifetime of a person, the heart beats 2.5 billion times [1]. The heart usually beats at a rate of about 75 beats per minute [2]. Any interruptions to this normal heartbeat are caused by different kinds of heart diseases. Among various heart diseases, arrhythmia is one of the most critical conditions. Arrhythmia is a term that refers to any abnormal heart rhythm, which is not physiologically justified [3]. The Sinoatrial (SA) node generates electrical impulses that regulate the heart's rhythm. When these electrical signals aren't functioning properly, arrhythmia may occur. This can lead to various life-threatening conditions such as stroke, heart failure, and even death. Arrhythmia is one of the leading causes of death all over the world, which claims (15–20) % of all deaths [4].

There are different types of arrhythmias such as ventricular fibrillation (VF), atrial fibrillation (AF), and ventricular tachycardia (VT). AF is the most common form of arrhythmia. In 2010, approximately 33.5 million people was affected by AF worldwide [5]. According to 2017 database, 3.046 million new cases of AF were recorded globally [6]. Among the high-income countries like USA, around (2-6) million people are currently affected by AF and estimated to be doubled within 2060 [7]. According to recent statistics in the USA, AF causes more than 454,000 hospitalizations each year [8]. AF is the reason of about 26,535 deaths per year in USA [9]. People from underdeveloped and developing countries, like Bangladesh, are affected more than the people of developed countries, where 80% of all deaths occur due to this heart disease [10]. In addition to death and hospitalization, a variety of emotional anguish or behavioral issues may be experienced by the arrhythmia patient. Mental health problems are particularly common among them [11]. These diseases have an impact not only on patient’s personal life but also on their whole family and society. It is also a major cause of financial burden for the victims and their family members as approximately 45% patient with arrhythmia and other cardiovascular disease experience financial hardship due to medical expenses [12]. From 2011 to 2025, it is predicted that all noncommunicable diseases will cause $7.28 trillion cumulative

economic losses in low- and middle-income countries and nearly half of the estimated loss will be due to this disease [13]. Noncommunicable cardiac arrhythmias and related mortality have significantly increased in Bangladesh during the past few decades [14].

People of any age might be affected by this life-threatening disease. Therefore, early detection and prediction of arrhythmia is very crucial for reducing the risk of morbidity, hospitalization, and mortality. Many researchers have used different approaches to predict arrhythmias, which are discussed briefly in the following subsections. Machine learning (ML) can play a significant role in prediction. But compare to all feature, using only the influential features we can easily predict arrhythmia. Now a days in order to predict arrhythmia accurately, many researchers preferred a variety of feature selection techniques and classification algorithms. One of the earliest studies on arrhythmia classification is [15], which was led by L. Qiao, R. Cadathur, and D. C. Gari in 2013. They classified VF and tachycardia using ML based support vector machine (SVM) algorithm. For their research purpose they have used three famous ECG databases (the American Heart Association Database, the Creighton University Ventricular Tachyarrhythmia Database, and the MIT-BIH Malignant Ventricular Arrhythmia Database) and extracted 14 metrics from an ECG signal. They got accuracy of 98.1% from in-sample training data and 96.3% from out-sample training data.

In 2014 F. Alonso-Atienza et al. [16] and his team detected VF and shockable arrhythmia based on Creighton University ventricular tachycardia database that contains 13 ECG parameters. They used SVM classifier and feature selection method to identify shockable versus non-shockable and VF versus non-VF arrhythmias. They determined Hilbert transform (HILB) as the most important feature and threshold crossing sample count (TCSC), sample entropy (SpEn), and VF filter (VFleak) are the important features for detection. They achieved 99.7% detection accuracy for arrhythmia.

In order to detect cardiac arrhythmia B. Amina et al. [17] introduced a novel set of classifier in 2015. They performed their research on University of California Irvine (UCI) Cardiac Arrhythmia dataset which has 279 attributes. They applied 25 classifier and got highest accuracy from decision tree (DT) classifier. DT classifier returned 87.21% of classification accuracy.

In the year of 2016, V. Kalidas and L. S. Tamil [18] classified five distinct arrhythmias including extreme bradycardia, extreme tachycardia, asystole, VT and VF using the SVM technique, which is based on ML. They applied logical analysis methods on PhysioNet/Computing in Cardiology 2015 Challenge dataset to carried out their work. Based on real-time dataset and retrospective dataset they conducted their analysis. For both of the dataset they attained 94% sensitivity. In case of specificity 82% obtained from real time dataset and 86% from retrospective dataset.

For the classification of arrhythmia K. Yasin et al. [19] introduced an effective ML algorithm in 2017. Based on MIT-BIH arrhythmia database they evaluated their model using feature extraction, dimension reduction and classification techniques. Two different feature selection method (statistical and temporal features), three different feature deduction methods named Genetic Algorithms (GAs), Independent Component Analysis (ICA) and Principal Component Analysis (PCA) were used in this research. Finally, they applied DT, SVM, neural network (NN) and k-nearest neighbor (KNN) to classify arrhythmia. Among these four classifiers, the maximum classification accuracy 99.30% was attained by the KNN classifier using GA.

In 2018, M. Anam et al. [20] researched on Irvine Machine Learning Data Repository to classify arrhythmia into 16 subclasses. Among those 16 classes, 15 classes indicated the presence of different types of arrhythmias and the remained one indicated the absence of this disease. For this multiple classification, they used one-against-one (OAO), one against-all (OAA), and error-correction code (ECC), where all of them were based on SVM. Highest accuracy of 81.11% was obtained from OAO method. A well cited research published in 2019 by N. Singh and P. Singh [21] studied on ECG dataset from UCI machine learning repository for arrhythmia prediction. There were 279 features on this dataset and the aim of this research was to select important feature. They selected important features with the help of 3 feature selection techniques named Chi-square Statistic, Symmetric Uncertainty (SU) and Gain Ratio (GR). They applied SVM, random forest (RF), and joint reverse algorithm program (JRip) as ML algorithm. The best result was achieved from GR which returned 30 features and RF outperformed the other classifier with the accuracy of 85.58%.

In 2019, another study by G. T. Taye et al. [22] used two different feature selection methods named HRV and the QRS complex shape for improving the performance of predicting VF onset 30s before its occurrence. They studied on two types of databases e.g., normal datasets from paroxysmal AF prediction challenge database and the MIT-BIH normal sinus rhythm database. 11 features were extracted using QRS complex shape features, and traditional HRV features. They used artificial neural network (ANN) classifier and made a comparison between these two feature selection methods. QRS complex shape feature had a better accuracy of 98.6% whereas the HRV feature has an accuracy of 72%.

In the year of 2020, authors of the paper [23] developed a unique model for detecting AF by analyzing the raw ECG data. They used both ML and deep learning (DL) based algorithm and achieved better performance with DL as compared to traditional ML. In contrast to ML, no feature selection method was required in DL. DL classifiers, such as convolutional neural network (CNN) and long short-term memory (LSTM), outperformed classical machine learning classifiers, such as multilayer perceptron (MLP) and logistic regression (LR), on the PhysioNet and MIT-BIH AF datasets. LSTM achieved the highest accuracy of 82.9% among all of the classifiers.

S. Wang et al. [24] predicted arrhythmia after acute myocardial infarction (AMI) and this research was published in 2021. First Affiliated Hospital of Harbin Medical University has provided data of 2084 subject and they used three ML models named DT, RF and ANN to predict tachyarrhythmia after MI. ANN returned the best prediction accuracy after Gini impurity feature selection method. This research has returned 66.8% prediction accuracy with ANN and ventricular wall motion as the most important feature.

In 2022 R. Hu et al. [25] and his team has detected arrhythmias from continuous single lead ECG signals. For this purpose, they have used two well-known databases: MIT-BIH arrhythmia database and MIT-BIH AF database and introduced a novel transformer-based DL NN named ECG detection transformer (DETR). After using 10-fold cross validation (CV), 99.12%, 99.49%, and 99.23% accuracy was obtained from respectively from 8, 4, and 2 distinct labels.

A brief overview of different ML based arrhythmia prediction approaches is shown in Table 1.1.

Table 1.1: State-of-art-works on arrhythmia prediction

Reference	PY	Algorithm	Outcome	Dataset
[15]	2013	SVM	98.1% accuracy from in-sample training data and 96.3% accuracy from out-sample training data.	AHA Database, CUVT database, and MIT-BIH MVA Database
[16]	2014	SVM, FS	HILB as the most important feature and TCSC, SpEn, and VFleak are the important features for detection. 99.7% detection accuracy.	UCVT database with 13 ECG parameters
[17]	2015	25 classifiers	DT returned highest classification accuracy and it is 87.21%.	UCI CA data set.
[18]	2016	SVM	94% sensitivity, specificity 82% from real time dataset and 86% from retrospective dataset.	PhysioNet/Computing in Cardiology
[19]	2017	DT, SVM, NN and KNN	99.30% accuracy with KNN classifier using genetic algorithm.	MIT-BIH arrhythmia database
[20]	2018	SVM based OAO, OAA, ECC	81.11% accuracy from OAO	IML Data Repository
[21]	2019	SVM, RF, and JRip	85.58% is the highest accuracy with RF classifier and GR feature selection method.	UCI ML repository
[22]	2019	ANN	QRS complex shape feature had a better accuracy of 98.6%	PAF prediction challenge database and the MIT-BIH NSR database
[23]	2020	CNN, LSTM, MLP and LR	LSTM classifier returned best accuracy of 82.9%	PhysioNet and MIT-BIH AF datasets
[24]	2021	DT, RF, ANN	Highest accuracy was 66.8% with ANN classifier, the most important feature was ventricular wall motion	FAH of HMU
[25]	2022	a novel transformer-based deep learning neural network named ECG DETR	After using 10-fold CV, 99.12%, 99.49%, and 99.23% accuracy was obtained from respectively from 8, 4, and 2 distinct labels.	MIT-BIH arrhythmia database and MIT-BIH AF database

- PY → Publication Year
- AHA → American Heart Association
- CUVT → Creighton University Ventricular Tachyarrhythmia
- MVA → Malignant Ventricular Arrhythmia
- CA → Cardiac Arrhythmia
- IML → Irvine Machine Learning
- PAF → Paroxysmal Atrial Fibrillation
- NSR → Normal Sinus Rhythm
- FAH of HMU → First Affiliated Hospital of Harbin Medical University

From these studies, we can see that most of the researchers have used ML classifiers to predict and classify arrhythmias. In the earlier days, most of them preferred the SVM classifier as an ML algorithm, but in this thesis, we have used RF classifier to predict arrhythmia. From these state of the art works we also see that no one has worked on time complexity. We have also used two different feature selection approaches to find out the important features that are highly responsible for the occurrence of arrhythmia and reduce the time complexity.

1.2 Motivation

The diagnosis and treatment of cardiac arrhythmia are expensive and time consuming. Various clinical tests such as ECG, angiogram, chest x-ray is required to diagnose arrhythmias. Some of these tests are not only expensive but also complicated. Besides, this life-threatening disease puts the patient's life at risk of death in minutes [26]. In such a condition, patient needs immediate medical attention, which is not always possible. So, we should focus on the prevention of this disease. Clinical data analysis of patients on a regular basis can facilitate in estimating the likelihood of having arrhythmia. Manual processing of each patient's clinical data might be a time-consuming and inappropriate approach. However, there are several new technologies such as statistical analysis, ML, DL, artificial intelligence (AI) and internet of things (IOT) that easily process the enormous volume of data within a shortest possible time in the medical field. Technologies based on ML are more popular for analyzing medical data because of their usefulness. This System allows us to predict the exacerbation events of any particular disease with better accuracy. In the context of arrhythmia, many clinical features indicate the presence of arrhythmia [27]. But not all features have equal contribution in predicting this disease. Important features increase the accuracy of prediction. We can find out these features with the help of statistical analysis as well as ML. Using different statistical tests such as Chi-square test, gamma coefficient we can easily determine the important features that have close association with arrhythmia. We can also select the important features and predict arrhythmia with the help of ML. The use of statistical analysis and ML can detect arrhythmia early and reduce the risk of harm and mortality rates caused by this disease.

According to existing medical system different clinical tests are required to diagnose arrhythmia. Doctors confirm arrhythmia by analyzing the reports of these clinical tests. It is a very time-consuming process. Because of the lengthy diagnosing process, the patient's condition deteriorates. In order to solve this problem, in this thesis we have introduced an efficient approach for on-time prediction of arrhythmias using ML and statistical techniques.

1.3 Objectives of the Thesis

Compared to clinical diagnosis process, in this thesis, an effective method for predicting arrhythmia has been introduced. For completing this task, our first aim was to find out the important features that are responsible for the occurrence of arrhythmia. Instead of all features, we can easily predict arrhythmia with the help of important features only. Our next objective was to reduce the time complexity so that we can predict arrhythmia within a reasonably short time. Our third objective was to analyze the performance of the model that we have used in this thesis. The final aim is to investigate the AMI dataset to find out the reason why the prediction accuracy of the model does not meet our expectations.

1.4 Contribution of the Thesis

After completing this thesis, we have successfully determined the significant features for predicting arrhythmia. We have identified them using two different methods. These features will help all cardiac patients and doctor in early prediction of this life-threatening disease. It is a little positive contribution to society from our thesis.

Besides this, we have contributed to the reduction of time complexity for prediction. Time complexity is a great factor in arrhythmia prediction, as this disease puts the patient’s life at risk of death in minutes. As soon as this disease can be predicted, the hospitalization and death rates will decrease proportionally. We have used the same dataset and classifier that S. Wang et al. [24] used. But we are able to increase the prediction accuracy from them [28] which is one of our contributions. Finally, instead of using all features, we have predicted arrhythmias using selected features and the prediction accuracy in both cases is almost the same.

1.5 Organization of The Thesis

This thesis is organized into five chapters. An outline of this thesis is given below-

Chapter 1

The background and motivations of our thesis are discussed in the “Introduction” chapter. In addition, it also represents the objectives and contribution of our study.

Chapter 2

The bibliometric literature review of our related studies is described in the chapter under “Literature Review”. It includes all the details of paper collection, selection, and various performance analysis methods.

Chapter 3

The overall methodology of our work, our dataset, and various statistical and shallow learning techniques are described in the “Materials and Methods” chapter.

Chapter 4

The key features for predicting arrhythmia are discussed in the chapter “Result and Discussion”. It also includes the prediction performance of proposed model. In this chapter, we have also compared our work with other relevant studies.

Chapter 5

The conclusion and future scope of this research are discussed in the final chapter, titled "Conclusion and Future Work".

Chapter 2

Literature Review

A literature review is a complete overview of related research on a specific topic that summarizes present knowledge, relevant approaches, methods, knowledge gaps, and the future scope of existing research. Literature review can be performed through different ways such as systematic literature review, meta-analysis and bibliometric analysis. In this study, we have chosen bibliometric analysis for literature review because it provides a lot of information in a concise way. It is a quite beneficial tool for literature reviews as it analyzes, tracks, and evaluates the quantitative connections and effects of publications in a particular field of study using mathematical and statistical techniques. This technique is effective for quickly identifying significant research, scholars, journals, institutions, and countries within a certain time period, as well as for providing a quantitative overview of huge amounts of academic literature. In this research, we have done bibliometric analysis based on some significant factors, including relevancy analysis, trend review, citation evaluation and networking analysis. In addition, a wide range of indicators such as, bibliographic coupling, co-citation, co-occurrence of keywords and collaboration has been used for mapping the bibliographic data graphically.

2.1 Methods of Bibliometric Review

Due to the tremendous advancement in scientific technology, a variety of bibliometric methods and applications have been developed to help researchers to conduct their research. This rigorous approach for quantitative research analyzes the interconnections and the effects of publications in a particular field of research using statistical and mathematical method. The term bibliometric was first introduced by P. Otlet in 1934 and described as "the measurement of all aspects related to the publication and reading of books and documents" [29]. To perform bibliometric analysis, we have organized our work into five different parts named as topic selection, searching tool, PRISMA, data mining tool and performance analysis. The methodology of bibliometric analysis is shown in Fig. 2.1.

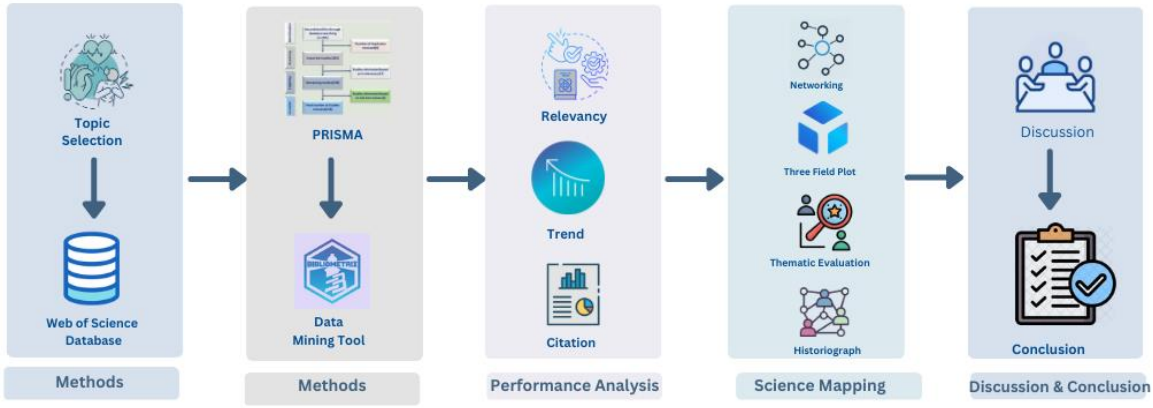


Fig. 2.1: Proposed methodology for bibliometric analysis

2.1.1 Topic Selection

For bibliometric analysis, we have selected “arrhythmia detection and classification” as our topic. It is a life-threatening disease across the whole world, and people of any age can be affected by this disease. So, we should pay more attention to the prevention, early detection, and prediction of arrhythmias. We have selected this topic to investigate more about the historical development and current situation of arrhythmia detection and prediction. This study will help the researchers to find those authors, institutions, journals, and members of the community who are working in this research field and who have the highest contribution to arrhythmia detection and prediction.

2.1.2 Searching Tool

Arrhythmia has already been the subject of extensive research so there are many research paper on arrhythmia detection and classification [30], [28], [31]. For identifying the relevant papers on a specific topic, there are various searching tools are available such as web of Science (WoS), Scopus and PubMed. As we don’t have academic access to Scopus, we have used WoS as a searching tool for this study. Web of Science is a research database which provides access to scientific and scholarly journals, articles, and conference papers. According to 2020, it was one of the largest databases in the world, with 74.8 million records [32]. For this analysis, we initially performed a search in the WoS database using the keywords "arrhythmia detection" and "arrhythmia classification" and the combined use of two keywords, such as "arrhythmia detection and classification." The two key search parameters used in this process were the document type "article" and the research years

"2005-2022." This searching tool revealed 283 papers on arrhythmia detection and classification.

2.1.3 PRISMA

Though there have been 283 papers on arrhythmia detection and classification published throughout this time, not all are directly related to our analysis. So, in order to improve the accuracy of our analysis, we have followed the principles of PRISMA proposed by the researches Moher et al. [33]. The information flow across the various phases of a bibliometric review is described in PRISMA (Fig. 2.2). This four-phase flow diagram represents the number of documents identified, selected and discarded, together with the justifications for exclusions.

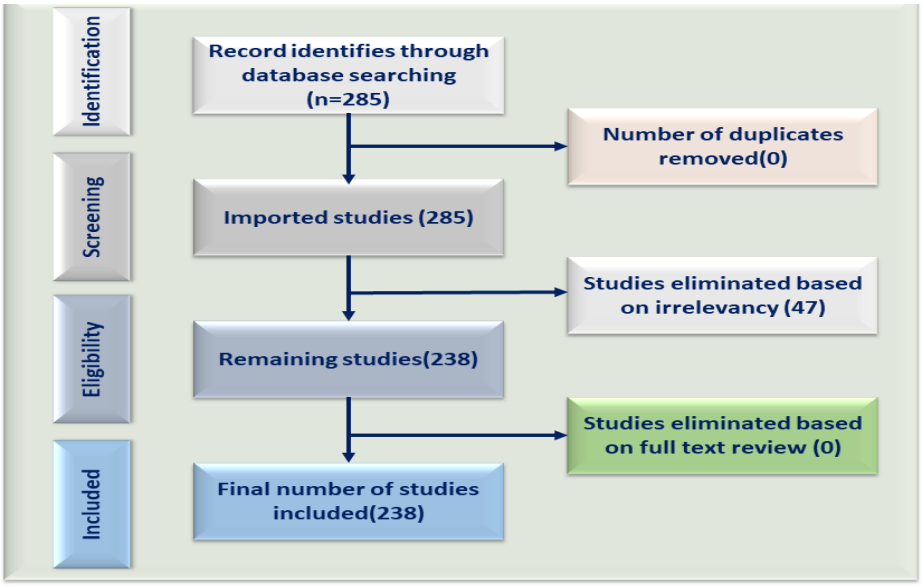


Fig. 2.2: The PRISMA flow diagram

Identification is the first step of PRISMA method. In this step, we have identified 285 papers from WoS database. Out of 285 papers, we did not have any duplicate paper. So, in the next stage, we have imported 285 papers as before. The following step was the elimination of irrelevant documents based on our research topic. In this step, we have omitted 47 articles, because of their irrelevancy and 238 articles remain in the database. As we did not have any articles based on full text review, 238 papers indicate the final selected papers for bibliometric analysis using PRISMA method.

2.1.4 Data Mining Tool

The selected papers can be analyzed with the help of different data mining tools such as Biblioshiny, VOSviewer, Gephi, HistCite, and CiteSpace. For this study, we have selected the Biblioshiny software to analyze, evaluate, and develop the graphical visualization from our database. Biblioshiny is a tool of R statistical programming language developed by M. Aria and C. Cuccurullo [34] and designed for quantitative evaluation. The user-friendly interface of Biblioshiny makes it simple for users to import, modify, and generate interactive visualizations of data. A variety of visualization options are offered by Biblioshiny, such as bar graphs, line plots, and maps, which help researchers in conducting their research. It can be integrated with other data mining software, including R and Excel that allow users to further evaluate and manipulate data. We have performed all of our analysis through this software.

2.1.5 Performance Analysis

In order to evaluate the research performance, growth, scientific trending in the field of arrhythmia detection and classification we have used different bibliometric attribute such as citation analysis, trend analysis, network analysis and others. Additionally, we performed this analysis from a variety of bibliometric aspects, including the performance of the authors, journals and institutions. In this paper, we have conducted two types of bibliometric analysis. First, we have done performance analysis and next we have focused on science mapping. Performance analysis primarily takes into consideration the contributions of an individual, a group, or an organization to a certain research topic while scientific mapping emphasizes on visualizing the relationships and interconnections between different authors, journals, or institutes. Performance analysis is usually used to identify significant authors, sources, countries, or affiliations in a certain research field and science mapping is used to figure out the historical development, identify gaps in the literature, and identify emerging or declining research trends. In the next sections, we have explained performance analysis and science mapping on selected 238 papers.

2.2 Bibliometric Performance Analysis for Arrhythmia Detection and Classification

In bibliometric studies, the contribution of researchers in a particular research field and overall progress of that field are investigated through bibliometric performance analysis. A bibliometric performance analysis was conducted in this section based on various performance indicators, such as the number of publications, trends, number of citations, leading authors, institutions, countries or regions etc. It provides a comprehensive knowledge and understanding of the myocardial arrhythmia detection and prediction research from 2005 to 2022.

2.2.1 Leading Countries, Authors, Affiliation and Sources

This section represents the most contributing countries, authors, institutions, and sources on myocardial arrhythmia detection and prediction research based on the number of their publications.

2.2.1.1 Most Productive Countries

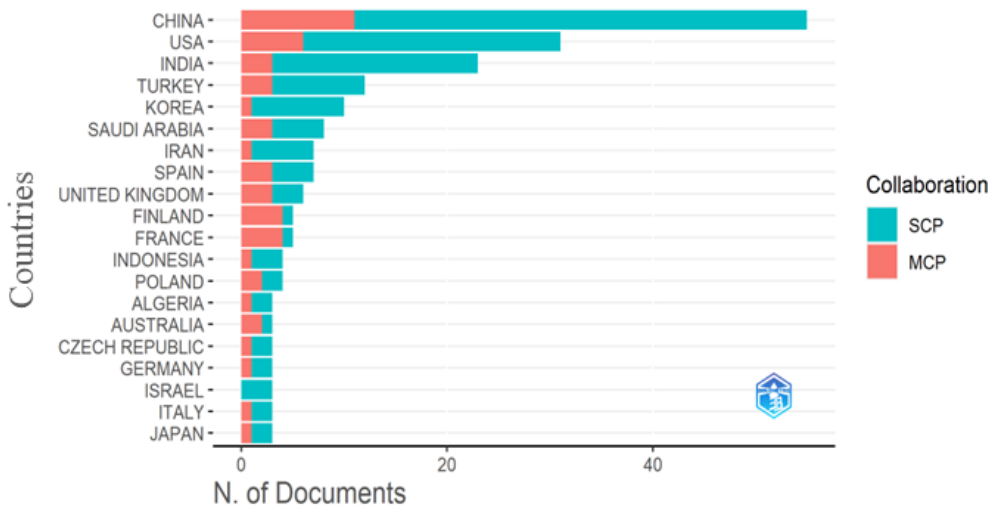


Fig. 2.3: Most productive countries

Fig. 2.3 represents the top 20 most productive countries or regions for arrhythmia detection and classification. The leading countries in terms of arrhythmia detection and classification are those that have produced the highest number of publications on this area. In the figure, the blue box indicates single country production whereas the red box denotes multiple country production from 2005 to 2022. The total amount of documents includes both those

that come from a single country and those that are created in association with other countries. China and USA outperform all other countries, showing that they began their research efforts before most other countries around the world. According to this analysis, China is the most productive country and USA is the second. China has a total of 55 articles, whereas the total number of articles in the USA is 31, which is not close to the number of articles in China. India is the third most productive country. It has 23 articles in total, which is almost two-thirds of the USA in terms of numbers. After India, other countries have a less significant number of articles on arrhythmia detection and classification. Korea and Turkey have published 12 articles and 10 articles respectively. Out of these twenty countries, the last eight countries have only 3 articles, which is very less compared to China. From this analysis, we can conclude that China has performed more research on arrhythmia detection and classification than any other country and after China, the USA has shown more interest in that area.

2.2.1.2 Most Relevant Authors

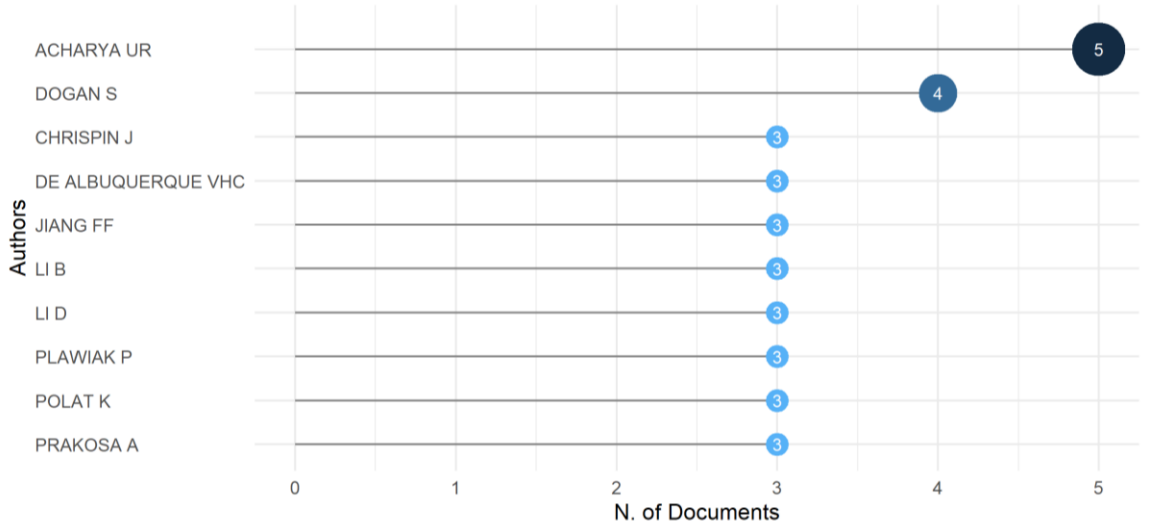


Fig. 2.4: Most relevant authors

Regarding the analysis at the author's level, the most relevant author is one of them. In the context of academic research or writing, a relevant author is one whose work is directly related to the topic being discussed and can provide valuable insights or information on that topic. Several authors around the world have conducted research in the field of arrhythmia. Fig. 2.4 lists the top 10 most relevant authors based on their contributions to

arrhythmia research in terms of published articles. The author named U. R. Acharya [35], [31], [36], [37], [30] ranks first with 5 contributions and N. S. Dogan [38], [39], [30], [40] contributes 4 publications. The remaining eight authors contributed equally with 3 publications. From this analysis, we can say that all authors' contributions in the research field of arrhythmia are closely related.

2.2.1.3 Most Relevant Affiliations

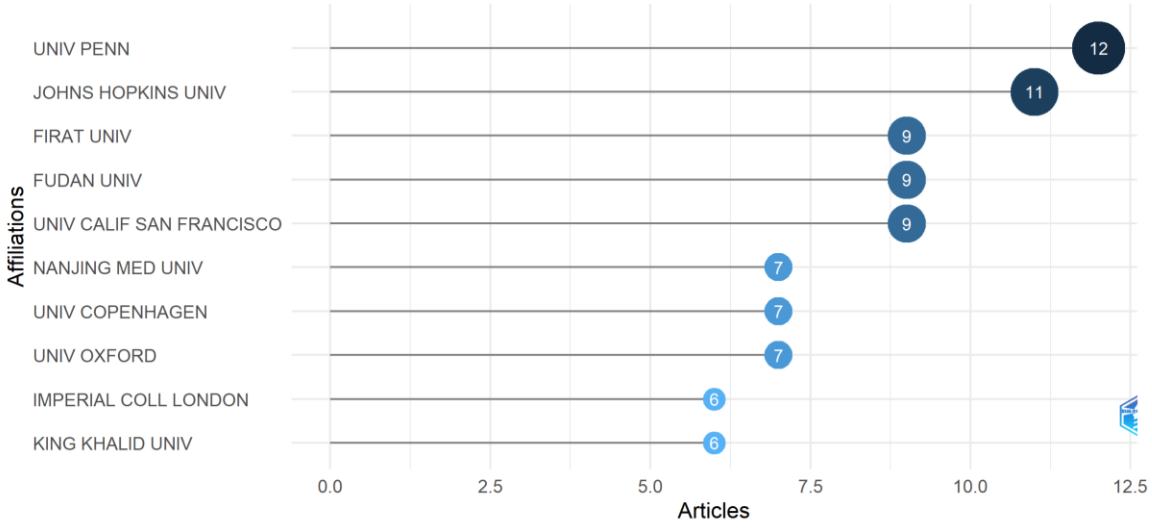


Fig. 2.5: Most relevant affiliations

Further, the most relevant affiliations were also investigated. A relevant affiliation is an organization or institution that is directly related to the place where the similar arrhythmia research is being done. We classify the top affiliations based on the number of research and published articles on arrhythmia. In Fig. 2.5, it has been shown that the USA is one of the most productive countries based on the number of published papers. Again, the analysis on most relevant affiliations reveals that the University of Pennsylvania, which is located in the USA is the most relevant affiliation with 12 published articles. The Johns Hopkins University in the USA is the second relevant institution with 11 publications. Firat University in Turkey, Fudan University in Shanghai, China, and the University of California in San Francisco were the next three productive institutions with 9 publications each. The rest of the top listed universities have fewer publications compared to those mentioned above. It is noticeable that out of the top 10 universities, 4 universities are from USA and China.

2.2.1.4 Most Relevant Sources

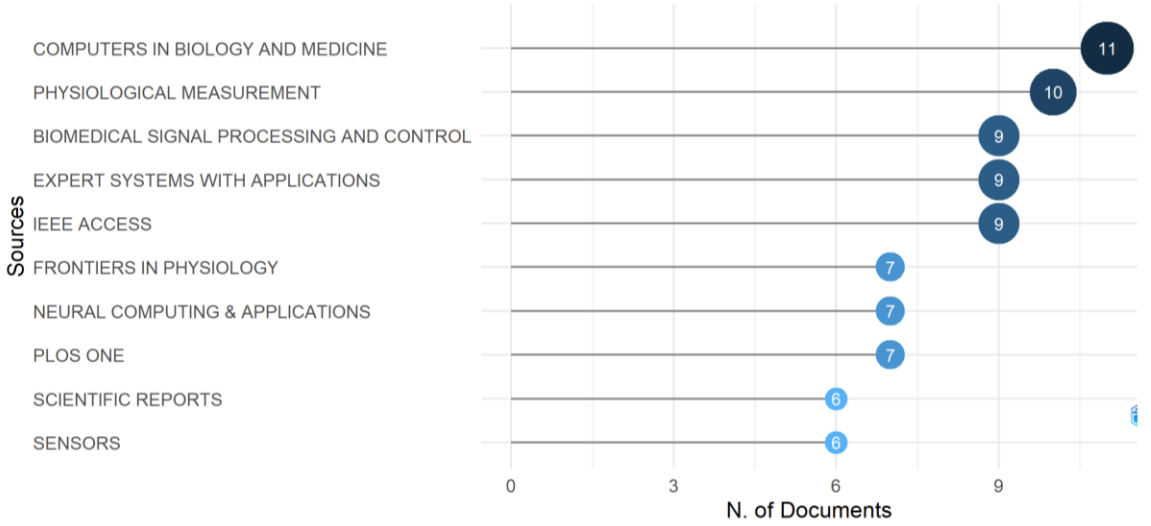


Fig. 2.6: Most relevant sources

The analysis of the most relevant sources is an important tool to evaluate the performance of a journal in a specific field. This Fig. 2.6 lists the name of the ten journals that have published the most papers in arrhythmia research. The journal named "Computer in Biology and Medicine" has achieved the top position with 11 publications. The second prolific source is the "Physiological Measurement" journal which focuses on clinical research and practice. This source has published 10 papers on arrhythmia detection and classification. The next three journals named "Biomedical Signal Processing and Control", "Expert System with Applications", and "IEEE Access" have achieved equal positions by publishing 9 papers. Interestingly, there is hardly any difference in performance among these top 5 journals. Compared to the leading 5 journals, rest of the journals named "Frontiers in Psychology", "Neural Computing and Applications", "PLOS One", "Scientific Reports and Sensors" have less publication. Among the top ten sources, the last two journals have published 6 papers each. This study reveals all the journals that publishes arrhythmia related paper mostly.

2.2.2 Trend Analysis

Trend analysis in academic research involves the identification and analysis of patterns or trends in specific research area over a period of time. It can be useful for a variety of purposes in academic research, such as determining whether certain factors or variables are changing over time or not that provide insights into the relationships between different variables in the data.

2.2.2.1 Word Cloud of Keywords

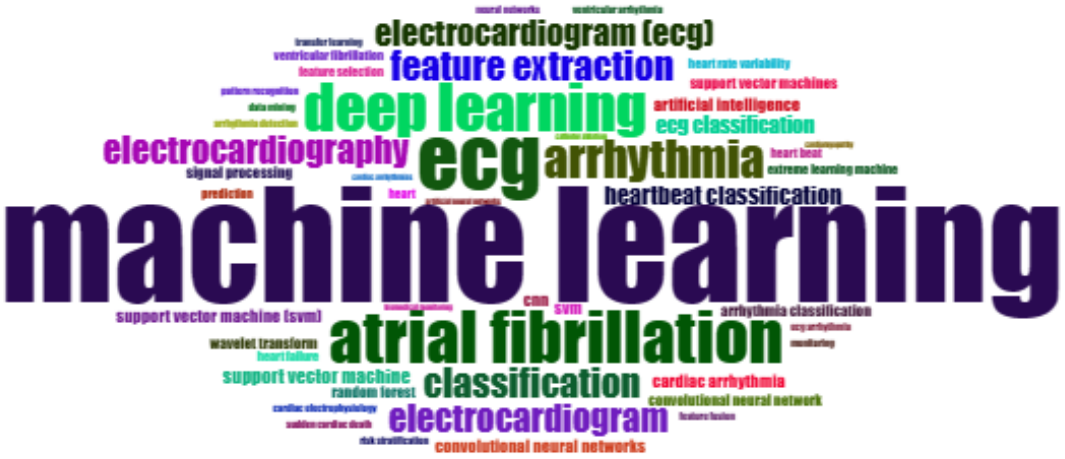


Fig. 2.7: Word cloud of keywords

Fig. 2.7 represents the 50 most frequently used author keywords using a word cloud where the size of each keyword represents its frequency. The term "machine learning" has been used most frequently in the field of arrhythmia research from 2005 to 2021. During that time, most of the researchers preferred machine learning approach for their arrhythmia research. “ECG” is the second most explored keyword after machine learning. “Deep learning”, “atrial fibrillation”, and “arrhythmias” all have smaller physical dimensions than the “ECG”, indicating that these three terms are used less frequently than the term “ECG”. In addition, the other available keywords in the word cloud aren't used very often; their occurrence rate is very less.

2.2.2.2 Growth of Top 10 Author's Keywords

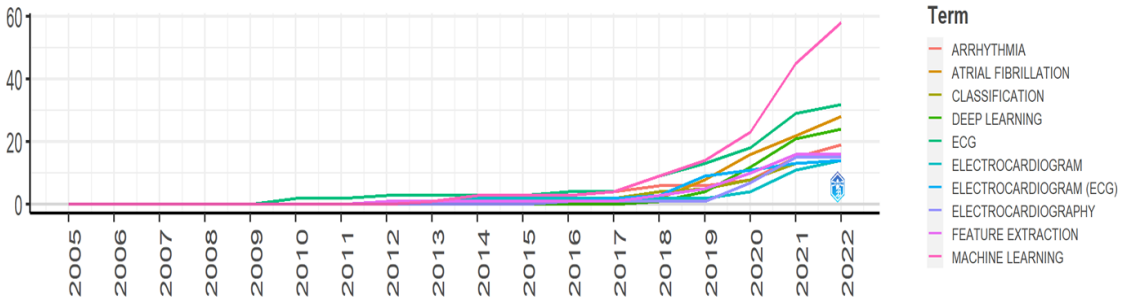


Fig. 2.8: Growth of top 10 keywords

Fig. 2.8 depicts the growth of the top 10 author's keywords from 2005 to 2022 and the keywords are "arrhythmia," "atrial fibrillation," "classification," "deep learning," "ECG," "electrocardiogram," "electrocardiogram (ECG)," "electrocardiography," "feature extraction," and "machine learning." During that time, every keyword has experienced a significant gradual increase in growth rate. The keyword has an annual growth, and researchers need to be aware of the most recent development. In this figure X-axis represents year and the y-axis represents the number of cumulative occurrences. In the research field of arrhythmia detection and classification, the keyword "machine learning" has a great performance curve. This keyword has the highest growth rate in this field. Though the term "machine learning" has started its journey from 2005, its growth rate has reached to 58 by the end of 2022. None of the keywords except "machine learning" we chose to analyze had ever been used from 2005 to 2009. Further, in 2010, the term "ECG" first came to the author's attention, which is currently, in the growth list, and it is in the second position. Its number of cumulative occurrences is 32, which is very low compared to the keyword "machine learning". In 2022, the third highest growth keyword is "atrial fibrillation", which has an occurrence number of 28. "Deep learning" has the fourth highest growth rate, and it is very close to "atrial fibrillation". After "deep learning", arrhythmia obtains the next growth rate. It's noteworthy that "feature extraction", "classification", "electrocardiography", "electrocardiogram", and "electrocardiogram (ECG)" have almost equal growth rates.

2.2.2.3 Trending Topics

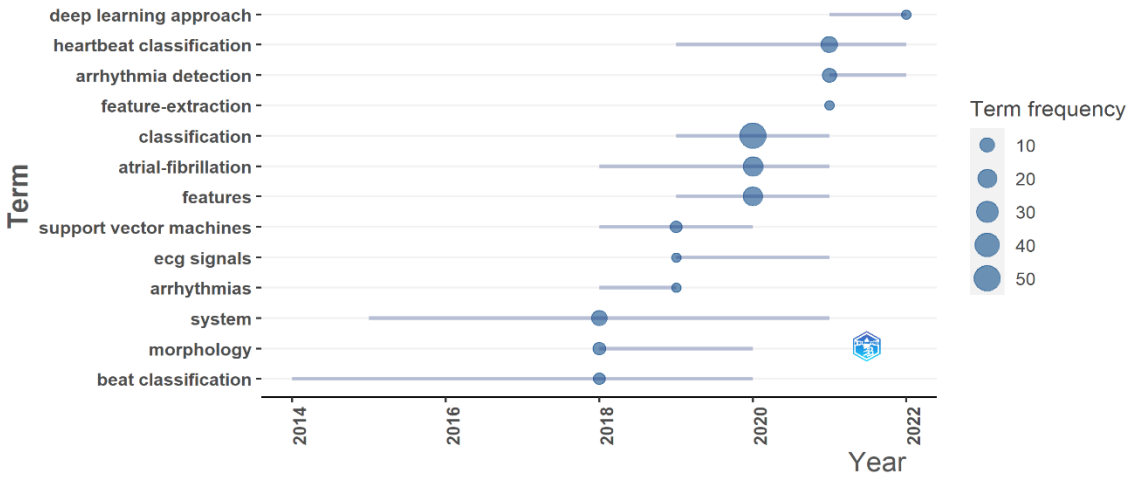


Fig. 2.9: Trending topics

Trend analysis is a useful tool for researchers because it reveals people's preferences as well as what is currently occurring. Fig. 2.9 shows the trending topics from 2014 to 2022 in the field of arrhythmia detection and classification. Every year, trending topics change based on their use in the research field. The size of the blue circle in the figure denotes the topic's frequency. It starts from 10, which is represented by the smallest circle, and the highest frequency is 50, which is denoted by the largest circle. At the beginning of the arrhythmia research, the trending topic was "beat classification," and it predominated from 2014 to 2020. In 2016, the "support vector machines" came into trend, and its trending duration lasted till 2020. In 2018, "beat classification", "morphology" and "system" got researchers' attention. However, among these three topics, "system" was able to attain the highest attention. It is interesting that in 2019, three topics, "ecg signals", "support vector machine" and "arrhythmias", had the same popularity. In the next year, the most significant topic was "classification", which beats the two important topics "atrial fibrillation" and "feature". Again in 2021, three topics equally attract scholars' attention. These three topics were "neural network", "heartbeat classification" and "arrhythmia detection". "Deep learning approach" is the only hot topic of 2020, but its popularity is very poor. From this analysis, we can say that the topic "beat classification" was on trend for the longest period. Also, we can hope that "arrhythmia detection" and "deep learning approach" will be on

trend in the future because among all the topics only these two topics became visible since 2021.

2.2.3 Citation Analysis

Citation analysis is an important factor in the field of bibliometrics. It is performed by looking at how frequently a work is cited, where and by whom it is cited, and the context in which it is cited. It is the study of the effect and significance of research and scholarly output. This analysis can be used for investigating the influence of a particular author and work. Also, it evaluates the quality and determines the importance and relevance of a research work to a particular topic.

2.2.3.1 Most Cited Countries

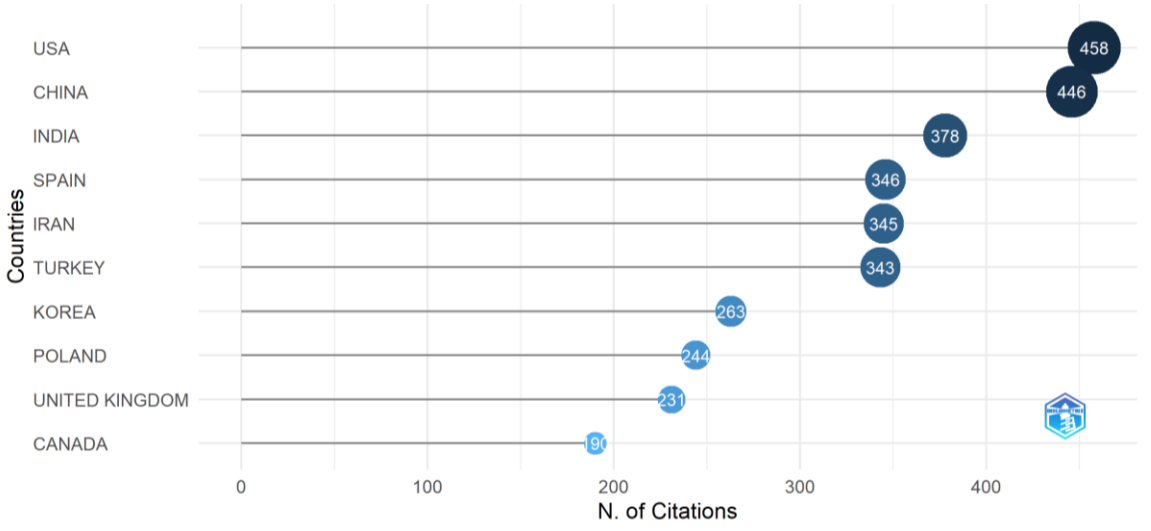


Fig. 2.10: Most cited country

Citation analysis is a very effective tool for determining the research significance of an article. Here the list of the most-cited country highlights the countries that make the biggest contributions to seedling research in the field of arrhythmia detection and classification. Fig. 2.10 represents the top 10 most cited countries in the world. Although China published the highest number of papers on arrhythmia detection and classification, the papers published in the USA are the most cited. This may be because the paper published in the USA is more relevant to arrhythmia classification and detection and contains more significant information. When ranked by the number of citations, the USA has the greatest scholarly impact with 458 citations, followed by China with 446 citations. India is the

third most cited country in the world for arrhythmia research. 378 papers have published on arrhythmia detection and classification from India. An interesting thing is that middle eastern countries like Iran and Turkey are ranked in the middle of the list. They have around 350 citations each. Canada ranks last among the 10 countries with 190 citations. It is noteworthy to mention that the majority of articles on the detection and classification of arrhythmia have come from the USA and China.

2.2.3.2 Most Cited Authors

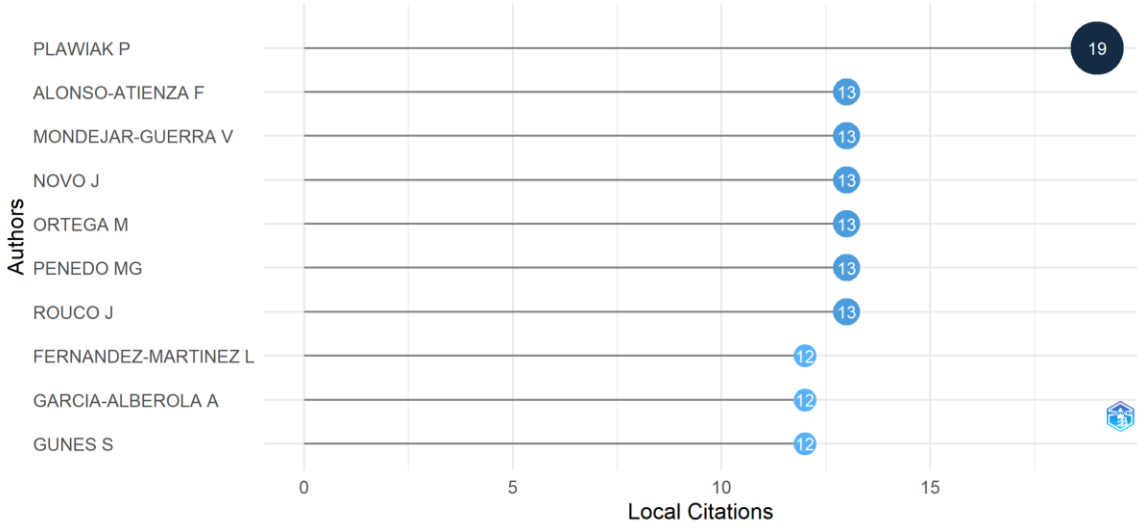


Fig. 2.11: Most cited author

The next interesting analysis is the most cited author, which carries a noteworthy significance to evaluate the performance of an author in a specific research field. Fig. 2.11 displays the name of the 10 most local cited authors in the field of arrhythmia classification and detection with their citation numbers. The author named P. Plawiak [31] ranks first. He has 19 citations, which is the highest number of citations between the years 2005 and 2022. It is very interesting that after P. Plawiak, six authors achieve the second position. They all have 13 citations each. The names of these six authors are F. Alonso-Atienza [16], V. M. Mondéjar-Guerra [41], J. Novo [41], M. Ortega [41], M. Gonzalez Penedo [41] and J. Rouco [41]. Lastly, we can see that three of the remaining authors have been cited 12 times in several publications. From this figure we can conclude that each author has at least 12 publications but not more than 20.

2.2.3.3 Most Cited Sources

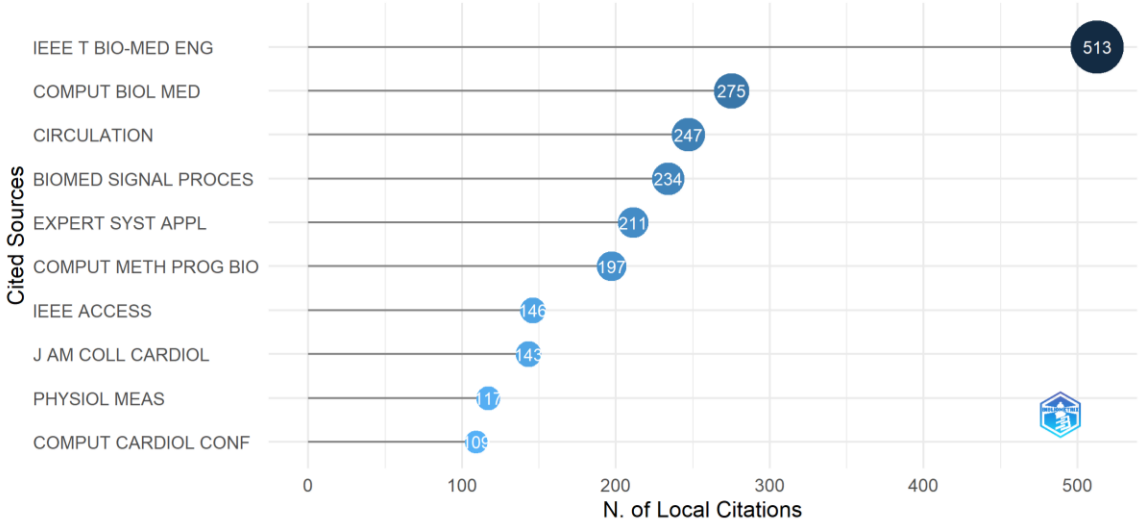


Fig. 2.12: Most cited source

Analyzing most cited source plays an important role in citation analysis. The most cited sources refer to those that have been cited most frequently by other authors in their research. We can understand the impact of a journal in specific research work by analyzing most cited sources. Above Fig. 2.12 shows that, journal “IEEE Transaction on Biomedical Engineering” has the most citation of 513 times. A lot of research has referred the publications of this journal, which makes it one of the major journals in the field of arrythmia detection and classification. Another major journal in this field is "Computers in Biology and Medicine ". Around 275 research paper on arrhythmia has cited the publications of this journal. Close to “Computers in Biology and Medicine”, “Circulation” is another journal with 247 citations. Besides the three most cited journals, there are “Biomedical Signal Processing and Control”, “Expert Systems with Applications”, and “Computer Methods and Programs in Biomedicine” with citations of 234, 211, and 197 each. Though the number of citations in these journals is not high as “IEEE Transaction on Biomedical Engineering”, the papers of these journals also have a decent number of citations, which makes them one of many major publications. A few noteworthy journals are “IEEE Access”, “Journal of the American College of Cardiology”, “Physiological Measurement and Computing in Cardiology” with citations of 146, 143, 117, and 109,

respectively. Analyzing this figure, we can conclude that most of the authors have cited the journal named “IEEE Transaction on Biomedical Engineering”.

2.3 Science Mapping

Science mapping is an approach for identifying and visualizing the connections between various scientific concepts and research domains. Usually, science mapping represents a network diagram. These network diagrams and the connections between them can be based on a variety of analyses, including the patterns of citations between papers, the collaboration of institutes in research publications, or the co-citation of authors.

2.3.1 Networking Analysis

Network analysis is a method of analyzing the relationships between various journals, affiliations and countries. In academic research, network analysis can be carried out in a variety of ways such as co-citation network of journals, collaboration network of institutes and so on. One of the main advantages of network analysis is that it helps researchers to identify chances for collaborative learning or for teamwork with other researchers who are working on related subjects.

2.3.1.1 Co-citation Network of Journals

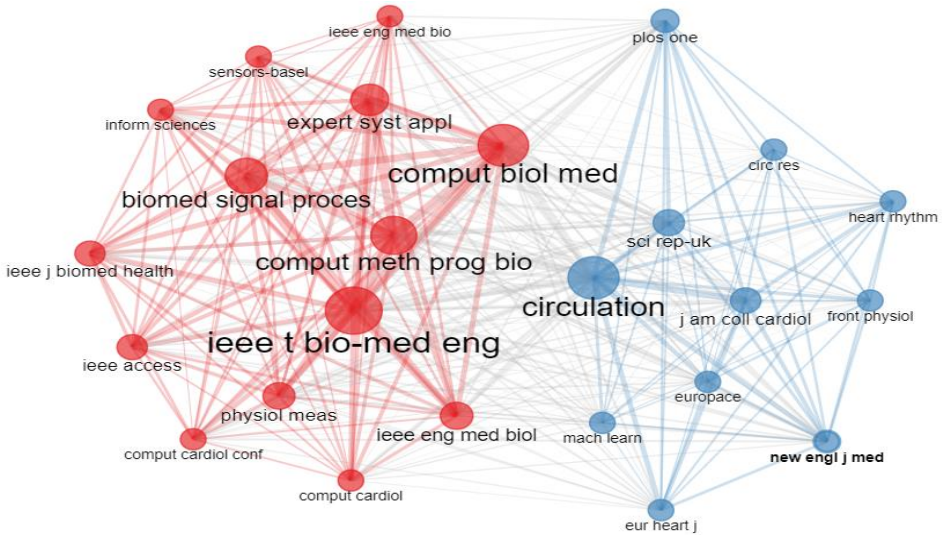


Fig. 2.13: Co-citation network of journals

Now in order to understand the relationships between different journals we will explore the co-citation network of journals. When two publications from separate journals are cited in

the same article, this is known as a co-citation of journals. Fig. 2.13 shows the co-citation network of the journal, and it contains 25 journals. This figure has two clusters, which are made of two different colors, one cluster is red and the another one is blue red cluster may represent journals that focus on biology with technology, while another cluster may represent journals that focus on machine learning, cardiovascular diseases and so on. The journals are represented by the colored nodes, and the co-citation among the two journals is indicated by the line between them. The size of the nodes represents their citation frequency. From Fig. 2.13, we can see that the red clustered journals "IEEE Transaction on Biomedical Engineering (ieee t bio-med eng.)", "Computers in Biology and Medicine (comput biol med)", and the blue clustered journal "Circulation" are the most cited journals, and their network of citations is vast. Some journals, such as "Computer Methods and Programs in Biomedicine (comput meth prog bio)", "IEEE Transaction on Biomedical Engineering (ieee t bio-med eng.)", "Computers in Biology and Medicine (comput biol med)" and so on, have cited each other papers and made the red cluster. The blue cluster also develops according to their citation and linkage with one another. In case of arrhythmia detection and classification, researchers can track changes in the relationships between different journals by creating co-citation networks and can find out the most influential journals in this research field. Analyzing this figure, we can conclude that there exists a strong relationship between the journals named "IEEE Transaction on Biomedical Engineering (ieee t bio-med eng.)", "Computers in Biology and Medicine (comput biol med)", “Computer Methods and Programs in Biomedicine (comput meth prog bio)” and “Biomedical Signal Processing and Control”.

2.3.1.2 Collaboration Network of Institutions

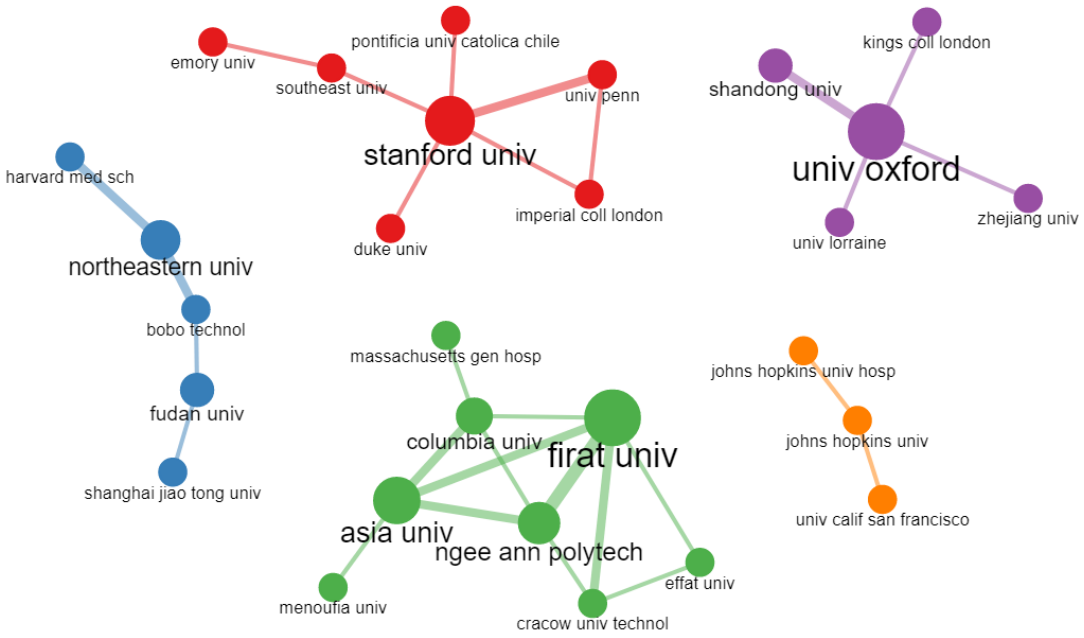


Fig. 2.14: Collaboration network of institutions

The collaboration network of the institutions in the research field of arrhythmia detection and classification is shown in Fig. 2.14. In this figure, the institutions are represented by colored circles. Lines of the same color represent the collaboration between institutions. Based on their collaborative research activities, there are 5 clusters with 5 different colors such as red, blue, green, orange and purple. Every institution in a specific cluster has conducted their research activities in collaboration with other institutions and they may be located in the same of different geographical areas. The size of the circle indicates the frequency of publication from a specific institution. In the figure, green cluster is the largest cluster. In this cluster, there are eight universities that are from the different geographical area such as Firat University from Turkey, Columbia University from USA Effat University from Saudi Arabia, and they have worked in close collaboration with each other. Firat University has developed the highest collaborative effort with other universities. In the red cluster, Stanford University, Duke University, University of Pennsylvania, and Emory University are located in the USA whereas Imperial College London is located in UK, Pontifical Catholic University in Chile and Southeast University is located at China. The rest of the clusters are also developed in the same way. Therefore, from this figure, we

can say that the institutions of green clusters have a strong collaboration between them and Firat University from Turkey is the most collaborative institution.

2.3.1.3 Collaboration Network of Countries

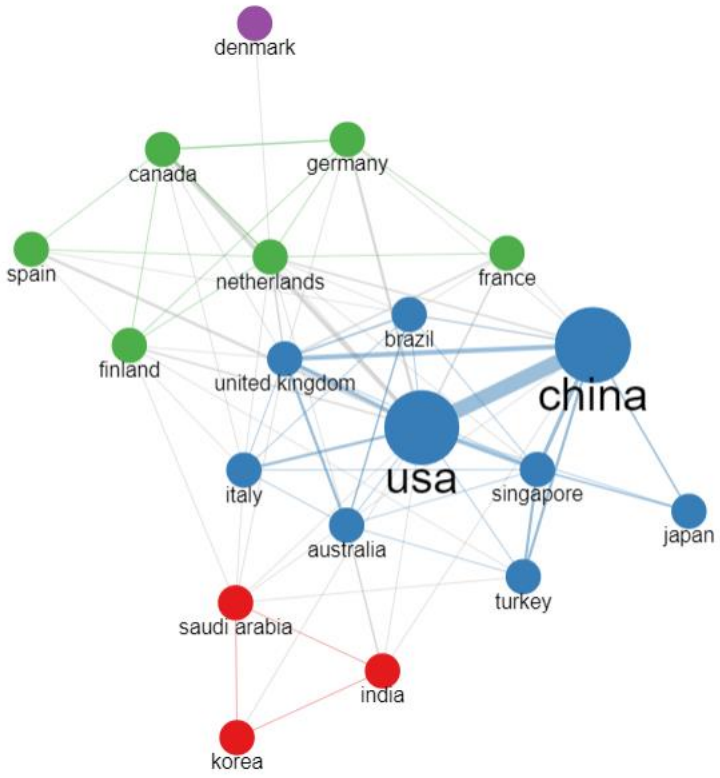


Fig. 2.15: Collaboration network of countries

Fig. 2.15 the collaboration network between the countries that have worked on arrhythmia detection and classification. In this figure, different colored (red, blue, green and purple) circles and lines are used to identify the collaboration among the countries. The connection between countries of the same color indicates their collaboration and the thickness of the line represents the strength of collaboration. At the same time, size of the circle also indicates their collaboration frequency. According to this figure, China and the USA have more collaborative partner than any other country and the collaboration between them is strongest. In addition, the blue cluster is the largest cluster than other clusters, which indicates that the number of countries cooperating together for detecting and classifying arrhythmia under this cluster is more. There are six countries under the green cluster, which

means that Canada, Spain, Finland, Germany, Netherlands and France are working together. The red cluster contains three countries and the purple cluster has only Denmark, which works in collaboration with only Netherlands. Therefore, it can conclude that both the largest and strongest network are in blue cluster.

2.3.1.4 Collaboration Map for Different Continents and Subcontinents

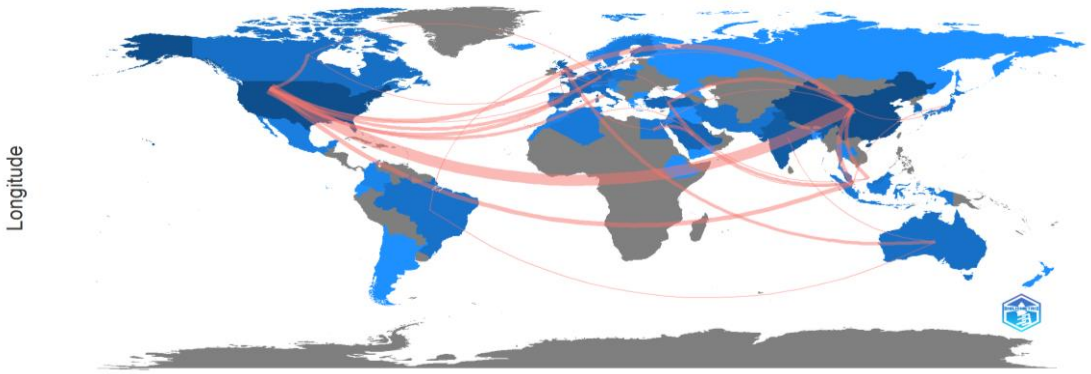


Fig. 2.16: Collaboration map for different continents and subcontinents

An overview of global collaboration on arrhythmia detection and classification research between different continents and subcontinents is provided in Fig. 2.16. The USA is located in North America and China is a country of East Asia but the collaboration between these two countries is the best. A similar collaboration network has been found between these two countries in Fig. 2.15. It is evidence of international collaboration. For global development in research, international collaboration is very important.

2.3.2 Overview

The complete overview of different performance analysis is given in this section using a thematic evaluation, three-field plot, and historiograph. It also shows relationship between different analysis.

2.3.2.1 Thematic Evolution

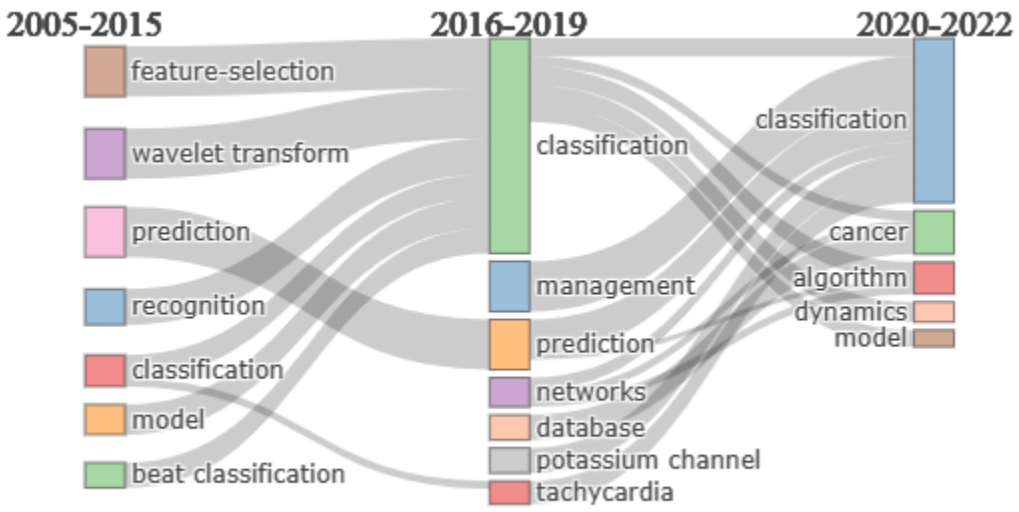


Fig. 2.17: Thematic evolution

The analysis of the thematic evolution is another important topic, which is shown in Fig. 2.17 using a Sankey diagram. A Sankey diagram is a graphical representation of a flow of information, resources and other quantities between nodes in a system [42]. The width of the lines that connect the nodes in a Sankey diagram is proportional to the amount of flow [42]. In this figure, each box, which is mainly called node denotes one theme. Height of the box denotes the occurrence rate of each theme. This diagram represents the connection between different themes.

It is noteworthy that there exists a variation between themes as different themes were preferred by the scholars at various times. We can see that though the “classification” emerged between 2005 and 2015, researchers have become more interested in this theme in the period from 2016 to 2019. Also, from 2020 to 2022, it was able to keep the researchers' interest almost in the same way. Initially, some themes were quite popular such as “feature selection”, “wavelet transform”, “recognition”, “model”, and “beat classification”. But over time, each of these melded into “classification”, which attracted additional researchers to the “classification”. Some themes that appear from 2016 to 2019 such as “management”, “network”, “database” and “tachycardia” did not develop later as an individual theme but rather combined with “classification”. Analyzing this figure, we

can see that from 2005 to 2022, the only theme that remained consistent was classification and it had the highest occurrence frequency in the time period from 2015 to 2019.

2.3.2.2 Three Fields Plot

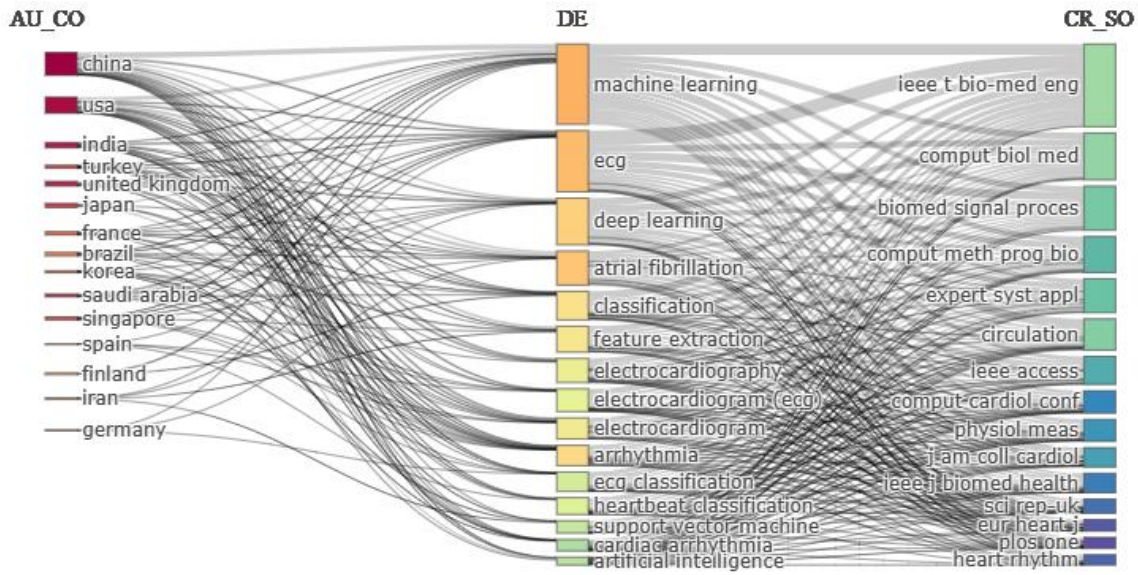


Fig. 2.18: Three fields plot

The three fields plot included in the Bibliometrix tool enables to comprehend the entire bibliometric analysis in a single figure. It represents a significant relationship between the different subjects of analysis for arrhythmia detection and classification. Fig. 2.18 shows the proportionality among the 15 most active countries, frequently used keywords of those countries, and main journal sources for detecting and classifying arrhythmia. The left field of the figure represents the 15 most active countries, the middle field shows the 15 keywords, which get maximum attention of the researchers of those countries, and the main journal sources are included in the right field.

This figure represents a meaningful correlation between the number of publications in the country, the most widely used keywords, and the main journal source. For arrhythmia detection and classification, the largest number of papers are published in China and the USA. After these two countries, India and Turkey have published large number of research articles, respectively, which has been shown in Fig. 2.3 as well. However, from this figure, we can say that “machine learning” seems to be the most important keyword to researchers.

"ECG", "deep learning", "atrial fibrillation" and "classification" are also quite popular among researchers, as represented in Fig. 2.8.

From the right field we can see that "IEEE Transaction on Biomedical Engineering (ieee t bio-med eng.)", "Computers in Biology and Medicine (comput biol med)", "Biomedical Signal Processing and Control (biomed signal proces)" are the main journal sources. They have published a lot of articles on arrhythmia detection and classification.

2.3.2.3 Historical Direct Citation Network

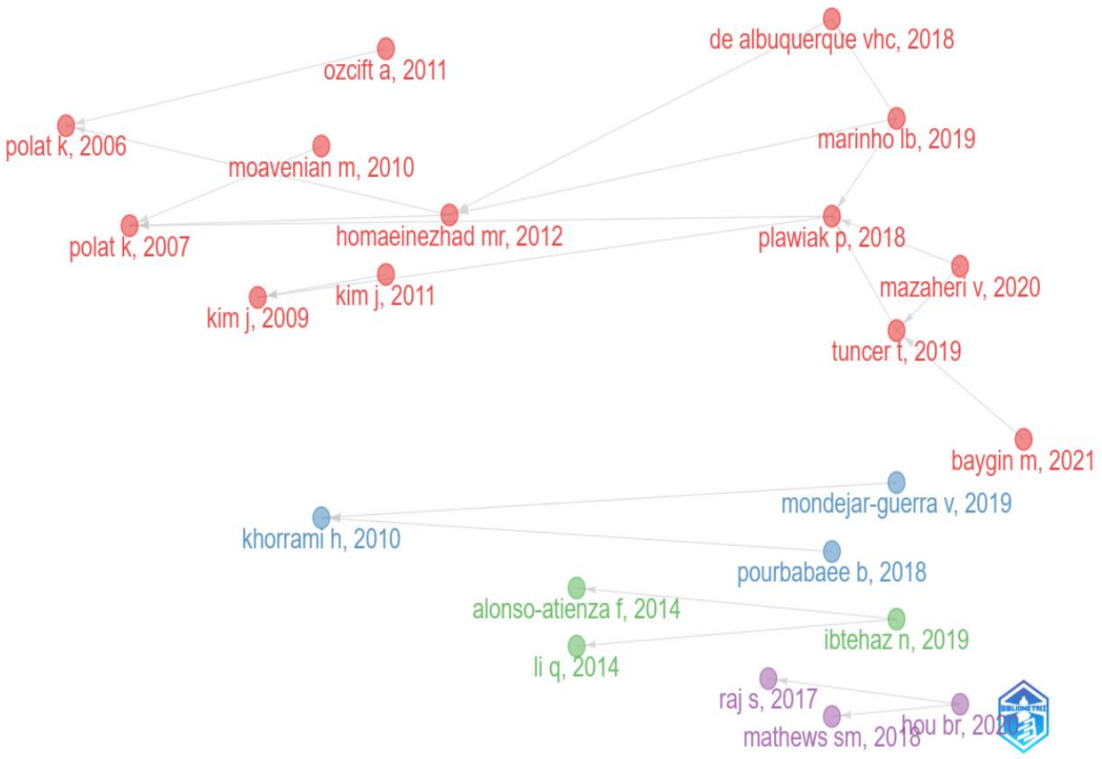


Fig. 2.19: Historical direct citation network

Fig. 2.19 displays the bibliometric historiography for arrhythmia detection and classification. A historical citation network is a graphical representation that visualize the relationships between authors based on the way they are cited by other authors. Each color stream denotes the direct citation with their historical development. They highlight the significance of a particular keynote. The size of the stream depends on the number of cited documents in the same concept. In this network, the nodes represent individual author's

name with the publication year, and the edges between the nodes represent the citations that one author makes to another. The length of the edge between two nodes denotes the time period. According to our figure, the red stream is the largest citation stream, which started with the research of K. J. E. S. A. Polat in 2006 [43] and ended with the research of M. Baygin in 2021[39]. This stream focuses on the arrhythmia classification. The most cited author P. Pławiak [31], who has a total of 12 citations is from the red steam. He proposed a novel methodology named "1 type of normalization x 2 Hamming window widths x 4 types of classifiers' in the field of arrhythmia classification in 2018. The remaining three streams have three nodes each. But among them, the blue stream is bigger than the other two. This was started by the researcher H. Khorrami [44], who has contributed to the identification of ECG arrhythmias using discrete wavelet transform (DWT), continuous wavelet transforms (CWT), and discrete cosine transform (DCT) transformations, and ended with V. M. Guerra's [45] study on heartbeat categorization. The smallest stream is the purple stream. The time period of this stream was from 2017 to 2020. This stream is based on ECG analysis for arrhythmia classification. Researcher S. Raj [35] has worked on ECG signal analysis through PSO (Particle Swarm Optimization) optimized SVM (Support Vector Machines) and DCT based DOST (Discrete Orthogonal Stockwell Transform). In 2018, S. M. Mathews [46] worked on ECG classification. Both of these publications were referenced by the author B. Hou [47] in his paper in 2020. According to this graph, some researchers are working on detecting arrhythmia, while others are working on classifying arrhythmia for a particular time period, and they are mentioning the relevant papers that they used for their research.

2.4 Discussions and Conclusions

Arrhythmia is a serious condition that can result in a number of complications, including stroke, heart failure and even death. Accurate detection and prediction of arrhythmia can help to determine the prognosis of an individual, which can help to reduce the risk of complications and improve outcomes. A lot of research work has been done so far to accurately predict and classify arrhythmia [31], [24], [23]. This paper describes the bibliometric analysis of arrhythmia detection and classification. Two different bibliometric methods named "performance analysis" and "science mapping" have been used in this paper.

Three issues are given the highest priority in this research for performance analysis. The first of the three issues are leading countries, authors, affiliation and sources. The result of performance analysis provides us those countries, authors, affiliations and journals who have made significant contributions for detecting and classifying arrhythmia. By analyzing 238 papers on arrhythmia detection and classification from 2005 to 2022, we can see that the contribution of China beats all of the other countries in the world. After China, the contribution of the USA and India are noteworthy. The author named U. R. Acharya [35], [31], [36], [37], [30] from India is the most relevant author over the past 17 years with 5 papers in the research field of arrhythmia detection and classification. When it comes to the contribution of individual institution, the University of Pennsylvania in the USA has done and published the maximum number of arrhythmia detection and classification-related research. The journal named “Computer in Biology and Medicine” has published more articles on arrhythmia detection and classification than any other journals and the number of published articles of this journal is 11.

The second major aspect of performance analysis is the trend analysis of this paper. It gives us the knowledge about different topics and keywords, which have drawn the attention of researchers in contemporary times. For detecting and predicting arrhythmias, "machine learning" has been the mostly used keyword in this technological era for the past 17 years. However, the other popular keywords are “ECG”, “atrial fibrillation” and “deep learning”. According to recent statistics, "deep learning approach,” “heart beat classification," and "arrhythmia detection" are now in trend, and researchers are paying more attention to those topics.

Our final performance analysis topic is citation analysis. In this paper citation analysis make us aware about the most cited countries, authors and sources in the field of arrhythmia detection and detection. The highly cited countries are USA, China and India respectively which means that many authors have cited the papers that were published in these countries. The author named P. Plawiak [31] has received the highest number of citations for his significant contribution to the field of arrhythmia detection and prediction. The most cited journal entitled “IEEE Transaction on Biomedical Engineering” has 513 citations.

Later we have used science mapping in to visualize and analyze the relationships and connections more thoroughly between different authors, countries, institutions, and journals. In order to explain science mapping, we have focused more on co-citation network of journals, collaboration network of institutions, collaboration network of countries and collaboration map for different continents and subcontinents. By creating a collaboration network for 20 countries, we can see that there are four unequal groups, and each group member is working in collaboration with each other to achieve accurate results in arrhythmia detection and classification. China and USA have developed the best collaboration network among themselves. Similarly, 28 institutions of different countries are continuing their research by organizing themselves into five clusters. The Turkish university, Firat University has established collaborative relationships with five other institutions, which makes it the most cooperative institution. The journals "IEEE Transaction on Biomedical Engineering," "Computers in Biology and Medicine," "Computer Methods and Programs in Biomedicine," and "Biomedical Signal Processing and Control" have a significantly close connection among themselves. The collaboration map for different continents provides scenario that countries from different continents, such as China from the Asian continent and the USA from the North American continent, are working in collaboration to enrich the research field of arrhythmia detection and classification. We have taken the help of thematic evaluation, a three-field plot, and a historical direct citation network to provide a comprehensive overview in the field of arrhythmia detection and classification. From 2005 to 2015, topics such as feature selection, wavelet transform, and prediction got the highest priority in the detection and classification of arrhythmia. However, in the following three years, "Classification" became the most popular topic, followed by "management" and "prediction" and “classification” has managed to retain its popularity in the last three years as well. This outcome gives us information on the emerging trends in the detection and prediction of arrhythmias. The three fields plot and historical direct citation network provide a summary of previous analyses.

This study gives the researchers, authors, and others a complete overview of different aspects of research for arrhythmia detection and classification by using a bibliometric method. Furthermore, this study has highlighted both new trends and past status, which

will make research easier for new researchers. This paper recognizes the significant authors, institutions, countries, journals, who are working day and night for better accuracy and improvement of arrhythmia detection and classification.

Chapter 3
Materials and Methods

We have discussed about the materials and methods of the thesis in this section. Arrhythmia is a life-threatening disease. From the statistics discussed in Chapter 1, it can be said that the number of deaths due to the occurrence of arrhythmia is enormous. The death tolls can be reduced if the occurrence can be classified and predicted earlier. Classification is the process to assign objects to undefined classes such that elements in the class are linked in some way [48]. It can be done via statistical analysis, machine learning, deep learning, data visualization, and other methods. In this thesis, we have employed two different approaches to predict AMI. These two approaches are the conventional statistical analysis and machine learning (ML) analysis. The Fig. 3.1 summarizes the methodology that we have used in this thesis to predict arrhythmia.

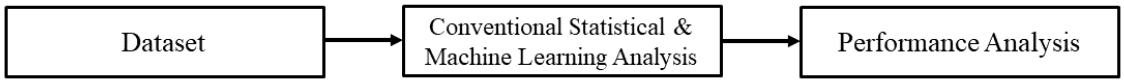


Fig. 3.1: Proposed methodology

3.1 Dataset

We have collected the dataset from the First Affiliated Hospital of Harbin Medical University [24]. The study covered individuals, who experienced an acute myocardial infarction (AMI) and received medical treatment in the hospital's cardiac care unit between January 2014 and January 2019. Each patient has had coronary angiography, 24-hour Holter monitoring, and three-dimensional echocardiography. The presence or absence of tachyarrhythmia was determined based on the result of the episodes [28]. The GitHub repository contains two excel files. The original dataset, named ‘Raw data.xlsx’ and the other, labeled as "variable conversion.xlsx" is a representation of the original variable conversion.

The dataset is comprised of 2086 AMI patients’ data and has a total of 45 features and the target label is arrhythmia. It is binary classification problem. The risk factors for

tachyarrhythmia following AMI identified in the prior study were chosen, and other risk factors, such as demographics, admission baseline features, laboratory characteristics, echocardiographic parameters, and angiography, were added as potential variables. All the features are categorized in Table 3.1. All these data were gathered during hospitalization and before to percutaneous coronary intervention (PCI). The 24-h Holter record recorded data both before and after PCI as some of the patients received emergency PCI [28].

Table 3.1: Feature characteristics

Category	**Features**
Demographics & Medical History	Age, Sex, Smoker, Drinker, Pre-hypertension, Pre-diabetes Mellitus, Prior MI, Prior CI, Prior HF, Prior CHD
Baseline Characteristics of Admission	SBP, DBP, HR, Killip, NYHA
Laboratory Characteristics	Pro-BNP, CRP, Total cholesterol, Triglyceride, HDL, LDL, Cr, K+, TNI, CK-MB, UGLU, DDP
Findings on ECG	P-R, QTc, BBB (LAFB, LPFB, LBBB, RBBB)
Echocardiographic Parameters	LVEF, FS, E/A, Dt, LVEDD, IVST, LVPWT, LA, RA (Up and Down), RA (Right and Left), PA, Vpa, Vao, VWMA
Angiographic Characteristics	LAD, LCX, RCA, LM, LAD + LCX, LAD + RCA, RCA + LCX, Triple Vessels

- VWMA → Ventricular Wall Motion Abnormal

The Table 3.2 represents the abbreviations of the acronyms that has been used in the ‘Raw Dataset.xlsx’ file. We have used these acronyms for easier functionality in the codes and further processing of the dataset.

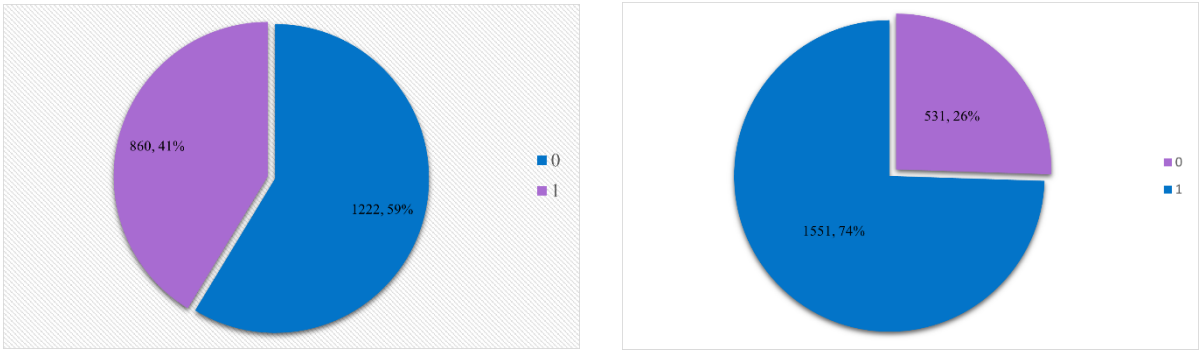
Table 3.2: Acronyms of the features

Features	Acronyms	Features	Acronyms
Age	-	Right Atrium Up and Down Diameter	RA (up and down)
Systolic Pressure	SBP	Right Atrium Right and Left Diameter	RA (right and left)
Diastolic Pressure	DBP	Pulmonary Artery Internal Dimension	PA
Heart Beats	-	Pulmonary Peak Flow Rate	Vpa
Pro-B-type Natriuretic Peptide	Pro-BNP	Peak Aortic Velocity	Vao
C-reactive Protein	CRP	PR Interval	P-R
Total Cholesterol	TC	QTc Interval	Q-Tc
Triglyceride	TG	Urine Glucose	UGLU
High-density Lipoprotein Cholesterol	HDL	Pre-hypertension	-
Low-density Lipoprotein Cholesterol	LDL	Pre-diabetes Mellitus	-
Creatinine	Cr	Smoker	-
k+		Drinker	-
Troponin I	TNI	Prior Myocardial Infarction	Prior MI
Creatine Kinase Isoenzyme	CK-MB	Prior Cerebral Infarction	Prior CI
D Dimer	DD-P	Prior Heart Failure	Prior HF

Features	Acronyms	Features	Acronyms
Left Ventricular Ejection Fraction	LVEF	Prior Coronary Heart Disease	Prior CHD
Fraction Shortening	FS	Killip	-
Mitral Valve Peak Velocity Early Diastolic Filling (E Wave) to Peak Velocity of Late Diastolic Filling (A Wave) Ratio	E/A	New York Heart Association	NYHA
E Deceleration Time	dt	Sex	-
Left Ventricular End-diastolic Diameter	LVEDD	Ventricular Wall Motion Abnormal	-

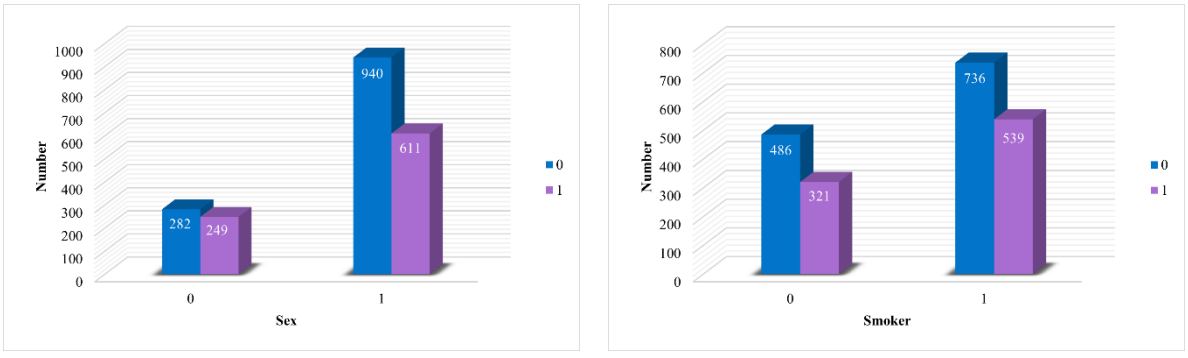
Features	Acronyms	Features	Acronyms
Interventricular Septum Thickness	IVST	Percutaneous Coronary Intervention	PCI
Left Ventricular Posterior Wall Thickness	LVPWT	Bundle-branch-block	BBB
Left Atrium Diameter	LA	Arrhythmia	-

It is a binary classification dataset where our goal is to classify arrhythmia. This feature has two classes, yes (1) and no (0) interpreting whether arrhythmia occurred or not. Fig. 3.2 (a) is the illustration of arrhythmia classes. We can see that the dataset has 1222 (59%) cases in of total 2082 where the subjects did not suffer from arrhythmia and 860 (41%) cases where they suffered from it. Fig. 3.2 (b) visualizes data on the sex of the subjects. 74% (1551) of total subjects are male population annotated by class 1 and the remaining 26% (531) population is female marked by class 0. We have investigated the occurrence of arrhythmia based on sex and smoking. The statistics of the investigation have been visualized in Fig. 3.3.



(a) (b)

Fig. 3.2: Representation of (a) arrhythmia classes (b) sex of patients



(a) (b)

Fig. 3.3: Occurrences of arrhythmia based on (a) sex (b) smoker

Fig. 3.3 (a) represents the occurrence of arrhythmia based on the sex of the subjects. We can see from the figure that, 249 female subjects of the total female population suffered from arrhythmia whereas 611 male patients of the total male population suffered from it. Fig. 3.3 (b) represents the occurrence of arrhythmia based on smoking. The dataset has data from 807 non-smoker patients and 1,275 smoker patients. 0 and 1 indicate non-smoker and smoker respectively. From the figure, it can be seen that 321 non-smoker patients suffered from arrhythmia, and 539 smoker patients suffered from it. The 0 and 1 in the legend indicate the cases of arrhythmia and they contain usual meaning.

3.2 Conventional Statistical Analysis

In this section, we have discussed statistical analysis method to find out the important features, which are mainly responsible for causing myocardial arrhythmia. Statistical analysis is a technique that makes statistical judgments about a quantitative feature of a population [49]. With the use of statistical analysis, we are able to reach a probabilistic conclusion from the information we have. Different types of statistical tests give different assumptions and decisions [50]. Combining the results of different tests and their observed patterns, we can come to a decision for our dataset, which is a very important part of the research.

Statistical analysis investigates the relationship between two features. For our dataset, statistical analysis allows us to determine whether there is a relationship between arrhythmia and other features and, if so, what kind of relationship it is. It enables us to test the hypotheses, such as whether there is a significant difference between two variables or whether one variable is related to arrhythmia. Statistical analysis allows us to gain a better understanding of those features that causes arrhythmia so that we can increase the performance of myocardial arrhythmia predictions and classification. For all kind of analysis, we have used SPSS, winch is a statistical software. This user free software provides advance statistical analysis option. Fig. 3.9 provides an overview of statistical analysis process.

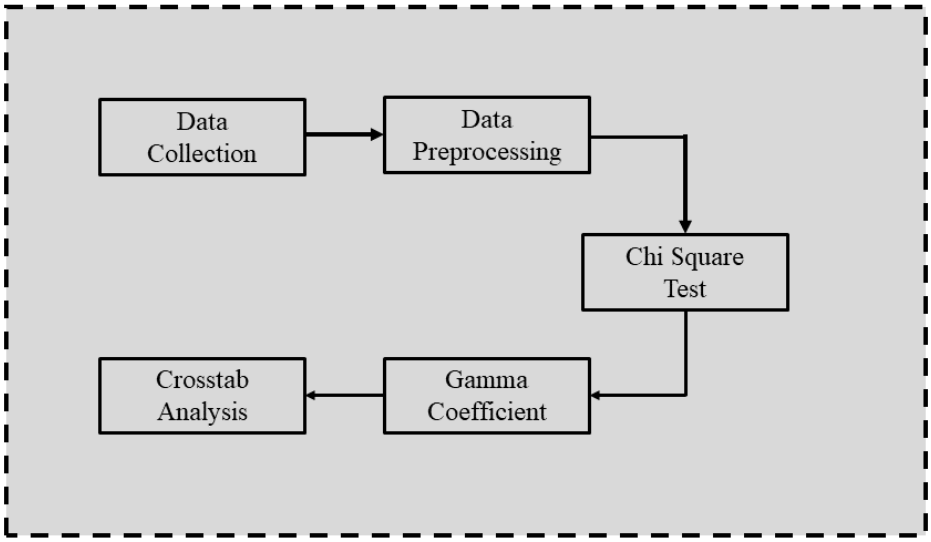


Fig. 3.4: Proposed methodology for conventional statistical analysis

3.2.1 Data Preprocessing

Data preprocessing is an initial stage in the statistical analysis approach that converts the raw data into a suitable format for further analysis. It improves the accuracy and reliability of the analysis. In the context of statistical analysis, data preprocessing involves several steps, such as data cleaning, data conversion and encoding, data integration, data normalization and noise identification [51]. But in this study, we have used only data cleaning and categorical conversion for data preprocessing. Except these two steps, other steps were not necessary to follow, as we have conducted our research with a secondary dataset.

3.2.1.1 Data Cleaning

Real-world data may have a variety of problems, including missing data, outliers and repetitions. Data cleaning refers to the method of finding and fixing errors and inconsistencies in data. Errors are sensitive to many analyses and can cause completely wrong conclusion. So, data cleaning is very important to obtain accurate and reliable result. This process involves finding outliers and resolving them, replacing missing values and removing duplicates.

Using ‘boxplot’ function we have identified outliers from our dataset. This function graphically shows the value of outliers. In order to fix these outliers, we have used ‘go to

case' option and change the value according to original dataset. In our dataset, there were some missing values. Missing data can cause the results of statistical analysis to degrade in accuracy. One common approach to dealing with missing data is to replace them with the series mean or median of the non-missing data in the same variable. In this study, we have used series mean to replace missing value. This process is called "imputation" [52].

3.2.1.2 Categorical Conversion and Encoding

In this preprocessing step, we have converted our continuous data into different categories. This process is known as categorical conversion. The raw dataset that obtained from the patient was in continuous form. Data in this form is not suitable for analysis. So, in accordance with the excel file "variable conversion.xlsx" in the GitHub repository, we have converted them into categorical form. For the convenience of our analysis, we have performed label encoding after categorical conversion. Label encoding is a method where each individual category is transformed into an integer number [53]. Label encoding is very important for ordinal data because each integer value carries a special meaning. As we have converted our scaled feature (continuous) into ordinal feature (categorical), label encoding is a prerequisite condition for statistical analysis. In the following section, we have performed relationship analysis and non-parametric test with this modified dataset.

3.2.2 Non-Parametric Test

Standard statistical tests are typically used to confirm a hypothesis [54]. Statistical tests come in a variety of forms based on the research questions and the research purpose. According to the data, both parametric and nonparametric tests are commonly used in medical research. The main parameters for recommending a parametric test or a non-parametric test are the normal distribution and skewed distribution of data [55]. As a general guideline, a parametric test is used when our available dataset has a normal distribution [56]. On the other hand, when our available dataset follows a skewed distribution, a non-parametric test is applied. The dataset which we use in our research follows skewed distribution rather than a normal distribution. As a result, we have chosen nonparametric test for our statistical analysis. In this study, we have performed different non-parametric tests (Chi-square test and gamma analysis) to confirm the association of different independent features with arrhythmia.

3.2.2.1 Chi Square Test

We have performed 'Chi squared test for independence' as a nonparametric test. Chi squared test is a hypothesis test and used to determine how closely the observed distribution matches the expected distribution [57]. It is a means of finding out whether there is any dependency between two categorical variables. That is, two variables are checked whether they are independent of each other through Chi squared test for independence. Chi square test follows a general formula. We can write the formula as:

$$X^2 = \sum \frac{(u - v)^2}{v} \tag{3.1}$$

where, X^2 denotes the Chi square value, u represents the observed value, and v indicates expected value.

In general, the parameter we use for the Chi squared test is referred to as the null hypothesis [57]. The Chi square test's null hypothesis states that there is no association or dependency between two categorical features [58]; they are independent of each other. The minimum value for the null hypothesis to be accepted is 0.05 [59]. If the value of the null hypothesis is less than 0.05, then the null hypothesis will be rejected and the alternative hypothesis will be accepted. This means that there is a dependency or association between two features. This association can be weak, moderate or strong.

In this study, we have performed the Chi square test to see if there is any dependency between each of the independent features and the only dependent feature, named arrhythmia. This result will help us to check the strength of association later.

3.2.2.2 Gamma Coefficient

The strength of correlation between two ordinal variables is measured symmetrically by the gamma coefficient and the value of the gamma coefficient varies from -1 to 1 [60]. It is used to measure the monotonic relationship between two features [61]. The strength of the association depends on the value of gamma and this value of gamma can be positive or negative. The positive gamma value indicates that there is a proportional association between two ordinal features and the negative value implies inverse association between two features. The value ranges from 0.01 to 0.29 represents for weak association, range

from 0.3 to 0.49 defines the moderate association, a range from 0.50 to 0.69 represents strong association, and lastly gamma value > 0.70 signifies the strongest association. A value of 1.00 indicates a perfect association [62].

In this study, for determining the association strength of those feature on arrhythmia we have performed the gamma coefficient test.

3.2.3 Association Analysis

Association analysis is a useful method for summarizing and analyzing the association between two or more categorical features. It helps us to know the distribution of one feature based on the category of another feature. There are several types of association analysis, such as regression analysis, crosstab analysis and association rule mining [63]. In this research, we have used only crosstab analysis to analyze the association between two features.

3.2.3.1 Crosstab Analysis

One of the most used tools for analyzing the association between two categorical features is crosstab analysis. Crosstabs analysis also known as crosstab, is a procedure that generates contingency tables, which represents the frequency of distribution for two different categorical features [64]. Crosstab analysis helps researchers to understand data sets at a deeper level. Crosstab occurs between two categorical variables, where one variable is in a row, and the other is in a column of the table [65]. The cells contain the distribution of occurrences.

In our study, we have performed crosstab analysis of each independent categorical feature with only dependent feature named arrhythmia. This analysis provides information about how arrhythmia and other independent categorical variables are related. From this analysis we have obtained a frequency distribution table that represents the distribution of arrhythmia occurrence in numerical value and percentage for each feature.

3.3 Machine Learning (ML) Analysis

In this section, we have discussed the machine learning approach to predict the occurrence of arrhythmia after AMI. For ML analysis, initially, data preprocessing has been performed in order to obtain an accurate and efficient classification model. To increase the accuracy

and the reliability of the model, we have optimized the hyperparameters of the random forest (RF) classifier. We have trained the model several times with different sets of features before and after the model has been optimized. Finally, we have analyzed the performances of the model which we have discussed briefly in chapter 4. The Fig. 3.5 summarizes the analysis process.

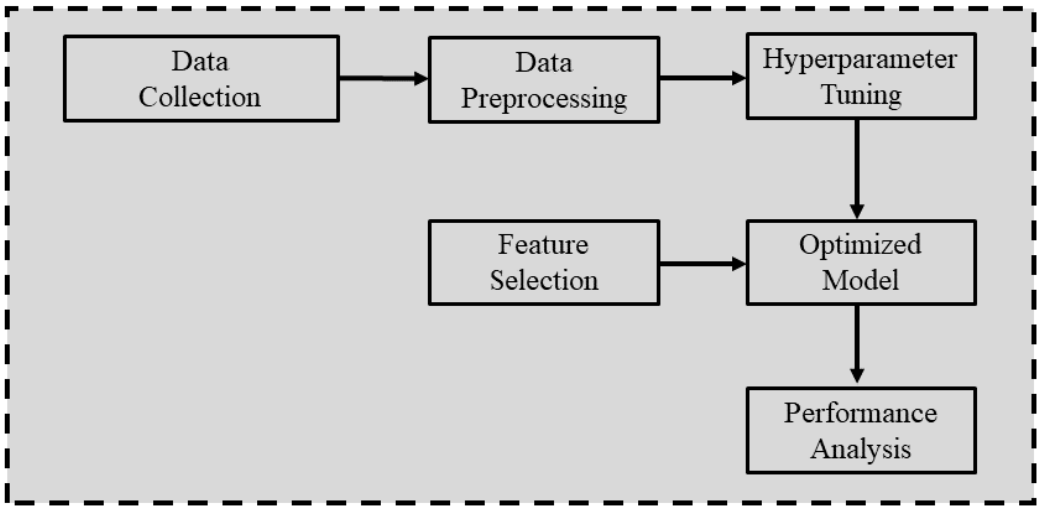


Fig. 3.5: Proposed methodology for machine learning analysis

3.3.1 Data Preprocessing

Data preprocessing can be described as the processing of raw data to transform the dataset for further processing in an easy and efficient format. This step aims to transform real-world data, which is frequently imperfect, unclean, and inconsistent, into a format that can be read by computers. The basic data preprocessing model includes several steps for transforming the dataset such as data integration, data cleaning, data normalization, and data transformation [66]. In this thesis, we only performed data cleaning, data normalization, and data transformation in the data preprocessing step. Data integration is the process of combining all the data from different data sources into one dataset. We worked with a secondary dataset, so we did not have to perform data integration. The steps, data normalization, and data transformation have been performed under the data preprocessing block (Fig. 3.5).

3.3.1.1 Data Cleaning

The real-world data, which is integrated is not free from errors. The integration might lead to an excessive amount of uncleaned data [67]. Unclean data include missing data, wrong data, and non-standard representation [66]. If the data is not clean enough, training machine learning (ML) models will not be accurate and reliable. So, to have accurate results and to obtain a reliable ML model, we must clean the data. Unclean data or dirty data can be classified as missing data and noisy data [66].

We have used the 'isnan' function from the NumPy library to find all the missing values in the dataset. This function returns the result as a Boolean array indicating true for NaN values and false otherwise [68]. NaN stands for not a number, which is the representation of missing values. The function returned the indexes of the missing values. To clean the data, we then used the DataFrame.dropna function from the Pandas library. This function removes missing values from a given index array or columns [69, 70]. Then this cleaned dataset was used for further processing.

3.3.1.2 Data Encoding

Data normalization and data transformation preprocessing steps will be discussed in this sub-section. The raw data, in most cases, is in a human-readable format. So, it is not useful enough for training accurate ML algorithms. Typically, to improve the predictive capacity of the ML algorithms the original dataset is transformed [66]. This transformed data will be easier to read for machines. Casting the data into a certain range, such as between 0 and 1 or between -1 and +1, is the data normalization. When there are significant variations between the ranges of several features, data normalization is necessary. Generally, data normalization concentrates on the transformations that change the distribution of the initial values into a new set of values with the required characteristics, rather than generating new values.

3.3.1.2.1 Min-Max Normalization

One of the well-known data normalization techniques are min-max normalization. This method scales all the numerical values, x of a feature, F into a predetermined range indicated by $[new\_min_F, new\_max_F]$, where min_F, max_F are minimum and maximum

values of F respectively [66]. The scaled values can be obtained by the following expression,

$$x' = \frac{x - min_F}{max_F - min_F}(new\_max_F - new\_min_F) + new\_min_F \quad (3.2)$$

where x' is the new scaled value, new_min_F is the minimum value of scaling range, and new_max_F is the maximum value of scaling range.

The normalization can be done between any specified range. Typically, it is done in the interval of 0 to 1. In this thesis, normalization or scaling will refer to the case where the specified range is [0, 1]. In this case the new_min_F is 0 and new_max_F is 1.

To normalize our dataset, we have used the min-max normalization technique. We have applied the MinMaxScaler method of class sklearn to perform the task [71]. We have a procesed dataset that we will use only for feature selection (FS) as an outcome of this stage.

3.3.1.2.2 Categorical Conversion

A categorical variable is a value that has a restricted and defined set of potential values, allowing a data unit to be classified by assigning it to a wide category [72]. As aforementioned, the raw dataset comprises data in its raw form that was obtained when the patients were hospitalized. This raw form is not suitable for training predictive ML algorithms. As a result, in order to create robust ML models, we turned the raw information into categorical values. This conversion was carried out in accordance with the excel file "variable conversion.xlsx" in the GitHub repository. For each feature, we built a function to convert raw data into categorical ones. The function defines the range of each category value for a given feature. After defining the function, pandas.DataFrame.apply [70, 73] function has been used to apply the function for a feature of the dataset. This processed dataset has been used to train our ML models.

3.3.2 Hyper Parameter Tuning

ML models generally have two types of parameters such as model parameters and hyper parameters (HP). Model parameters are initialized and updated through the learning process. Algorithms update these model parameters based on the data provided for learning

[73, 74]. On the other hand, HP are higher level parameters that must be set manually before training the model based on the properties of the data because these parameters cannot be estimated from the data [73, 74]. HP are used to optimize model for better performance or used to minimize the loss function [75]. A wide range of alternatives must be investigated in order to develop an optimum ML model. HP tuning refers to the process of constructing the best model architecture with an optimal HP configuration [76]. HP tuning is considered as a major component in developing an efficient ML model, particularly for tree-based ML models and deep neural networks, which have several hyper-parameters [77]. Manual tuning process is very time consuming and lengthy process and other factors such as large number of HP, complex models make the task more difficult. HP optimization (HPO) techniques are very handy tools to solve these difficulties. From a wide range of available HPO techniques, it is an important task to find a suitable optimization method for the model. Optimization techniques such as decision-theoretic approaches, Bayesian optimization models, multi-fidelity optimization techniques, and metaheuristics algorithms are more suitable for HPO than traditional optimized methods [78]. In this thesis we have used cuckoo search algorithm (CSA), a metaheuristics algorithm base approach to perform the HPO task. The CSA and metaheuristics algorithms have been discussed briefly in section 3.3.3.1.

3.3.3 Feature Selection

The objective of ML algorithms is prediction or classification by harvesting data. To make the algorithms more accurate, a large amount of data is provided so that machines can learn better. The size of the data is increasing day by day. This increase in data size has an impact on the computational cost and prediction accuracy of ML algorithms [79]. As the size of the data is increasing, the variables of the dataset are also increasing. The variables, which are used as input for training ML models are called features. Each column of the dataset is a feature. In prediction or classification problems, ML algorithms are used to predict or classify target feature(s). If the data is big, it will have a lot of features, that may not always be important to train an efficient and optimal model. It means, the ML model will learn unnecessary patterns and information. This unnecessary information is noise. If we put noise as input to the model, the output can also become noisy. This is where feature selection (FS) comes into the picture. FS is a widely used approach in data preprocessing,

and it has become an essential component of the ML process [80]. It is the process of identifying and reducing unnecessary, duplicated, or noisy data. It helps us to shrink the size of the data size. The reduction is data size accelerates data mining (DM) and ML algorithms, improves learning accuracy, and enhances comprehension. The objective of FS is to identify a subset of features that will maximize prediction accuracy or minimize the size of the structure without significantly reducing the prediction accuracy of the classifier with only the selected features [81]. FS methods can be categorized as filter methods, wrapper methods, and embedded methods. Filter methods are independent of ML algorithms and are focused on the performance of features in several statistical tests. A correlation test is performed to see whether the features are positively or negatively correlated to the output. Information gain, Chi-square test, etc. are examples of filter methods. Wrapper methods use a subset of features and train learning algorithms to test the performance of each subset. Based on the performance of the algorithms, features are added or subtracted and evaluated again. The subsets are formed using a greedy approach and all possible combinations of features are evaluated. Embedded methods combine both filter and wrapper methods to deal with search and classification at the same time [82, 83]. Finding out the best subset of feature is a NP-Hard problem [84, 85]. Metaheuristic algorithms are one of the best approaches to solve these types of problems [83, 86, 87]. It is a search procedure, designed to find the best solution. Metaheuristics can find good results with less computational cost than conventional optimization algorithms [88]. Recently metaheuristic algorithms are widely used for FS [89-91]. In this thesis, we will use a well-known swarm-based metaheuristic algorithm, CSA for FS.

3.3.3.1 CSA

CSA is a well-known metaheuristic algorithm for optimization problems [83]. CSA was first introduced in 2009 and this algorithm is based on the brood parasitism of some species of cuckoos [92]. It is a known fact that, some species of cuckoos always lay eggs in the nest of host birds. Generally, these host birds are of different species. When a cuckoo lay egg in host bird nest, there is a possibility that the host bird will discover the alien egg. This possibility of discovery of alien eggs is denoted by P_a. Upon the discovery of egg, two possible scenarios can take place. The host bird can destroy the alien egg or it will

abandon its nest to build a new one. For the sake of simplicity, three idealized rules are assumed,

i. Each cuckoo lays only one egg in a randomly selected nest
ii. Only the best nests with high-quality eggs will pass down to the next generations and
iii. Only the best nests with high-quality eggs will pass down to the next generations $P_a \in [0, 1]$.

When the egg is discovered two possible scenarios can take place which we discussed earlier in this sub-section. For less complexity, some nests are replaced by new ones with a fraction of P_a of the total n nests. Some assumptions are made to simply describe the algorithm easily. Each egg in a nest represents a solution, a cuckoo egg represents a new solution, a high-quality egg refers to the best solution near optimum value and discovered eggs indicate worse solution. The target of the algorithm is to replace a bad solution with potentially better solutions. New solutions are generated via Lévy flight. It is a stochastic equation for a random walk. Generating new solutions $x_i^{(t+1)}$ for cuckoo i, Lévy flight is performed governing by the following equation,

$$x_i^{(t+1)} = x_i^t + \alpha \oplus \text{Lévy}(\lambda) \tag{3.3}$$

where x_i^t is current solution and α is step size.

The random walk of Lévy flight is given by Lévy distribution [92],

$$\text{Lévy} \sim \text{u} = t^{-\lambda} \tag{3.4}$$

The working procedure of the CSA can be described simply by the following steps,

i. Total n nests are randomly initialized in the search space.
ii. Each nest is assigned with a fitness value.
iii. New solutions are generated via Lévy flight.
iv. Locations update followed by Lévy flight.
v. Repeating step 3 and 4 until global best position is found.

In this thesis, we implemented CSA with the help of a python library, Mealpy [93].

3.3.4 Random Forest

In supervised learning technique, random forest (RF) is one of the well-known learning algorithms. The RF is an ensemble learning technique used for classification, regression and other problems. Ensemble learning technique is a process to solve complex problems and enhance the performance of the model by using multiple classifiers. RF uses multiple decision trees to reach a conclusion. In the case of classification tasks, the output is the class selected by most of the trees. The average or mean prediction of the different trees is returned for regression tasks [94, 95]. The RF algorithm was first developed by T. K. Ho [94]. It was developed using random subspace method [95]. Later, the algorithm was developed based on Brieman's bagging idea and random selection of features introduced by Ho. It was trademarked by L. Breiman and A. Cutler in 2006 [96]. RF is an effective algorithm for general-purpose classification and regression problems [97].

Bagging, also known as bootstrap aggregating, was first introduced by L. Brieman [96]. It is an ensemble learning technique used to improve accuracy of ML algorithms. If a training set, D is provided of size n, the bagging method will generate m new training sets D_i of size n'. These new sets will be generated by sampling uniformly from D with replacement. Some data may repeat in each D_i while sampling with replacement. If $n' = n$, D_i is expected to have approximately 63.2% of unique data samples of D for large n [98, 99]. It is known as bootstrap sample. When multiple data samples are generated, the individual tree are then trained independently and by averaging those trained multiple trees, a decision is made by the RF algorithm. Random forests are another sort of bagging method that use a modified tree learning algorithm that picks a random subset of the features at each candidate split in the learning process. It is referred to as 'feature bagging' from time to time. The reason for it is the correlation of the trees in a regular bootstrap sample: if one or a few features are extremely good predictors of the response variable (target output), these features will be picked in many of the B trees, causing them to become correlated. In this study, we have used the RF algorithm to perform classification task on the dataset because it outperforms other classification model such as support vector machine (SVM), classification and regression trees (CART), naïve Bayes (NB), and K-nearest neighbors (KNN) [100].

3.3.5 t-SNE Plot

The t-distributed stochastic neighbor embedding (t-SNE) technique is a statistical approach for displaying high-dimensional data by assigning a position in a two or three-dimensional map to each datapoint. It is based on S. Roweis and G. Hinton's [101] Stochastic Neighbor Embedding, to which L. Maaten offered the t-distributed variation [102]. It is a nonlinear dimensionality reduction approach that is well-suited for embedding high-dimensional data for display in a two or three-dimensional environment. It specifically describes each high-dimensional item by a two- or three-dimensional point in such a way that comparable things are described by adjacent points with high probability and different objects are modeled by distant points.

The t-SNE algorithm is divided into two parts. To begin, t-SNE creates a probability distribution over pairs of high-dimensional objects in which comparable items are allocated a greater probability and dissimilar points are assigned a lower likelihood. Second, t-SNE creates a comparable probability distribution over the points in the low-dimensional map and minimizes the Kullback-Leibler divergence (KL divergence) between the two distributions with regard to the map locations. While the original approach bases its similarity metric on the Euclidean distance between objects, this may be altered as needed. The dataset on which we have performed our work, has 45 features in total. To see the distribution of the dataset, we used t-SNE to project this high dimensional data into lower dimensional space.

3.3.6 Performance Analysis Metrics

To measure the performance of the model we have used accuracy, precision, recall, F score, confusion matrix, and area under the curve (AUC).

Accuracy: The accuracy of ML models is the ratio of true positives (TP) and true negatives (TN) to all the positive and negative observations. In other words, accuracy is the number of correctly classified data over the total amount of data [103].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

Precision: It is the ratio of TP and the total number of positively predicted labels. It is also known as positive predictive value [104].

$$Precision = \frac{TP}{FP + TP} \tag{3.6}$$

Recall: The ratio of all correctly categorized classes to actual classes is known as recall [104].

$$Recall = \frac{TP}{FN + TP} \tag{3.7}$$

F1 score: It is the model score as a function of accuracy and recall. F-score is a machine learning model performance statistic that assigns equal weight to Precision and Recall when calculating accuracy, making it an alternative to accuracy metrics. It is the harmonic mean of precision and recall [105].

$$F1\ Score = \frac{2 \times Racall \times Precision}{Recall + Precision} = \frac{2TP}{2TP + FN + FP} \tag{3.8}$$

Confusion Matrix: A confusion matrix is a representation of classification or prediction outcomes. The number of right and wrong predictions is represented with count values and split by class. It provides information not only on the faults produced by the classifier, but also about the categories of errors made [106].

ROC: A receiver operating characteristic (ROC) curve is a graph that depicts the performance of a classifier over all classification thresholds. It describes the capacity to differentiate between distinct classes. The AUC is the area that is covered by the ROC curve. The greater a ROC curve's AUC, the better the model can predict different classes [107].

Chapter 4

Result and Discussion

In this chapter we have discussed about the outcomes of our study. We have used two different approaches, conventional statistical machine learning analysis in our study. The dataset that we used has 2086 entries. In the preprocessing stage, we discarded entries which had null and missing values and converted the instances of each entry into categorical variables. After preprocessing, the obtained processed dataset has 2082 subjects and 43 features. We performed both analysis processes on this processed dataset.

4.1 Performance of Conventional Statistical Analysis

In this thesis, we have used three different statistical tests to find out the important feature that have great significance in arrhythmia prediction. For this purpose, initially, we have used Chi-square test to determine the association or dependency between arrhythmia and other features. In the next stage, to identify the strength of association between arrhythmia and arrhythmia related features, we have applied gamma test. Using first two tests, we have obtained our important features for arrhythmia classification. Finally, to observer the categorical relation between arrhythmia and important features we have conducted crosstab analysis.

4.1.1 Chi-Square Test

As mentioned previously, we have used the Chi-squared test to determine, which features are related to arrhythmia. The result of the Chi-square test is shown in Table 4.1.

Table 4.1: Outcomes of chi-square test

Features Name	Chi-Square Test (P value)	Association
PCI	0.000	Yes
Heart Beats	0.000	Yes
VWMA	0.000	Yes
P-R	0.000	Yes
DBP	0.000	Yes
Age	0.000	Yes
BBB	0.000	Yes
DD-P	0.000	Yes
RA (right and left)	0.000	Yes
SBP	0.000	Yes
RA (up and down)	0.001	Yes
sex	0.002	Yes
TNI	0.013	Yes

Features Name	Chi-Square Test (P value)	Features Name
UGLU	0.018	Yes
TG	0.021	Yes
VPA	0.023	Yes
CK-MB	0.028	Yes
Pro BNP	0.030	Yes
LA	0.039	Yes
Prior MI	0.049	Yes
FS	0.053	No
LDL	0.057	No
Prior CHD	0.062	No
Cr	0.065	No
EA	0.074	No
k+	0.090	No
PA	0.113	No

Features Name	Chi-Square Test (P Value)	Features Name
Prior HF	0.230	No
Pre-diabetes Mellitus	0.241	No
Vao	0.242	No
smoker	0.259	No
LVPWT	0.296	No
TC	0.568	No
drinker	0.580	No
LVEDD	0.608	No
Pre-hypertension	0.712	No
dt	0.758	No
CRP	0.847	No
IVST	0.875	No
HDL	0.920	No

- VWMA → Ventricular Wall Motion Abnormal

The null hypothesis of the Chi-square test for independence is that there is no dependency between two variables, and the alternative hypothesis states that there is a dependency between two variables. The alternative hypothesis will be accepted, and the null hypothesis will be rejected when the estimated p value from the chi-square test is less than 0.05. Table 4.1 shows that the alternative hypothesis is excepted for 21 variables, as the value of p for these variables is less than 0.05. This result indicates, arrhythmia has dependency over 21 variables among 43 variables. Fig. 4.1 represents the name of these 21 associated features. According to this Fig. 4.1 and Table 4.1, p is estimated to have a value of 0 for the following 10 features: PCI, Heart Beats, Ventricular Wall Motion Abnormal, P-R, DBP, Age, BBB, DD-P, RA (right and left), and SBP. From this analysis, we can assume that these features are closely related to arrhythmia. Further, their strength of association with arrhythmia will help us to determine the most important feature for the prediction of arrhythmia.

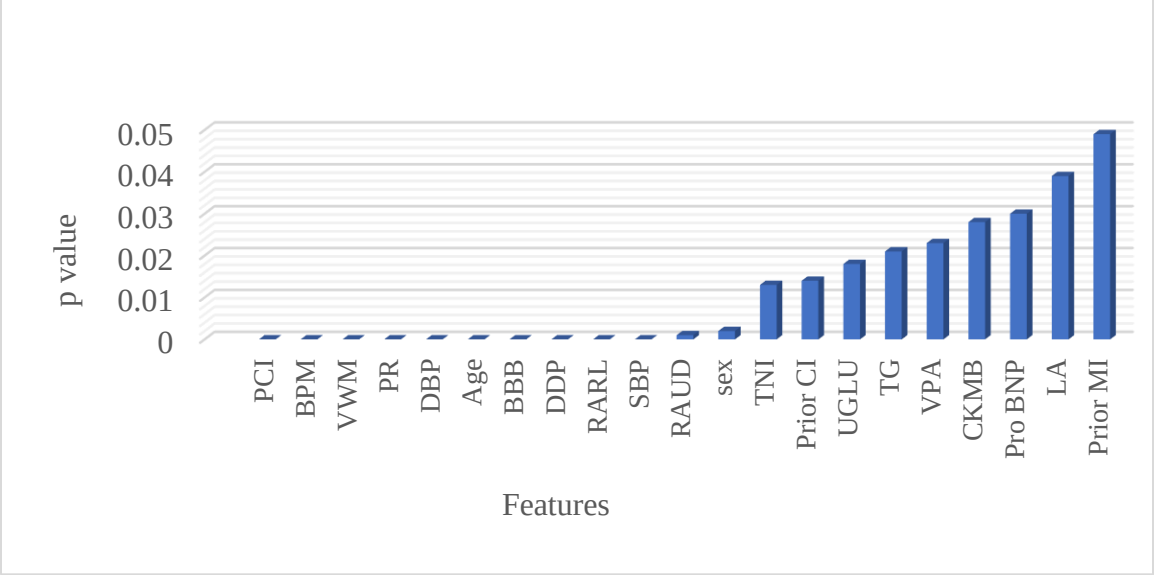


Fig. 4.1: Features related to arrhythmia

4.1.2 Gamma Test

In this study, we have used the gamma coefficient to observe the strength of association between each of the features and arrhythmia. Table 4.2 shows the outcomes of the gamma test.

Table 4.2: Outcomes of gamma coefficient

Features Name	Value of Gamma Coefficient	Strength of Association
DBP	-0.175	Negligible
sex	-0.152	Negligible
SBP	-0.143	Negligible
Heart Beats	-0.142	Negligible
k+	-0.123	Negligible
Vpa	-0.110	Negligible
TG	-0.107	Negligible
FS	-0.086	Negligible
LVEF	-0.076	Negligible
Vao	-0.072	Negligible
LDL	-0.031	Negligible
drinker	-0.030	Negligible
TC	-0.030	Negligible
dt	-0.022	Negligible
IVST	-0.018	Negligible
CRP	-0.009	Negligible

Features Name	Value of Gamma Coefficient	Strength of Association
HDL	-0.004	Negligible
Pre-hypertension	0.016	Negligible
LVEDD	0.042	Negligible
smoker	0.052	Negligible
Pre-diabetes Mellitus	0.061	Negligible
VWMA	0.061	Negligible
UGLU	0.083	Negligible
EA	0.084	Negligible
Q-Tc	0.086	Negligible
PA	0.091	Negligible
Pro-BNP	0.093	Negligible
CK-MB	0.101	Negligible
TNI	0.104	Negligible
Cr	0.105	Negligible
LA	0.118	Negligible

Features Name	Value of Gamma Coefficient	Strength of Association
PCI	0.136	Negligible
LVPWT	0.156	Negligible
Prior MI	0.157	Negligible
Prior CI	0.171	Negligible
Prior HF	0.174	Negligible
Prior CHD	0.193	Negligible
Age	0.221	Negligible
DD-P	0.250	Negligible
RA (up and down)	0.317	+ Moderate Association
BBB	0.342	+ Moderate Association
P-R	0.424	+ Moderate Association
RA (right and left)	0.503	+ Strong Association

- VWMA → Ventricular Wall Motion Abnormal

From Table 4.2, we can see that out of 43 features, 39 have a negligible association with arrhythmia. These features are less significant for predicting arrhythmia. Three features named RA (up and down), BBB, and P-R have a positive and moderate association with arrhythmia. The only feature that has strong association in the case of predicting arrhythmia is RA (right and left). These features have a significant association with arrhythmia. According to other literature and clinical guideline, these four features are also primarily responsible for causing arrhythmia. Further, by analyzing the crosstab, we have determined the most important category of these four features that are significantly related to arrhythmia.

4.1.3 Crosstab Analysis

From the preceding analysis, we have seen the strength of the association between arrhythmia and each categorical variable. In this section, we have analyzed their categorical relationship with arrhythmia through crosstab analysis. Even though we have done a crosstab analysis for each categorical feature, we have only represented the relationships between arrhythmia and four features in this section. Each of these described features has a moderate or strong association with arrhythmia. Their relationship is illustrated in the Fig. 4.2. Crosstab analysis for the rest of the features is included in the appendix A in Table A.1.

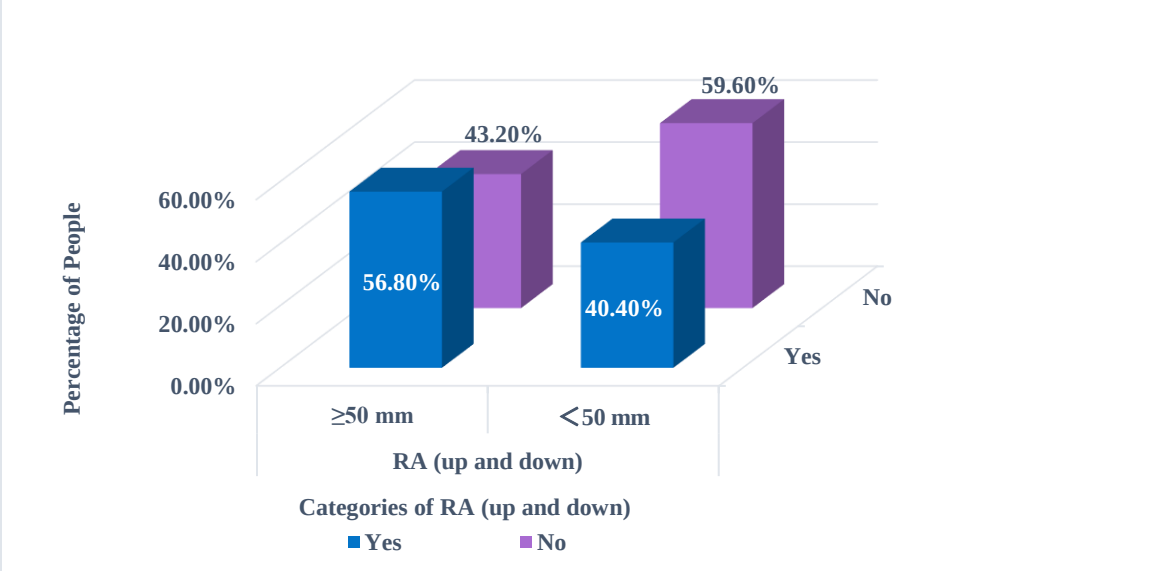


Fig. 4.2: Occurrence of arrhythmia based on RA (up and down)

RA (up and down) has a moderate association with the arrhythmia. Fig. 4.2 represents the distribution of arrhythmia for the different categories of RA (up and down). As outlined in the figure, RA (up and down) has two categories. The first category contains the diameter of RA (up and down) that is greater than or equal to 50 mm, and the second is when it is less than 50 mm. 56.80% of people in the first category have an arrhythmia. So, the first category is an essential predictor of arrhythmia. In the second category, 59.60% of people are not affected with arrhythmia, and 40.40% are affected. From this figure, we assume that though RA (up and down) has a moderate association with arrhythmia, the first category has more significance for predicting this disease.

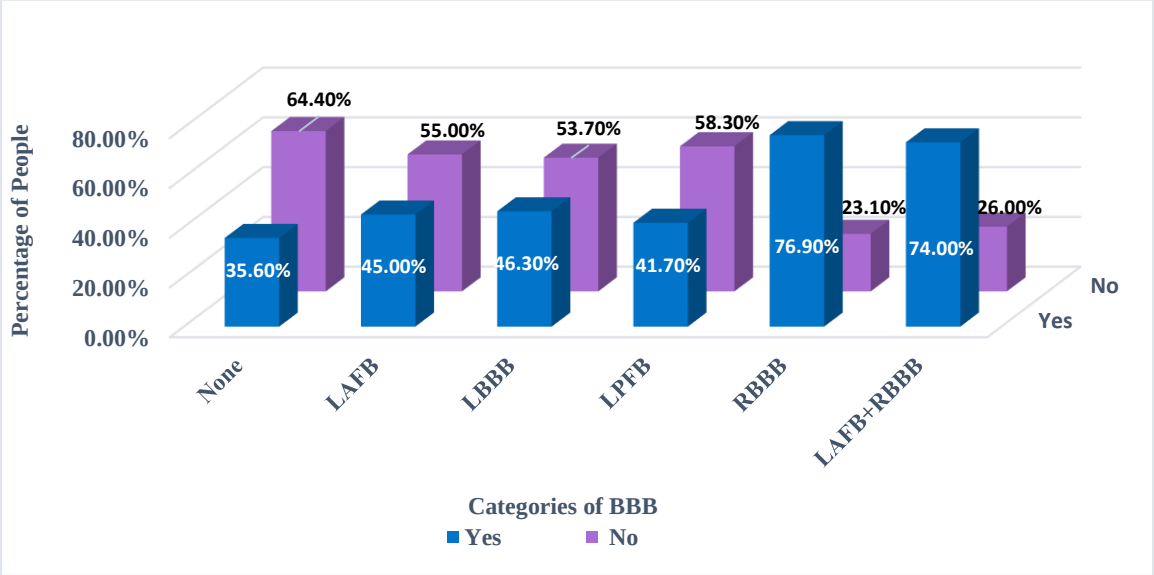


Fig. 4.3: Occurrence of arrhythmia based on BBB

BBB also maintains a moderate relationship with arrhythmia. Fig. 4.3 represents the categorical relationship between BBB and arrhythmia. From Fig. 4.3, we can see that there are six categories for this feature. But not all categories of this feature have an equal relationship with arrhythmia prediction. It is clear from this picture that the importance of right bundle branch block (RBBB) and, combination of left anterior fascicular block and right bundle branch block (LAFB+RBBB) is highest in arrhythmia prediction. For the category named RBBB, 76.90% of people have a relationship with an arrhythmia. Again, 74% of people with LAFB+RBBB have an arrhythmia. So, we can say that RBBB is more related to arrhythmia than any other category of BBB.

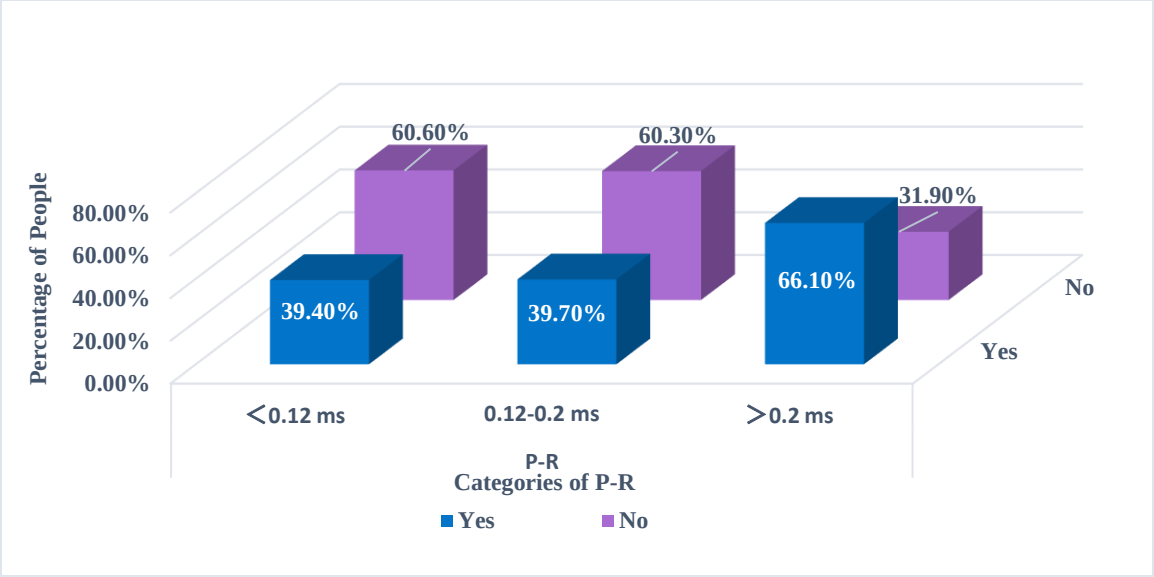


Fig. 4.4: Occurrence of arrhythmia based on P-R

Fig. 4.4 depicts the relationship between P-R and arrhythmia. The association between these two features is moderate. P-R is divided into three different categories. The first category is when the P-R value is less than 0.12 ms; the second category has a value of 0.12 ms to 0.20 ms; and the last category contains a value greater than 0.2 ms. For the first two categories, fewer people have arrhythmias, but in the third category, 66.10% of people have an arrhythmia. So, from this analysis, we conclude that the third category of P-R has a close relationship with arrhythmia.

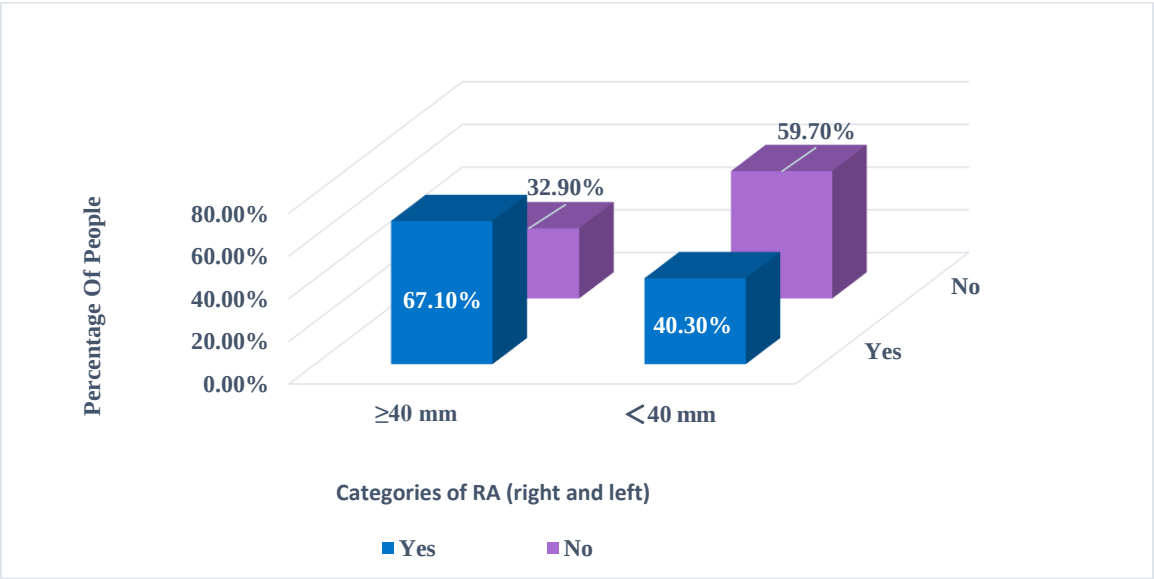


Fig. 4.5: Occurrence of arrhythmia based on RA (right and left)

From the previous analysis, we have found that RA (right and left) has a strong relationship with arrhythmia. In this section, we have analyzed its categorical relationship with arrhythmia. RA (right and left) has two categories. For the first category, the diameter of RA (right and left) is greater than or equal to 40 mm, and this diameter is less than 40 mm for the second category. From Fig. 4.5, we can see that 67.10% of people in the first category have an arrhythmia, whereas, in the second category, only 40.30% of people have this disease. This analysis reveals the first category of RA (right and left) as an essential predictor for arrhythmia.

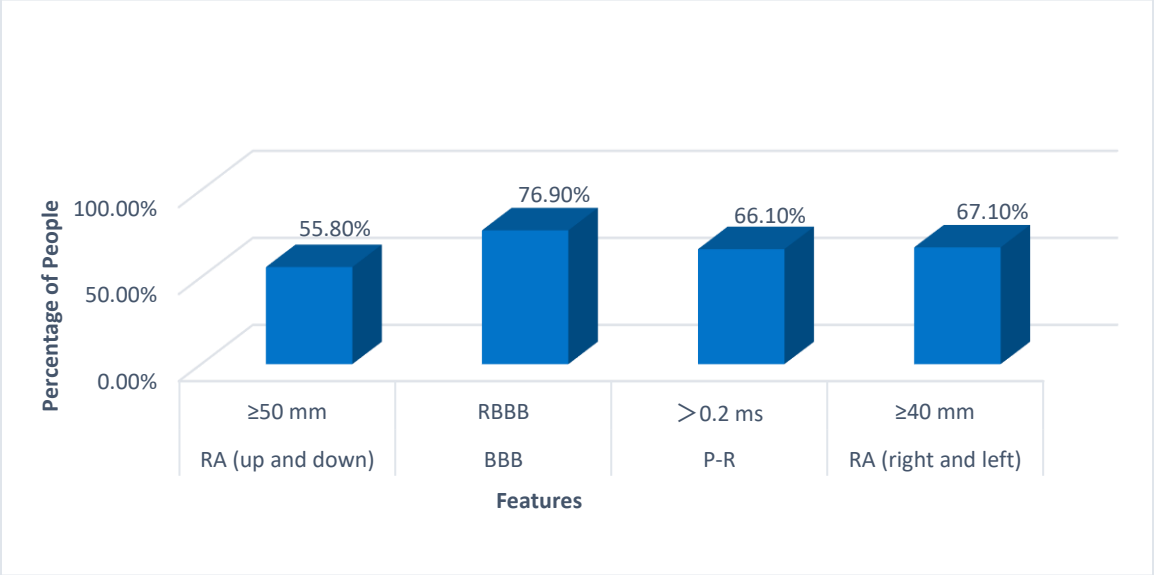


Fig. 4.6: Relationship of important features with arrhythmia

From the previous analysis (Table 4.2), we have found that RA (right and left) has a strong relationship with arrhythmia. In this section, we have analyzed its categorical relationship with arrhythmia. RA (right and left) has two categories. For the first category, the diameter of RA (right and left) is greater than or equal to 40 mm, and this diameter is less than 40 mm in the second category. From Fig. 4.5, we can see that 67.10% of people in the first category have an arrhythmia, whereas, in the second category, only 40.30% of people have this disease. This analysis reveals that the first category of RA (right and left) as an essential predictor for arrhythmia.

4.2 Performance of Machine Learning Analysis

We have trained our RF classifier several times. The Fig. 4.7 summarizes the result analysis process of this thesis. From the figure we can see that we have trained the model with all the 43 features of the processed dataset before and after hyper parameter optimization (HPO). Then after feature selection (FS), we have also trained the optimized model with each set of features. Finally, we have analyzed the results returned from the model each time. We have used stratified k fold cross validation (CV) to train the model 10 times (k = 10) and analyzed the best performed model and average performance of 10-fold.

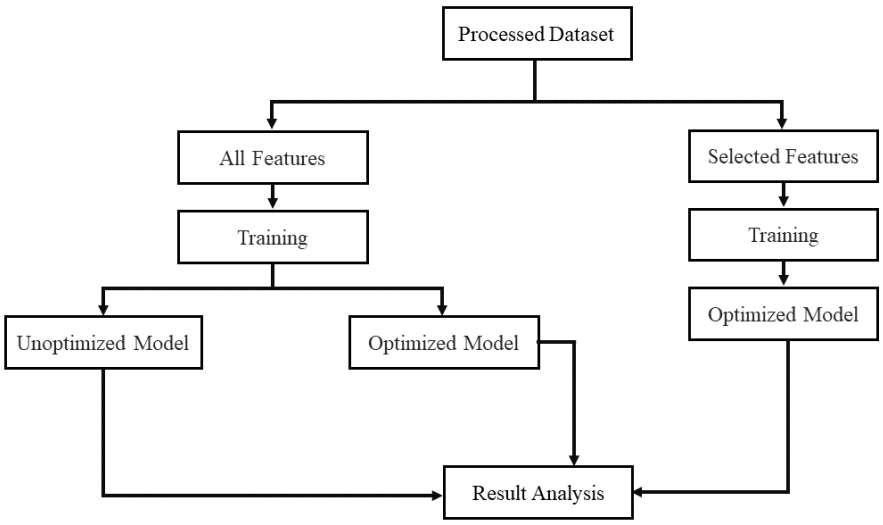


Fig. 4.7: Proposed methodology for ML result analysis

All the results were generated using a PC running Windows 10 Home Single Language having Intel Core i3-7020U CPU processor with clock speed 2.30 GHz and 20 GB of RAM.

4.2.1 Hyper Parameter Tuning Outcome

To increase the accuracy and reliability of the RF classifier we have optimized the values of the HP. Max depth, n estimators, min samples leaf, and max features of the RF classifier has been optimized. As discussed earlier, manual tuning process is very lengthy and time consuming. So, we have used 6 different algorithms to optimize the values of the HP. We have used Grid Search CV (GS), Random Search CV (RS), Bayesian Optimization (BO),

Bayes Search CV (BS), Optuna, and CSA. The Table 4.3 represents the outcomes of these 6 optimization algorithms.

Table 4.3: Outcomes of HPO algorithms

Algorithm	Accuracy	MD	MF	MSL	NE	MS
GS	62.50%	7	8	32	63	-
RS	58.70%	6	25	4	40	-
BO	62.20%	-	0.67	-	198	0.61
BS	63.80%	6	10	30	60	-
Optuna	62.10%	46	sqrt	24	70	-
CSA	69.71%	30	21	13	46	-

- MD → Max Depth
- MF → Max Features
- MSL → Min Samples Leaf
- NE → N Estimators
- MS → Max Sample

Among other algorithms the CSA has returned the optimum set of values of the HP for which the model performance is highest. So, we have used the set of values returned from the CSA and optimized the RF classifier. Along with these optimum values, the parameters random state, n jobs and criterion which have been used to tune the RF model. The values of these parameters are 2, -1 and gini respectively.

The fitness function was evaluated based on the average accuracy of the RF classifier obtained using 10-fold CV. It was a maximizing problem because the maximum accuracy of the model was expected. The global best fitness returned by the CSA was 63.84%. The range of values were different for each hyper parameter. For max depth the upper and lower bound was 1 and 100 respectively. The range differed for n estimators from 1 to 500, for min samples leaf from 1 to 200 and for max features 1 to 43. The search space can be varied more, but we have considered the mentioned range because the computational cost was very high. The parameters of the CSA are ‘epoch’, ‘pop_size’, and ‘p_a’ and their values are 100, 100, 0.25 respectively.

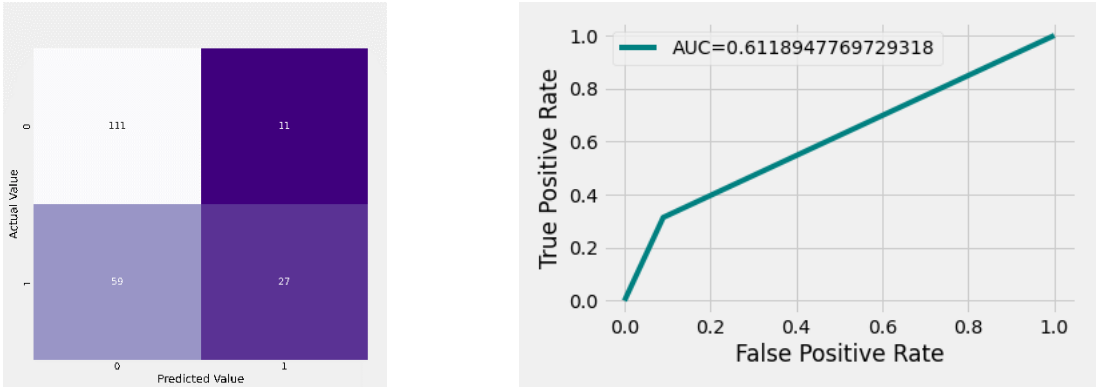
4.2.2 Performance of Unoptimized Model

We have trained the RF classifier before optimizing the model to analyze how the model performs. We have used all the features of the processed dataset to train the unoptimized

model. The average accuracy of the model after 10-fold CV was 62.87% and the best accuracy was 66.34%. The AUC corresponding to the best accuracy was 61.18%. Fig. 4.8 shows the confusion matrix and the ROC curve of the best performed unoptimized model in 10-fold CV respectively. All the performance metrices of the best performed model in 10-fold CV has been showed in Table 4.4.

Table 4.4: Performance metrices of best performed unoptimized model

Metrices	Accuracy	Precision	Recall	F1 Score	AUC
Values	66.34%	67.67%	66.34%	62.59%	61.18%



(a) (b)

Fig. 4.8: (a) Confusion matrix (b) ROC curve of the best performed unoptimized model

- 0 → Arrhythmia class no
- 1 → Arrhythmia class yes

From the Fig. 4.8 (a), it can be seen that the model has predicted successfully 111 cases where the subjects did not suffer from arrhythmia and 27 cases where they suffered from it and from the Fig. 4.8 (b), the AUC of the ROC curve is 61.18%.

4.2.3 Performance of Optimized Model

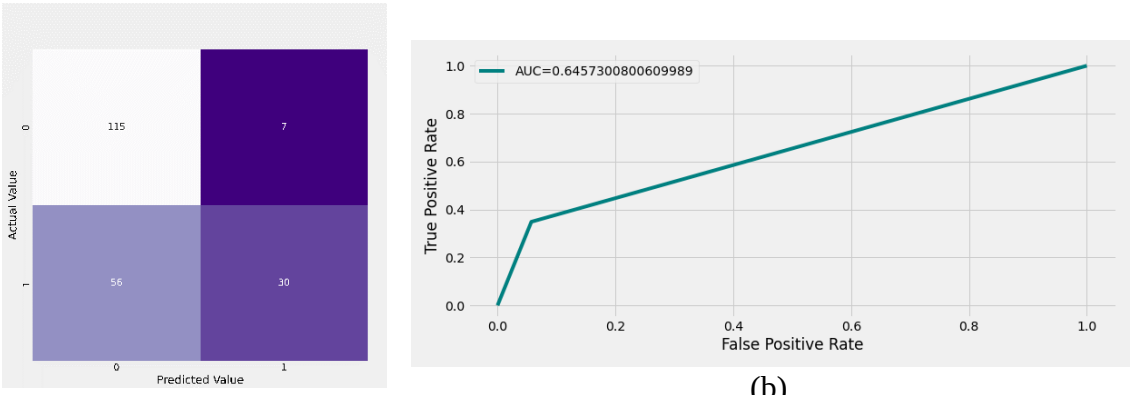
To increase the learning rate, the model has been optimized using the optimum values of the HP returned from the CSA. After HPO, we have trained the RF classifier with all the features and compared the results with the results of unoptimized model. The average accuracy of the optimized model after 10-fold CV was 64.55%, which is 2.67% more than the results of unoptimized model. The best performed model has an accuracy of 69.71%,

which is approximately 5.08% increase from the best performed classifier of unoptimized model. The computational cost was also drastically reduced. We will discuss about the computational complexity section 4.2.5. The Table 4.5 represents the performance metrices of the best performed model in the 10-fold.

Table 4.5: Performance metrices of best performed optimized model

Metrices	Accuracy	Precision	Recall	F1 Score	AUC
Values	69.71%	72.96%	69.71%	66.21%	64.57%

The confusion matrix and ROC curve of the best performed model has been represented in the Fig. 4.9.



(a) (b)

Fig. 4.9: (a) Confusion matrix (b) ROC curve of the best performed optimized model

- 0 → Arrhythmia class no
- 1 → Arrhythmia class yes

We can see from the Fig. 4.9 (a) that, the model has predicted 30 cases where arrhythmia occurred and 115 cases where it did not. The AUC of the ROC curve is 64.57% (Fig. 4.9 (b))

4.2.4 Feature Selection

We have performed FS to reduce the dimension of data and increase the accuracy of the RF classifier. FS was performed by using a meta heuristics algorithm, CSA. The fitness function was constructed based on the evaluation of the accuracy of the RF classifier. Other performance metrices, such as F1 score, precision, and recall were not taken into account. We have varied the parameter ‘p_a’ (probability of discovery of egg) of the CSA from 0.21 to 0.3. The ideal value for ‘p_a’ is considered 0.25 [92]. The ‘pop_size’ and ‘epoch’ were

100 for both parameters. The table 4.6 represents the sets of features returned from the CSA for different values of ‘p_a’.

Table 4.6: Features returned from CSA

Values of ‘p_a’	Returned Features	Feature Set
0.21	Heart Beats, DD-P	A
0.22	DBP, RA (right and left)	B
0.23	DBP, DD-P	C
0.24	P-R, BBB	D
0.25	RA (right and left), PCI	E
0.26	E/A, BBB	F
0.27	Heart Beats, E/A, LVEDD, RA (right and left), BBB	G
0.28	Ventricular Wall Motion Abnormal, PCI, BBB	H
0.29	SBP, P-R	I
0.3	Heart Beats, Pre-diabetes Mellitus, BBB	J

For different values of ‘p_a’, a different set of features was returned from the CSA. We have trained the optimized RF model with these 10 sets of features.

The comparison of average and best accuracy of each set of features has been shown in the Fig. 4.10. The highest average accuracy was returned by the feature set I, which is 62.72% and the best performance is 67.78%. The feature set H was returned from CSA for ‘p_a’ value 0.28. But the highest best accuracy can be seen for the feature set A returned from CSA for the value of ‘p_a’ equal to 0.21. The best performance of the model for this set of features is 68.89% and the average accuracy is 61.81%. The Table 4.7 represents the performance metrices of the 10 different sets of features returned from the CSA.

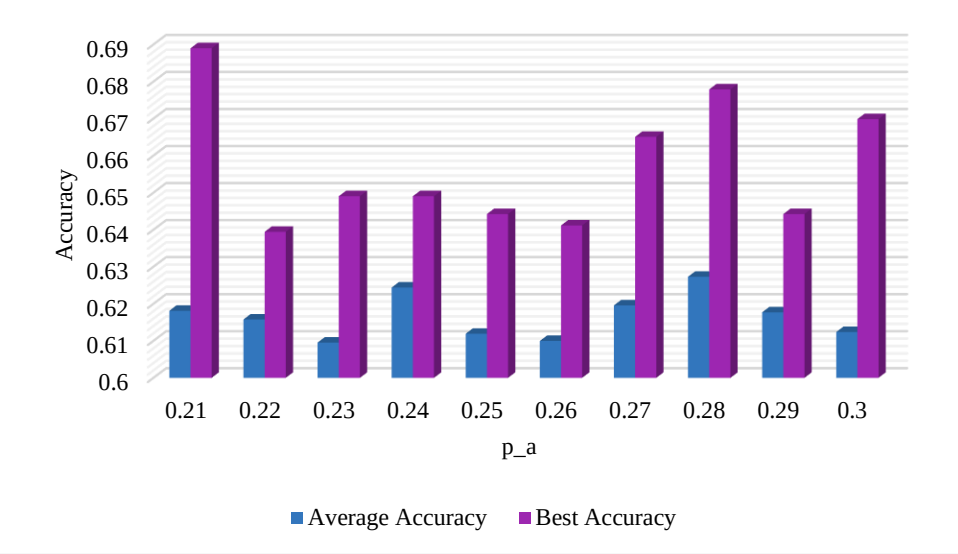


Fig. 4.10: Accuracy comparison of each set of features

Table 4.7: Performance metrices for best performed models of different sets of features

Feature Set	Accuracy	Precision	Recall	F1 Score	AUC
A	68.89%	70%	68.89%	66.25%	64.30%
B	63.94%	66.62%	63.94%	57.64%	57.59%
C	64.90%	65.39%	64.90%	61.14%	59.78%
D	64.90%	68.60%	64.90%	58.76%	58.58%
E	64.42%	65.75%	64.42%	59.50%	58.69%
F	64.11%	66.14%	64.11%	58.21%	57.79%
G	66.50%	66.26%	66.50%	64.41%	62.45%
H	67.78%	67.47%	67.78%	67.55%	66.19%
I	64.42%	67.27%	64.42%	58.40%	58.17%
J	66.98%	67.25%	66.98%	64.40%	62.50%

The Fig. 4.11 visualizes number of times a single feature was returned from the CSA. It can be seen from the figure that BBB was returned 5 times from CSA, Heart Beats and RA (right and left) were returned 3 times. From the statistics we can sum up that these 3

features, returned frequently may play a significant role in arrhythmia classification. So, to see their significance in arrhythmia classification we have trained our optimized model using BBB, Heart Beats and RA (right and left) features only.

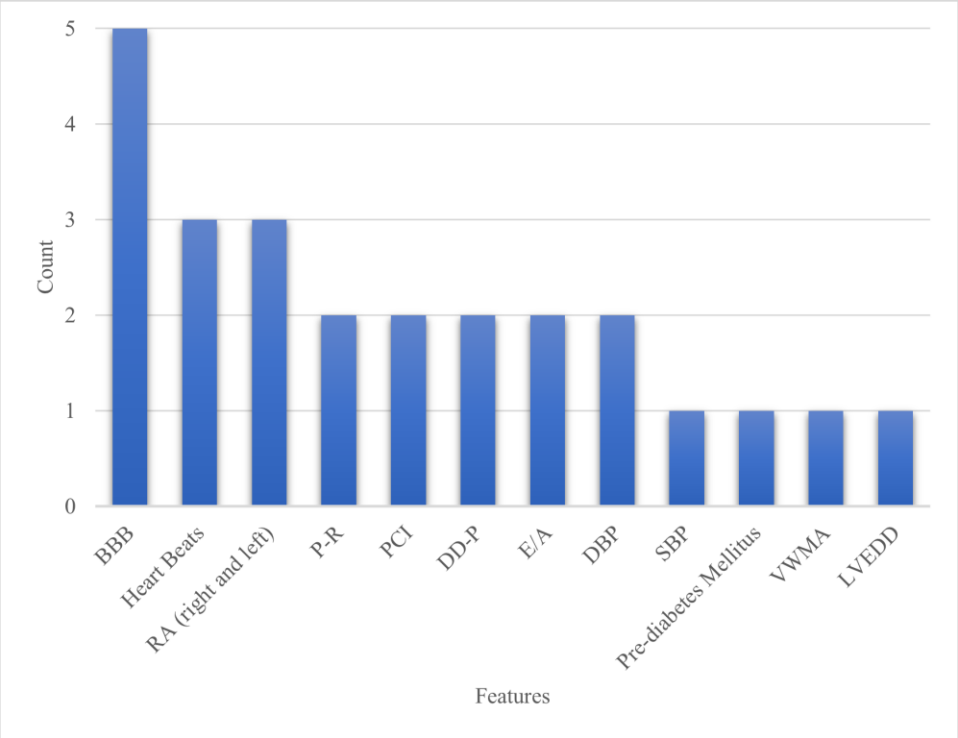


Fig. 4.11: Number of times same feature returned from CSA

- VWMA → Ventricular Wall Motion Abnormal

We have also considered the outcomes of the conventional statistical analysis. From the gamma coefficient analysis (Section 4.1.2), we have found four best features, P-R, BBB, and RA (right and left) and RA (up and down), which have correlation with arrhythmia. We have trained our model using these features to analyze their performance for predicting arrhythmia.

The average accuracy after 10-fold CV was 62.34% and the best performance shown was 67.46% by the most returned features. From the most associated features of statistical analysis, the average accuracy was 63.15% and the best performance was 65.55%. The performance of most returned features is slightly lower than the performance measured for the feature set 8. Table 4.8 provides insights of the best performed model of feature set A, the most returned features and the associated features returned from the statistical analysis.

Table 4.8: Performance metrices for best performed models

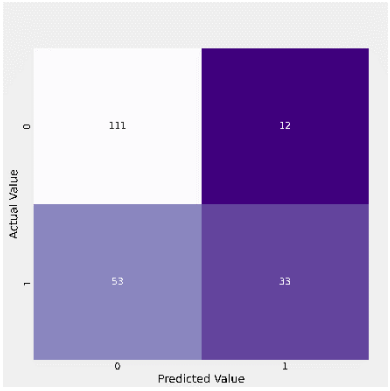
Feature Set	Accuracy	Precision	Recall	F1 Score	AUC
A	68.89%	70%	68.89%	66.25%	64.30%
K	67.46%	67.43%	67.46%	65.43%	63.43%
L	65.55%	67.47%	65.55%	60.76%	59.71%

- A → Feature set A [Table 4.8]
- K → Most returned features (BBB, Heart Beats, RA (right and left))
- L → Associated features returned from statistical analysis (P-R, BBB, and RA (right and left), RA (up and down))

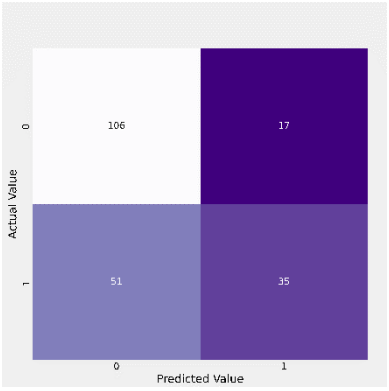
We have reduced the dimension of the dataset through FS. The aim was to improve the accuracy while reducing the dimension of the dataset. But after FS, the accuracy of the model decreased slightly from the accuracy of the optimized RF classifier trained with all the features.

The features returned from statistical and ML analysis is somewhat similar. BBB and RA (right and left) both features have been returned from the CSA and statistical analysis. The CSA has also returned P-R as an important feature in case of predicting arrhythmia similar to statistical analysis. But we did not take it into account as some other features were also returned twice as can be seen in the Fig. 4.11. But the statistical analysis did not extract heart beats as an important feature whereas the ML analysis did not extracted RA (up and down). A study shows that the dimension of left atria is an important factor for developing arrhythmia in patients [108]. Another study on the PR interval has shown it can be an important factor for predicting arrhythmia and it is easily obtainable from the echocardiography (ECG) data [109]. From the outcome of the study on bundle branch block (BBB) has shown that prevalence of arrhythmia is higher in patients who had BBB [110]. So, we can draw a conclusion to the fact that, BBB is also an important predictor for predicting arrhythmia. The studies [111] and [112] have shown heart beats as a significant factor for predicting arrhythmia. Both analysis approaches of this study have also extracted those features as important predictors of arrhythmia and the approaches individually has returned similar types of features.

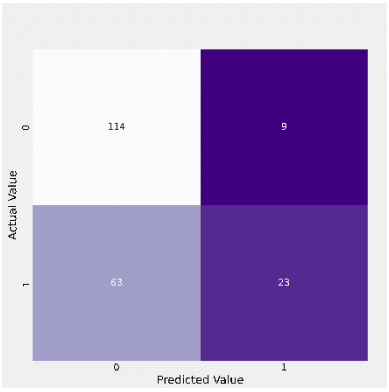
The Fig. 4.12 represents the confusion metrices of model trained by feature set A, most returned features and associated features returned from statistical analysis. From the figure we can see that the model trained by these features have accurately predicted 111, 106 and 114 cases respectively where the subjects did not suffer from arrhythmia. Similarly, 33, 35 and 23 cases respectively where they suffered from it. The model trained by associated features returned from statistical analysis has predicted the true negatives (TN) more accurately than the other two models. But in case of the true positives (TP), its performance is the lowest. Again, the model trained by most returned features from the CSA, predicted TP more accurately than other models and TN less accurately than other two models.



(a)



(b)



(c)

Fig. 4.12: Confusion matrix of models trained by (a) feature set a (b) most returned features (c) associated features returned from statistical analysis

The Fig. 4.13 represents the ROC curves of model trained by feature set A, most returned features and associated features returned from statistical analysis. From the figure, it can be seen that, the AUC of these models are 64.30%, 63.43% and 59.71% respectively. The

best performance has been shown by the model trained with feature set A and the worst has been shown by features returned from statistical analysis.

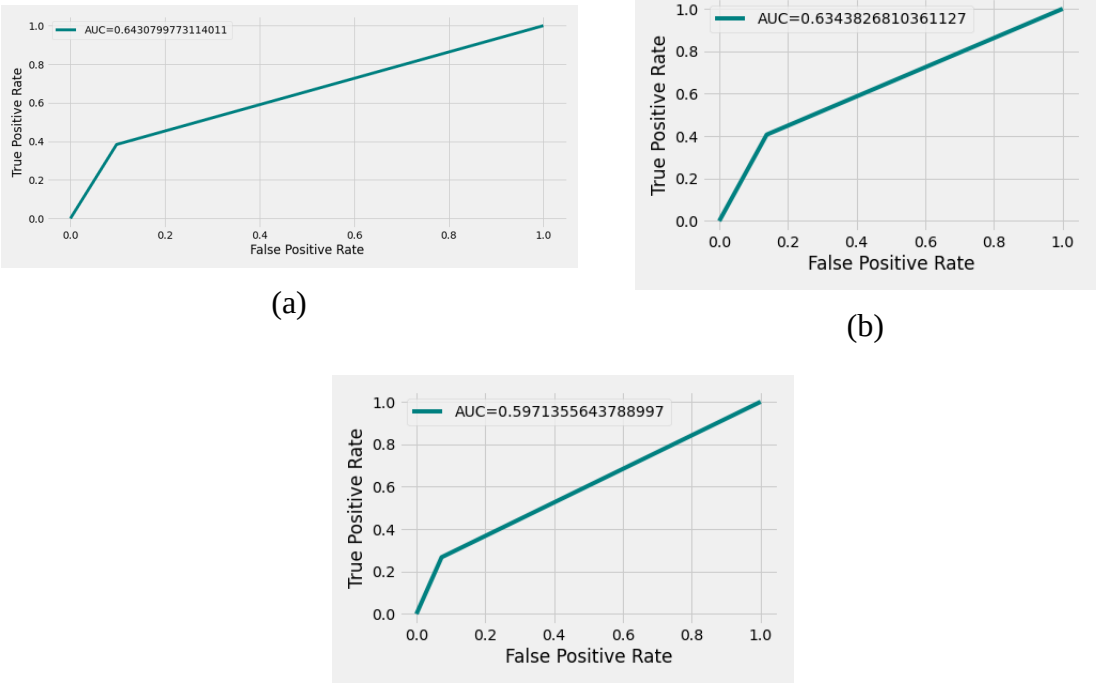


Fig. 4.13: ROC curve of models trained by (a) feature set A (b) most returned features (c) associated features returned from statistical analysis

4.2.5 Time Complexity Analysis

Time complexity is an important metric to measure the performance of an algorithm. Beside best accuracy of the RF classifier, we have considered time complexity to evaluate how efficiently and quickly our model can classify the problem. The unoptimized RF classifier needed an average of 429.10 ms to train the model. We calculated the average value from the train and test time measured at each fold CV. However, the train time significantly dropped while training the optimized RF classifier. The average time taken was 162.23 ms, almost 267 ms less time taken on average. In case of prediction time, unoptimized model took 41.02 ms and 14.05 ms by optimized model on average. Fig. 4.14 depicts the average time taken of all the models we have trained so far.

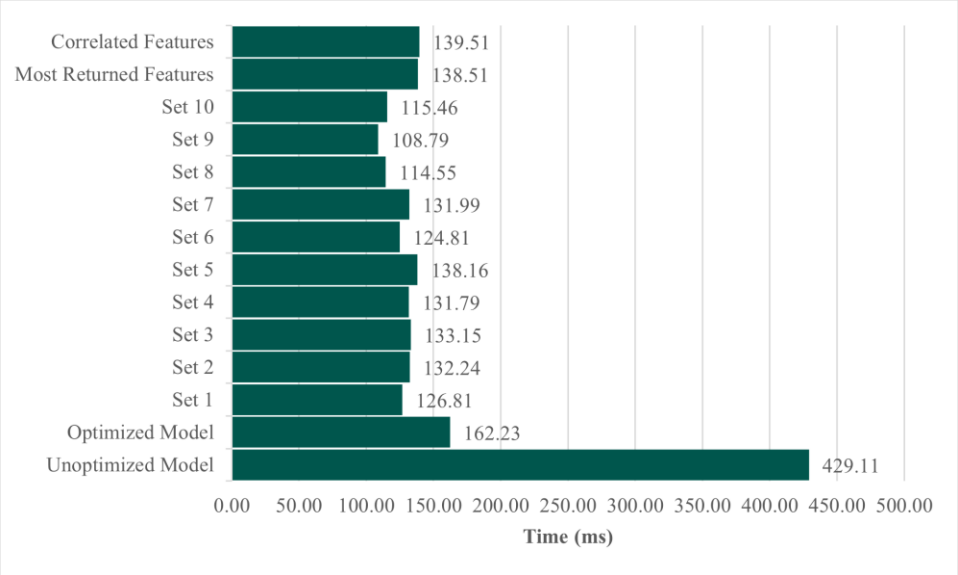


Fig. 4.14: Average train time of all models

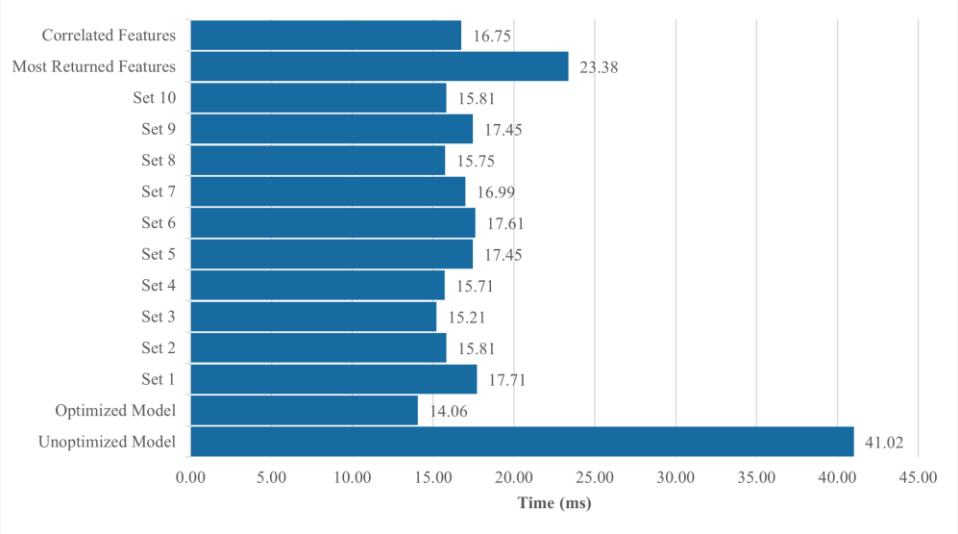


Fig. 4.15: Average test time of all models

We can say from the data that the time duration reduces as the model is optimized and the number of features i.e., data dimension and size is reduced. In the case of prediction time, we can validate this statement from the computed data. Fig. 4.15 shows the average time each model has taken to predict arrhythmia on test dataset. Once again, the unoptimized model has taken the longest to make predictions. The duration reduces as we optimized the model and lower dimensional data were provided.

4.2.6 Comparison with Other Researches

We have used the AMI dataset in this study. We have compared our work with some state-of-the-art works that used the RF classifier to predict arrhythmia as the goal of our study was to classify arrhythmia. The findings are listed in Table 4.9.

Table 4.9: Comparison of related works

Work	Classifier	Dataset	Accuracy	F1 - Score
[113]	RF	X	94.65%	-
[114]	RF	Y	-	97.7%
[115]	RF	Z	98.9%	-
[28]	RF/ANN	AMI Dataset	64.80%/66.80%	-
Our	RF	AMI Dataset	68.89%	66.25%

- X → Clinical Dataset of Children Undergoing Interventional Closure of ASD at the Heart Center of Qingdao Women and Children's Hospital
- Y → Physionet/Computing in Cardiology 2017 AF Challenge dataset and the Atrial Fibrillation Termination Database (AFTDB). The test datasets consist of the MIT-BIH Atrial Fibrillation Database (AFDB) and the MIT-BIH Arrhythmia Database (MITDB).
- Z → ECG, PPG, and accelerometry measurements in 40 patients undergoing a 24-hour Holter measurement as part of routine clinical care.
- ANN → Artificial neural networks

From the table, we can see that the works performed on dataset X, Y and Z has better model accuracy than our model. But we cannot directly compare our work with theirs because the dataset that has been used is different that the dataset we have used in this thesis. But the authors of [28], have used the same dataset that we have used in our thesis. They have used both RF and ANN, and the accuracy of RF and ANN was 64.80% and 66.80% respectively. Whereas, the RF classifier that we have used in this thesis, has shown an accuracy of 68.89% after FS and 69.71% while it was trained with all features. So, it can be concluded that we have improved the performance of the model than the previous work performed on the same dataset.

4.2.7 Dataset Analysis

To find out the reason of the poor performance of the model, we investigated the dataset. ML is a statistical method to classify or predict data. So, it depends on the quality of dataset. If the dataset is not in good shape, the prediction accuracy will be low and the model will not be reliable. To analyze the distribution of our dataset, we plotted the T-SNE plot. It visualizes higher dimensional data into 2D or 3D. We used a 2D to visualize the distribution of our dataset. Fig. 4.16 portrays the T-SNE distribution of the dataset.

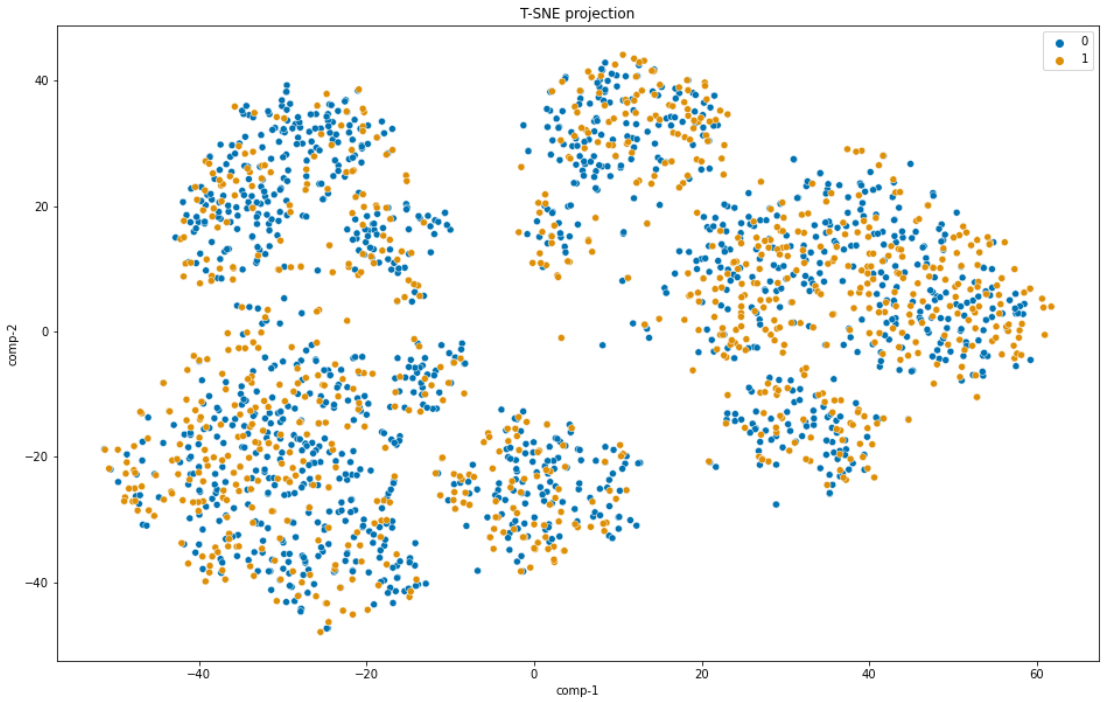


Fig. 4.16: T-SNE plot

In the Fig. 4.16, legend 0 indicates that arrhythmia did not occur and 1 indicates the opposite. It can also be observed the classes are not well distributed or clustered. The classes are spread out everywhere. Neighboring data points contain both classes frequently. This poor distribution of the data points affects the RF classifier, hence affecting the accuracy of classification.

Chapter 5
Conclusion and Future Work

5.1 Conclusion

Heart is a vital organ in human body. A human can die if the heart somehow stops or fails to pump blood. Arrhythmia is the irregular heart beat occurs when the bio-electric potentials do not work properly. It can cause so much pain in human body and sometimes it can be life threatening. The number of deaths can be reduced if it can be predicted earlier. The objectives of our thesis are (i) to identify important features for predicting arrhythmia (ii) to reduce time complexity (iii) to investigate the AMI dataset and (iv) to analyze the model performance. To achieve these objectives, we have used two different approaches to predict the occurrences of arrhythmia after AMI. From the statistical analysis approach, we have obtained P-R, BBB, and RA (right and left), RA (up and down) features which are associated for the occurrence of arrhythmia and from the machine learning analysis approach, we have obtained BBB, Heart Beats, RA (right and left) features, which has been repeatedly returned from the CSA. We have also validated the significance of these predictors for predicting arrhythmia by presenting studies that have also shown the significance of these predictors for predicting arrhythmia. Both approaches have returned almost similar types of features that are responsible for arrhythmia. The importance of these features has also been validated by medical journals. Using only these important features only, our model has shown an accuracy of 68.89% and while trained with all the features of the AMI dataset, it has shown 69.71% accuracy. Our model did not perform up to the mark compared to other models trained by different datasets, but it outperformed the model of previous work trained by the AMI dataset. Our goal was not only to predict arrhythmia accurately but also to make predictions as quickly as possible. We have minimized the time complexity by optimizing the RF classifier and reducing the number of features. Finally, we have investigated the dataset to find out the reason of the unsatisfactory performance of the model. We have found that the dataset is skewed and not well distributed. Hopefully this thesis will contribute in the field of biomedical. With the help of our model, arrhythmia can be predicted quickly and efficiently, also in a cost-effective manner.

5.2 Future Work

The drawback of this thesis was that we have used an imbalanced dataset that limits the performance of the model. In future our model will be trained with different datasets to increase the learning rate of the model. To improve the accuracy and reliability, genetic algorithms (GA) and neural networks (NN) will be applied. A telehealth-based patient monitoring system will be developed where this model will be embedded in a server to predict arrhythmia.

References

[1] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability," *Frontiers in psychology,* vol. 5, 2014.

[2] X. Jouven, J.-P. Empana, P. J. Schwartz, M. Desnos, D. Courbon, and P. Ducimetière, "Heart-rate profile during exercise as a predictor of sudden death," *New England journal of medicine,* vol. 352, no. 19, pp. 1951-1958, 2005.

[3] C. Antzelevitch and A. Burashnikov, "Overview of basic mechanisms of cardiac arrhythmia," *Cardiac electrophysiology clinics,* vol. 3, no. 1, pp. 23-45, 2011.

[4] C. J. Murray and A. D. Lopez, "Mortality by cause for eight regions of the world: Global Burden of Disease Study," *The lancet,* vol. 349, no. 9061, pp. 1269-1276, 1997.

[5] N. J. Patel, V. Atti, R. D. Mitrani, J. F. Viles-Gonzalez, and J. J. Goldberger, "Global rising trends of atrial fibrillation: a major public health concern," vol. 104, ed: BMJ Publishing Group Ltd and British Cardiovascular Society, 2018, pp. 1989-1990.

[6] G. Lippi, F. Sanchis-Gomar, and G. Cervellin, "Global epidemiology of atrial fibrillation: an increasing epidemic and public health challenge," *International Journal of Stroke,* vol. 16, no. 2, pp. 217-221, 2021.

[7] E. Nichols *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019," *The Lancet Public Health,* vol. 7, no. 2, p. 21, 2022.

[8] Y.-C. Wang *et al.*, "Current advancement in diagnosing atrial fibrillation by utilizing wearable devices and artificial intelligence: a review study," *Diagnostics,* vol. 12, no. 3, p. 689, 2022.

[9] K. R. Hassler and H. Ramakrishna, "Predicting Postoperative Atrial Fibrillation: The Search Continues," *Journal of Cardiothoracic Vascular Anesthesia,* vol. 36, no. 10, pp. 3738-3739, 2022.

[10] S. H. Habib and S. Saha, "Burden of non-communicable disease: global overview," *Diabetes Metabolic Syndrome: Clinical Research Reviews,* vol. 4, no. 1, pp. 41-47, 2010.

[11] J. Jaakkola *et al.*, "The effect of mental health conditions on the use of oral anticoagulation therapy in patients with atrial fibrillation: the FinACAF study," *European Heart Journal-Quality of Care Clinical Outcomes,* vol. 8, no. 3, pp. 269-276, 2022.

[12] R. Khera, J. Valero-Elizondo, and K. Nasir, "Financial toxicity in atherosclerotic cardiovascular disease in the United States: current state and future directions," *Journal of the American Heart Association,* vol. 9, no. 19, 2020.

[13] M. Z. I. Chowdhury *et al.*, "Prevalence of cardiovascular disease among Bangladeshi adult population: a systematic review and meta-analysis of the studies," *Vascular health risk management,* pp. 165-181, 2018.

[14] A. M. Islam, A. Mohibullah, and T. Paul, "Cardiovascular disease in Bangladesh: a review," *Bangladesh Heart Journal,* vol. 31, no. 2, pp. 80-99, 2016.

[15] L. Qiao, R. Cadathur, and D. C. Gari, "Ventricular fibrillation and tachycardia classification using a machine learning approach," *IEEE Transactions on Biomedical Engineering,* vol. 61, no. 6, p. 17, 2013.

[16] A.-A. Felipe, M. Eduardo, F.-M. Lorena, G.-A. Arcadi, and L. Jose, Rojo-Alvarez "Detection of life-threatening arrhythmias using feature selection and support vector machines," *IEEE Transactions on Biomedical Engineering,* vol. 61, no. 3, p. 9, 2013.

[17] B. Amina , C. M. Amine, and B. Sarra, "Classifier set selection for cardiac arrhythmia recognition using diversity," *Journal of Medical Imaging and Health Informatics,* vol. 5, no. 3, p. 7, 2015.

[18] V. Kalidas and L. S. Tamil, "Cardiac arrhythmia classification using multi-modal signal analysis," *Physiological Measurement,* vol. 37, no. 8, p. 20, 2016.

[19] K. Yasin, P. Hüseyin, and E. Mehmet, Tenekeci, "Effective ECG beat classification using higher order statistic features and genetic feature selection," *BIOMEDICAL RESEARCH-INDIA,* vol. 28, no. 17, 2017.

[20] M. Anam, M. Syed, Anwar , and M. Muahammad, "Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants," *Computational mathematical methods in medicine,* vol. 2018, 2018.

[21] N. Singh and P. Singh, "Cardiac arrhythmia classification using machine learning techniques," in *Engineering Vibration, Communication and Information Processing: ICoEVCI 2018, India,* 2019, pp. 469-480: Springer.

[22] G. T. Taye, E. B. Shim, H.-J. Hwang, and K. M. Lim, "Machine learning approach to predict ventricular fibrillation based on QRS complex shape," *Frontiers in physiology,* vol. 10, p. 1193, 2019.

[23] S. Liaqat, K. Dashtipour, A. Zahid, K. Assaleh, K. Arshad, and N. Ramzan, "Detection of atrial fibrillation using a machine learning approach," *Information,* vol. 11, no. 12, p. 549, 2020.

[24] S. Wang *et al.*, "AMI Dataset," ed. GitHub, 2021.

[25] R. Hu, J. Chen, and L. Zhou, "A transformer-based deep neural network for arrhythmia detection using continuous ECG signals," *Computers in Biology Medicine,* vol. 144, 2022.

[26] J. L. Anderson *et al.*, "Interaction of baseline characteristics with the hazard of encainide, flecainide, and moricizine therapy in patients with myocardial infarction. A possible explanation for increased mortality in the Cardiac Arrhythmia Suppression Trial (CAST)," *Circulation,* vol. 90, no. 6, pp. 2843-2852, 1994.

[27] Z. I. Attia *et al.*, "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *The Lancet,* vol. 394, no. 10201, pp. 861-867, 2019.

[28] S. Wang *et al.*, "Application of machine learning to predict the occurrence of arrhythmia after acute myocardial infarction," *BMC medical informatics and decision making,* vol. 21, pp. 1-14, 2021.

[29] K. Blakeman, "Bibliometrics in a digital age: help or hindrance," *Science progress,* vol. 101, no. 3, pp. 293-310, 2018.

[30] T. Tuncer, S. Dogan, P. Pławiak, and U. R. Acharya, "Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals," *Knowledge-Based Systems,* vol. 186, p. 104923, 2019.

[31] Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Computers in biology in medicine,* vol. 102, pp. 411-420, 2018.

[32] V. K. Singh, P. Singh, M. Karmakar, J. Leta, and P. Mayr, "The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis," *Scientometrics,* vol. 126, pp. 5113-5142, 2021.

[33] A. Liberati *et al.*, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration," *Annals of internal medicine,* vol. 151, no. 4, p. 30, 2009.

[34] M. Aria, C. Cuccurullo, and M. M. Aria, "Package 'bibliometrix'," ed, 2017.

[35] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network," *Information sciences,* vol. 405, pp. 81-90, 2017.

[36] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Computers in biology medicine,* vol. 102, pp. 278-287, 2018.

[37] O. Faust, A. Shenfield, M. Kareem, T. R. San, H. Fujita, and U. R. Acharya, "Automated detection of atrial fibrillation using long short-term memory network with RR interval signals," *Computers in biology medicine,* vol. 102, pp. 327-335, 2018.

[38] M. A. Kobat, O. Karaca, P. D. Barua, and S. Dogan, "Prismatoidpatnet54: an accurate ECG signal classification model using prismatoid pattern-based learning architecture," *Symmetry,* vol. 13, no. 10, 2021.

[39] M. Baygin, T. Tuncer, S. Dogan, R.-S. Tan, and U. R. Acharya, "Automated arrhythmia detection with homeomorphically irreducible tree technique using more than 10,000 individual subject ECG records," *Information Sciences,* vol. 575, pp. 323-337, 2021.

[40] T. Turker, D. Sengul, P. Pawel, and S. Abdulhamit, "A novel Discrete Wavelet-Concatenated Mesh Tree and ternary chess pattern based ECG signal recognition method," *Biomedical Signal Processing and Control,* vol. 72.

[41] V. Mondéjar-Guerra, J. Novo, J. Rouco, M. G. Penedo, and M. Ortega, "Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers," *Biomedical Signal Processing and Control,* vol. 47, p. 8, 2018.

[42] R. Patrick, H. Manfred, and F. Bernd, "Interactive Sankey diagrams," *IEEE Symposium on Information Visualization,* p. 8, 2005.

[43] P. Kemal, Ş. Seral, and G. Salih, "A New Method to Medical Diagnosis: Artificial Immune Recognition System (AIRS) with Fuzzy Weighted Pre-processing and Application to ECG Arrhythmia," *Expert Systems with Applications,* vol. 31, no. 2, pp. 264-269, 2006.

[44] H. Khorrami and M. Moavenian, "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification," *Expert systems with Applications,* vol. 37, no. 8, p. 8, 2010.

[45] V. Mondéjar-Guerra, J. Novo, J. Rouco, M. G. Penedo, and M. Ortega, "Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers," *Biomedical Signal Processing Control,* vol. 47, pp. 41-48, 2019.

[46] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "A novel application of deep learning for single-lead ECG classification," *Computers in biology medicine,* vol. 99, pp. 53-62, 2018.

[47] B. Hou, J. Yang, P. Wang, and R. Yan, "LSTM-based auto-encoder model for ECG arrhythmias classification," *IEEE Transactions on Instrumentation Measurement,* vol. 69, no. 4, pp. 1232-1240, 2019.

[48] R. M. Cormack, "A review of classification," *Journal of the Royal Statistical Society: Series A (General),* vol. 134, no. 3, pp. 321-353, 1971.

[49] J. L. Myers, A. D. Well, and R. F. Lorch, *Research design and statistical analysis*. Routledge, 2013.

[50] M. P. Fay and M. A. J. S. s. Proschan, "Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules," vol. 4, p. 1, 2010.

[51] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.

[52] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, p. 5, 2006.

[53] K. S. Sree, J. Karthik, C. Niharika, P. Srinivas, N. Ravinder, and C. Prasad, "Optimized Conversion of Categorical and Numerical Features in Machine Learning Models," in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, 2021, pp. 294-299: IEEE.

[54] A. J. Harrison *et al.*, "Recommendations for statistical analysis involving null hypothesis significance testing," vol. 19, ed: Taylor & Francis, 2020, pp. 561-568.

[55] V. Grech and N. J. E. h. d. Calleja, "WASP (Write a Scientific Paper): Parametric vs. non-parametric tests," vol. 123, pp. 48-49, 2018.

[56] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: a guide for non-statisticians," *International journal of endocrinology metabolism*, vol. 10, no. 2, p. 486, 2012.

[57] E. Volchok, "Clear-Sighted Statistics: Module 17: Chi-Square Tests," 2020.

[58] T. M. Franke, T. Ho, and C. A. Christie, "The chi-square test: Often used and more often misinterpreted," *American journal of evaluation*, vol. 33, no. 3, pp. 448-458, 2012.

[59] R. Rana and R. Singhal, "Chi-square test and its application in hypothesis testing," *Journal of the Practice of Cardiovascular Sciences*, vol. 1, no. 1, p. 69, 2015.

[60] J. Fang, L. Liu, and P. Fang, "What is the most important factor affecting patient satisfaction–a study based on gamma coefficient," *Patient preference adherence*, vol. 13, 2019.

[61] A. J. T. A. S. Raveh, "On measures of monotone association," vol. 40, no. 2, pp. 117-123, 1986.

[62] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[63] C. Rygielski, J.-C. Wang, and D. C. Yen, "Data mining techniques for customer relationship management," *Technology in society*, vol. 24, no. 4, pp. 483-502, 2002.

[64] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE transactions on visualization computer graphics,* vol. 12, no. 4, p. 11, 2006.

[65] F. J. B. Yates, "The analysis of contingency tables with groupings based on quantitative characters," vol. 35, no. 1/2, pp. 176-181, 1948.

[66] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.

[67] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee, "A taxonomy of dirty data," *Data mining and knowledge discovery,* vol. 7, pp. 81-99, 2003.

[68] C. R. Harris *et al.*, "Array programming with NumPy," *Nature,* vol. 585, no. 7825, pp. 357-362, 2020/09/01 2020.

[69] W. McKinney, "Data structures for statistical computing in python," vol. 445, pp. 51-56: Austin, TX.

[70] J. Reback *et al.*, "pandas-dev/pandas: Pandas 1.0. 5," *Zenodo,* 2020.

[71] F. Pedregosa *et al.*, "Scikit-learn: Machine Learn-ing in Python, Journal of Machine Learning Re-search, 12," 2011.

[72] D. S. Starnes, D. Yates, and D. S. Moore, *The practice of statistics*. Macmillan, 2010.

[73] T. Agrawal, *Hyperparameter optimization in machine learning: make your machine learning and deep learning models more efficient*. Springer, 2021.

[74] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013.

[75] G. I. Diaz, A. Fokoue-Nkoutche, G. Nannicini, and H. Samulowitz, "An effective algorithm for hyperparameter optimization of neural networks," *IBM Journal of Research and Development,* vol. 61, no. 4/5, pp. 9-1, 2017.

[76] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing,* vol. 415, pp. 295-316, 2020/11/20/ 2020.

[77] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[78] N. DeCastro-García, A. L. Munoz Castaneda, D. Escudero Garcia, and M. V. Carriegos, "Effect of the sampling of a dataset in the hyperparameter optimization phase over the efficiency of a machine learning algorithm," *Complexity,* vol. 2019, 2019.

[79] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *The New England journal of medicine,* vol. 375, no. 13, p. 1216, 2016.

[80] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems,* vol. 12, no. 1, pp. 95-116, 2007/05/01 2007.

[81] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab1996.

[82] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014.

[83] T. Dokeroglu, A. Deniz, and H. E. Kiziloz, "A comprehensive survey on recent metaheuristics for feature selection," *Neurocomputing*, 2022.

[84] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273-324, 1997.

[85] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on knowledge and data engineering*, vol. 17, no. 4, pp. 491-502, 2005.

[86] I. Boussaïd, J. Lepagnot, and P. Siarry, "A survey on optimization metaheuristics," *Information sciences*, vol. 237, pp. 82-117, 2013.

[87] T. Dokeroglu, E. Sevinc, T. Kucukyilmaz, and A. Cosar, "A survey on new generation metaheuristic algorithms," *Computers & Industrial Engineering*, vol. 137, p. 106040, 2019.

[88] A. Deniz, H. E. Kiziloz, T. Dokeroglu, and A. Cosar, "Robust multiobjective evolutionary feature subset selection algorithm for binary classification using machine learning techniques," *Neurocomputing*, vol. 241, pp. 128-146, 2017.

[89] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary ant lion approaches for feature selection," *Neurocomputing*, vol. 213, pp. 54-65, 2016.

[90] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371-381, 2016.

[91] E. Hancer, B. Xue, M. Zhang, D. Karaboga, and B. Akay, "Pareto front feature selection based on artificial bee colony optimization," *Information Sciences*, vol. 422, pp. 462-479, 2018.

[92] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," pp. 210-214: Ieee.

[93] N. v. Thieu and S. Mirjalili, "MEALPY: a Framework of The State-of-The-Art Meta-Heuristic Algorithms in Python," v2.4.2 ed: Zenodo, June, 2022.

[94] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, vol. 1, pp. 278-282 vol.1.

[95] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.

[96] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001/10/01 2001.

[97] G. Biau and E. Scornet, "A random forest guided tour," *TEST,* vol. 25, no. 2, pp. 197-227, 2016/06/01 2016.

[98] J. A. Aslam, R. A. Popa, and R. L. Rivest, "On Estimating the Size and Confidence of a Statistical Audit," *EVT,* vol. 7, p. 8, 2007.

[99] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, pp. 123-140, 1996.

[100] A. Majeed, "Improving time complexity and accuracy of the machine learning algorithms through selection of highly weighted top k features from complex datasets," *Annals of Data Science,* vol. 6, pp. 599-621, 2019.

[101] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems,* vol. 15, 2002.

[102] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9, no. 11, 2008.

[103] C. E. Metz, "Basic principles of ROC analysis," in *Seminars in nuclear medicine,* vol. 8, pp. 283-298: Elsevier.

[104] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[105] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC medical imaging,* vol. 15, no. 1, pp. 1-28, 2015.

[106] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote sensing of Environment,* vol. 62, no. 1, pp. 77-89, 1997.

[107] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27, no. 8, pp. 861-874, 2006.

[108] W. L. Henry *et al.*, "Relation between echocardiographically determined left atrial size and atrial fibrillation," *Circulation,* vol. 53, no. 2, pp. 273-279, 1976.

[109] K. Schumacher, N. Dagres, G. Hindricks, D. Husser, A. Bollmann, and J. Kornej, "Characteristics of PR interval as predictor for atrial fibrillation: association with biomarkers and outcomes," *Clinical Research in Cardiology,* vol. 106, pp. 767-775, 2017.

[110] M. Z. Khan *et al.*, "Association between atrial fibrillation and bundle branch block," *Journal of Arrhythmia,* vol. 37, no. 4, pp. 949-955, 2021.

[111] G. Sannino and G. De Pietro, "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection," *Future Generation Computer Systems,* vol. 86, pp. 446-455, 2018/09/01/ 2018.

[112] C. Philip de, M. O. Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196-1206, 2004.

[113] H. Sun, Y. Liu, B. Song, X. Cui, G. Luo, and S. Pan, "Prediction of arrhythmia after intervention in children with atrial septal defect based on random forest," *BMC pediatrics*, vol. 21, no. 1, p. 280, 2021.

[114] V. Kalidas and L. S. Tamil, "Detection of atrial fibrillation using discrete-state Markov models and Random Forests," *Computers in biology and medicine*, vol. 113, p. 103386, 2019.

[115] L. M. Eerikäinen *et al.*, "Detecting atrial fibrillation and atrial flutter in daily life using photoplethysmography data," *IEEE journal of biomedical and health informatics*, vol. 24, no. 6, pp. 1610-1618, 2019.

Appendix

We have used crosstab analysis to analyze the categorical association between arrhythmia and other 43 features. From this analysis we can find out the important categories of features which are highly responsible for the occurrence of arrhythmia. The outcome of crosstab analysis is given in Table A.1.

Table A.1: Outcomes of Crosstab Analysis

Features Name	**Measurement Type**	**Categories**	**Labeled As**	**Arrhythmia Status**	
				Yes	**No**
Age	Ordinal	＜30	0	3 20.0%	12 80.0%
		30-39	1	15 22.1%	53 77.9%
		40-49	2	102 33.2%	205 66.8%
		50-59	3	228 36.9%	390 63.1%
		60-69	4	310 45.6%	370 54.4%
		70-79	5	167 50.5%	164 49.5%
		80-89	6	34 56.7%	26 43.3%
		＞90	7	1 33.3%	2 66.7%
Systolic Pressure	Ordinal	Under 80	0	20 69.00%	9 31.00%
		80-99	1	78 58.20%	56 41.80%

		100-119	2	207 43.50%	269 56.50%
		120-139	3	276 39.50%	423 60.50%
		140-159	4	178 39.00%	278 61.00%
		160-199	5	99 36.00%	176 64.00%
		200 and older	6	2 15.40%	11 84.60%
Diastolic pressure	Ordinal	＜60	0	86 63.20%	50 36.80%
		60-69	1	114 46.90%	129 53.10%
		70-79	2	227 44.30%	285 55.70%
		80-89	3	194 36.40%	339 63.60%
		90-99	4	130 38.20%	210 61.80%
		100-109	5	70 38.20%	114 61.80%
		≥110	6	39 29.10%	95 70.90%
Heart beats	Ordinal	＜50	0	60 75.00%	20 25.00%
		50-69	1	318 44.20%	402 55.80%
		70-89	2	327 36.40%	571 63.60%

		90-109	3	124 38.60%	197 61.40%
		110-149	4	25 45.60%	31 54.40%
		150-199	5	5 83.30%	1 16.70%
		≥200	6	1 100.00%	0 0.00%
Pro-BNP	Ordinal	＜300	0	151 38.20%	251 61.80%
		300-449	1	60 38.70%	95 61.30%
		450-899	2	158 39.00%	247 61.00%
		900-1799	3	156 38.20%	252 61.80%
		1800-4499	4	222 46.90%	251 53.10%
		4500-8999	5	75 44.60%	93 55.40%
		＞9000	6	34 50.70%	33 49.30%
CRP	Ordinal	＜8	0	340 41.60%	478 58.40%
		≥8	1	520 41.10%	744 58.90%
Total cholesterol	Ordinal	＜3.1	0	148 42.00%	204 58.00%
		3.1-4.1	1	273 42.10%	375 57.90%

		4.2-5.1	2	237 41.30%	337 58.70%
		5.2-7.2	3	187 40.80%	271 59.20%
		＞7.2	4	15 30.00%	35 70.00%
Triglyceride	Ordinal	＜1.7	0	496 44.00%	631 56.00%
		1.7-2.3	1	180 39.00%	281 61.00%
		＞2.3	2	184 37.20%	310 62.80%
HDL	Ordinal	≤1	0	427 41.40%	604 58.60%
		＞1	1	433 41.20%	618 58.80%
LDL	Ordinal	＜1.8	0	108 38.20%	175 61.80%
		1.8-2.5	1	249 46.50%	287 53.50%
		2.6-3.3	2	270 39.80%	408 60.20%
		3.4-4.9	3	211 39.20%	327 60.80%
		＞4.9	4	22 46.80%	25 53.20%
Cr	Ordinal	0-34.48	0	23 45.30%	29 54.70%
		35.36-69.84	1	362 38.40%	580 61.60%

		70.72-105.20	2	387 42.10%	532 57.90%
		106.08-140.55	3	61 51.70%	57 48.30%
		141.44-175.92	4	13 54.20%	11 45.80%
		176.8-352.72	5	10 50.00%	10 50.00%
		＞353.6	6	4 57.10%	3 42.90%
k+	Ordinal	＜3.5	0	72 50.00%	72 50.00%
		3.5-5.0	1	754 40.70%	1100 59.30%
		＞5.0	2	34 40.50%	50 59.50%
TNI	Ordinal	＜0.03	0	104 40.30%	154 59.70%
		0.03-1	1	82 35.20%	151 64.80%
		1.01-10	2	111 36.80%	191 63.20%
		10.01-50	3	331 41.50%	467 58.50%
		50.01-200	4	202 46.40%	233 53.60%
		＞200	5	30 53.60%	26 46.40%
CK-MB	Ordinal	＜25	0	177 36.60%	307 63.40%

		25-41	1	81 40.70%	118 59.30%
		42-81	2	152 40.50%	223 59.50%
		82-164	3	189 40.60%	277 59.40%
		165-300	4	152 45.10%	185 54.90%
		＞300	5	109 49.30%	112 50.70%
DD-P	Ordinal	＜0.55	0	600 38.20%	970 61.80%
		≥0.55	1	260 50.80%	252 49.20%
LVEF	Ordinal	＜30	0	10 50.00%	10 50.00%
		30-44	1	187 45.00%	229 55.00%
		45-54	2	328 41.50%	462 58.50%
		≥55	3	335 39.10%	521 60.90%
FS	Ordinal	＜26	0	407 43.60%	526 56.40%
		≥26	1	453 39.40%	696 60.60%
E/A	Ordinal	＜1	0	559 40.00%	840 60.00%
		≥1	1	301 44.10%	382 55.90%

dt	Ordinal	＜167	0	471 41.70%	658 58.30%
		167-231	1	322 41.30%	458 58.70%
		≥231	2	67 38.70%	106 61.30%
LVEDD	Ordinal	≤55	0	790 41.10%	1130 58.90%
		＞55	1	70 43.20%	92 56.80%
IVST	Ordinal	＜12	0	826 41.30%	1172 58.70%
		≥12	1	34 40.50%	50 59.50%
LVPWT	Ordinal	＜12	0	838 41.10%	1199 58.90%
		≥12	1	22 48.90%	23 51.10%
LA	Ordinal	＜40	0	689 40.30%	1022 59.70%
		≥40	1	171 46.10%	200 53.90%
RA (up and down)	Ordinal	＜50	0	797 40.50%	1172 59.50%
		≥50	1	63 55.80%	50 44.20%
RA (right and left)	Ordinal	＜40	0	805 40.30%	1195 59.70%
		≥40	1	55 67.10%	27 32.90%

PA	Ordinal	＜26	0	691 40.50%	1015 59.50%
		≥26	1	169 44.90%	207 55.10%
VPA	Ordinal	＜0.6	0	40 55.60%	32 44.40%
		0.6-0.9	1	669 41.40%	946 58.60%
		＞0.9	2	151 38.20%	244 61.80%
Vao	Ordinal	＜1	0	241 44.20%	304 55.80%
		1-1.7	1	606 40.20%	902 59.80%
		＞1.7	2	13 44.80%	16 55.20%
P-R	Ordinal	＜0.12	0	13 39.40%	20 60.60%
		0.12-0.2	1	768 39.70%	1165 60.30%
		＞0.2	2	79 66.10%	37 31.90%
Q-Tc	Ordinal	＜340	0	1 25.00%	3 75.00%
		340-439	1	395 39.20%	612 60.80%
		≥440	2	464 43.30%	607 56.70%
UGLU	Nominal	(-)	0	566 39.70%	860 60.30%

		1+	1	56 45.90%	66 54.10%
		2+	2	42 51.90%	39 48.10%
		3+	3	56 47.90%	61 52.10%
		4+	4	74 36.60%	128 63.40%
		(±)	5	66 49.30%	68 50.70%
Pre-hypertension	Nominal	Yes	1	442 41.70%	618 58.30%
		No	0	418 40.90%	604 59.10%
Pre-diabetes mellitus	Nominal	Yes	1	214 43.60%	277 56.40%
		No	0	646 40.60%	945 59.40%
Smoker	Nominal	Yes	1	539 42.30%	736 57.70%
		No	0	321 39.80%	486 60.20%
Drinker	Nominal	Yes	1	173 40.10%	258 59.90%
		No	0	687 41.60%	964 58.40%
Prior MI	Nominal	Yes	1	81 48.50%	86 51.50%
		No	0	779 40.70%	1136 59.30%

Prior CI	Nominal	Yes	1	110 48.90%	115 51.10%
		No	0	750 40.40%	1107 59.60%
Prior HF	Nominal	Yes	1	7 58.30%	5 41.70%
		No	0	853 41.20%	1217 58.80%
Prior CHD	Nominal	Yes	1	48 50.50%	47 49.50%
		No	0	812 40.90%	1175 59.10%
Sex	Nominal	Male	0	249 46.90%	282 53.10%
		Female	1	611 39.40%	940 60.60%
Ventricular wall motion abnormal	Nominal	≥2 walls	0	213 44.20%	269 55.80%
		Anterior	1	217 33.50%	430 66.50%
		Apex	2	3 13.60%	19 86.40%
		Anteroseptal	3	2 9.50%	19 90.50%
		Posterior	4	135 47.70%	148 52.30%
		Inferior	5	268 47.50%	296 52.50%
		None	6	22 34.90%	41 65.10%

PCI	Nominal	LAD	0	95 26.60%	262 73.40%
		LCX	1	20 30.30%	46 69.70%
		RCA	2	112 58.90%	78 41.10%
		LM	3	35 47.90%	38 52.10%
		LAD+LCX	4	64 28.80%	158 71.20%
		LAD+RCA	5	166 45.60%	198 54.40%
		RCA+LCX	6	60 48.00%	65 52.00%
		Triple vessels	7	308 45.00%	377 55.00%
BBB	Nominal	None	0	553 38.10%	899 61.90%
		LAFB	1	175 45.00%	214 55.00%
		LBBB	2	19 46.30%	22 53.70%
		LPFB	3	35 41.70%	49 58.30%
		RBBB	4	51 67.10%	25 32.90%
		LAFB+RBBB	5	27 67.50%	13 32.50%

- The AMI Dataset, that we have used in this thesis is available on GitHub. Link: https://github.com/wangsuhuai/AMI-database1