

# **An Efficient Hybrid Classification Model for Heart Disease Prediction**



By

**Maaham Munsif**

**Fall-2019-MS-CS 00000319159 SEECS**

Supervisor

**Dr. Mehvish Rashid**

**Department of Computing**

A thesis submitted in partial fulfillment of the requirements for the degree of Masters  
of Science in Computer Science (MS CS)

In

School of Electrical Engineering & Computer Science (SEECS) ,

National University of Sciences and Technology (NUST),

Islamabad, Pakistan.

(June 2023)

## THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "An efficient hybrid classification model for Heart disease prediction" written by MAAHAM MUNSIF, (Registration No 00000319159), of SEECs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: \_\_\_\_\_  \_\_\_\_\_

Name of Advisor: Dr Mehvish Rashid \_\_\_\_\_

Date: 16-Jun-2023 \_\_\_\_\_

HoD/Associate Dean: \_\_\_\_\_

Date: \_\_\_\_\_

Signature (Dean/Principal): \_\_\_\_\_

Date: \_\_\_\_\_

## Approval

It is certified that the contents and form of the thesis entitled "An efficient hybrid classification model for Heart disease prediction" submitted by MAAHAM MUNSIF have been found satisfactory for the requirement of the degree

Advisor : Dr Mehvish Rashid

Signature:  \_\_\_\_\_


Date: 16-Jun-2023

Co-Advisor: Dr Farzana Jabeen

Signature:  \_\_\_\_\_

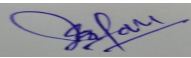
Date: 15-Jun-2023

Committee Member 1:Dr. Asad Waqar Malik

Signature:  \_\_\_\_\_

Date: 15-Jun-2023

Committee Member 2:Dr. Rabia Irfan

Signature:  \_\_\_\_\_

Date: 16-Jun-2023

# Dedication

To my beloved parents, whose unwavering dedication and countless sacrifices have shaped me into the person I am today. Your unwavering support, love, and guidance have been the driving force behind my journey. Thank you for your unconditional love, for instilling in me the values of hard work, never giving up, and perseverance, and for always being there for me at every step of my life.

To "Izzah," I dedicate this thesis to you.

This thesis is dedicated to all of you, whose contributions have been invaluable in my academic and personal growth. Thank you for believing in me, for inspiring me to reach new heights, and for being the pillars of strength in my life.

## Certificate of Originality

I hereby declare that this submission titled "An efficient hybrid classification model for Heart disease prediction" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name:MAAHAM MUNSIF

Student Signature: 

# Acknowledgments

Glory be to Allah (S.W.A), the Creator, the Sustainer of the Universe. Who only has the power to honor whom He pleases, and to abase whom He pleases. Verily no one can do anything without His will. From the day, I came to NUST till the day of my departure, He was the only one Who blessed me and opened ways for me, and showed me the path of success. There is nothing that can pay back for His bounties throughout my research period to complete it successfully.

**Maaham Munsif**

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation for the Research . . . . .	6
1.2	Problem Statement . . . . .	7
1.3	Research Objectives . . . . .	7
1.4	Research Questions . . . . .	8
1.5	Solution Statement . . . . .	9
1.6	Major Contributions . . . . .	10
1.7	Research Significance . . . . .	12
1.8	Thesis Structure . . . . .	14
<b>2</b>	<b>Literature Review</b>	<b>16</b>
2.1	Feature Selection in Heart disease prediction . . . . .	16
2.2	Classification in Heart disease prediction . . . . .	20
2.3	A hybrid of Feature Selection and Classification in Heart Disease Prediction	22
<b>3</b>	<b>Design and Methodology</b>	<b>31</b>
3.1	Dataset Description . . . . .	31
3.2	Framework of Proposed Model . . . . .	32
3.2.1	Data Pre-processing . . . . .	33
3.2.2	Feature Selection . . . . .	34
3.2.3	Classification Using Hybrid Model SVM-CNN . . . . .	35

## CONTENTS

3.3	Proposed Methodology . . . . .	35
3.3.1	Problem Formulation Phase . . . . .	35
3.3.2	Data Collection Phase . . . . .	38
3.3.3	Proposed Solution Phase . . . . .	39
3.3.4	Results and Validation Phase . . . . .	41
3.3.5	Evaluation Method and Criteria . . . . .	41
<b>4</b>	<b>Implementation</b>	<b>44</b>
4.1	Proposed Solution . . . . .	44
4.2	Components of proposed Solution . . . . .	45
4.2.1	Input Phase . . . . .	45
4.2.2	Feature selection Phase . . . . .	45
4.2.3	Classification Phase . . . . .	47
4.3	Experimental setup: . . . . .	50
4.3.1	Experiment NO.1 Using UCI Dataset 1: . . . . .	50
4.3.2	Experiment NO.2 Using Z-Alizadeh Sani Dataset 2: . . . . .	51
4.3.3	Experiment NO.3 Using Cardiovascular Disease Dataset 3: . . . . .	52
<b>5</b>	<b>Results and Discussions</b>	<b>54</b>
5.1	Feature Selection Using GA on Datasets: . . . . .	54
5.2	Comparison of datasets: . . . . .	56
5.3	Confusion Matrix and AUR-ROC Curve: . . . . .	57
5.4	Accuracy Comparison: . . . . .	61
5.5	Comparison of Precision, Recall, and F1-score: . . . . .	63
<b>6</b>	<b>Conclusion and Future Work</b>	<b>65</b>



# List of Figures

1.1	Mapping among Research Objectives, Research Questions, and Chapters	9
1.2	Thesis Structure . . . . .	15
3.1	Workflow of Methodology . . . . .	33
3.2	proposed methodology . . . . .	36
4.1	Architecture of Proposed Model . . . . .	45
5.1	Comparison among datasets . . . . .	56
5.2	Confusion Matrix of Heart Disease uci dataset . . . . .	58
5.3	Confusion Matrix of Heart Disease Z-Alizadeh Sani dataset . . . . .	58
5.4	AUC-ROC Cruve . . . . .	59
5.5	Confusion Matrix of Cardio Dataset . . . . .	60
5.6	AUC-ROC Curve Cardiovascular . . . . .	60
5.7	Accuaracy Comparision of heart disease uci . . . . .	61
5.8	Accuaracy Comparision of Z-Alizadeh Sani . . . . .	62
5.9	Accuaracy Comparision Cardiovascular Disease Dataset . . . . .	63

# List of Tables

2.1	Existing Feature Selection Techniques . . . . .	19
2.2	Existing Techniques of Classification . . . . .	22
2.3	Existing Hybrid Techniques . . . . .	25
3.1	Datasets Description . . . . .	32
5.1	Feature Selection Using GA: Total Features, Selected Features, and their names . . . . .	55
5.2	Comparison of Proposed and Existing Technique . . . . .	64

# Abstract

Heart disease prediction is a critical task in healthcare, aiming to identify individuals at risk and enable timely intervention. In this study, we propose a novel approach that combines a genetic algorithm for feature selection with a hybrid SVM-CNN model (GA-SVM-CNN) for heart disease prediction. The approach is evaluated on three diverse datasets: UCI, Z-Alizadeh Sani, and Cardiovascular Disease Dataset. First, the genetic algorithm is employed to select the most informative features from the datasets, reducing dimensionality and eliminating irrelevant or redundant features, and selecting the most appropriate features. Next, the hybrid SVM-CNN model is trained using the selected features, leveraging the strengths of both techniques for accurate prediction. The performance of the GA-SVM-CNN approach is evaluated using three benchmark datasets. On the UCI dataset, the approach achieves an impressive accuracy of 98%, indicating its effectiveness in accurately predicting heart disease. On the Z-Alizadeh Sani dataset, the approach achieves an accuracy of 97%. On the Cardiovascular disease Dataset, the approach achieves an accuracy of 86%. The high accuracy achieved by the GA-SVM-CNN approach demonstrates its efficacy in heart disease prediction across different datasets. The combination of the genetic algorithm's feature selection and the hybrid SVM-CNN model's predictive power contributes to superior performance. These results underscore the potential of this approach in supporting personalized healthcare solutions and improving patient outcomes.

## CHAPTER 1

# Introduction

Cardiovascular disease is a critical medical condition that ranks as the third leading cause of death worldwide, resulting in approximately 18 million deaths annually, according to a report by the World Health Organization [1]. The increasing burden of this disease underscores the urgent need to develop accurate predictive models that can identify individuals at risk of developing this deadly condition. Early detection and prevention of heart disease are crucial for improving patient health outcomes and reducing healthcare costs. Early diagnosis and treatment can help reduce the risk of heart disease and save lives. However, detecting heart disease at an early stage is notoriously challenging.

Heart disease is a significant global health issue, accounting for a substantial burden of morbidity and mortality worldwide [1]. These diseases encompass various conditions affecting the heart and blood vessels, such as coronary artery disease, myocardial infarction (heart attack), heart failure, and arrhythmias.

Heart disease has a significant impact on both individuals and healthcare systems. It contributes to a decreased quality of life for patients, leading to physical limitations, reduced productivity, and increased healthcare utilization. Moreover, heart disease poses a considerable economic burden. These costs encompass expenses related to hospitalization, medications, diagnostic procedures, and long-term management. The significance of early detection and prediction of heart disease cannot be overstated.

Early identification of individuals at risk or those who already have the disease allows for timely interventions, preventive measures, and targeted treatments, leading to improved patient outcomes and reduced healthcare costs [2].

Early prediction of heart disease plays a vital role in improving patient outcomes, pre-

venting complications, and reducing healthcare costs. By identifying individuals at risk or those who already have the disease, healthcare providers can implement timely interventions, preventive measures, and personalized treatment plans to effectively manage heart disease and promote cardiovascular health [3].

Traditional risk factors are well-established factors that have long been recognized as contributing to the development and progression of heart disease. These factors provide valuable information for assessing an individual's likelihood of developing heart disease. Here are some key traditional risk-factors associated with heart disease: "age, gender, family history, smoking, high blood pressure, etc". [4, 5, 6, 7].

Machine learning algorithms have demonstrated promising results in heart disease prediction. Studies have utilized algorithms such as "logistic regression", "decision trees", "random forests", "support vector machines", and "neural networks" to develop prediction models that accurately identify individuals at risk of heart disease [8, 9, 10].

Researchers have explored the integration of various data sources, including electronic health records, genetic information, wearable devices, and medical imaging data, to enhance the accuracy of heart disease prediction models. Combining multiple data types has shown promise in capturing a more comprehensive view of an individual's health status and improving prediction accuracy [10, 11].

Researchers have also focused on developing hybrid Machine Learning prediction models along with feature selection algorithms to get better features from the dataset and improve the accuracy and performance of the model.

Based on research published in renowned journals like *Circulation* and *Circulation: Cardiovascular Quality and Outcomes*, it has been identified that individuals experiencing heart disease commonly exhibit various symptoms. These include chest pain or discomfort, known as angina, along with shortness of breath, cold sweat, feelings of nausea, and episodes of lightheadedness or sudden dizziness. These findings highlight the importance of recognizing and understanding these symptoms, as they can serve as crucial indicators of potential heart-related issues [5].

Heart disease symptoms can vary and may present in subtle ways. Extensive research conducted at the University of Michigan Health System indicates that individuals are also prone to encounter symptoms such as shortness of breath, nausea, back or jaw pain, and unexplained fatigue, which can be indicative of a heart attack. These findings

emphasize the significance of recognizing a broader range of symptoms associated with heart disease [6]. They found that people delay seeking treatment due to these differences in symptoms. By the time people arrive at the emergency room, the heart muscle has already sustained significant damage.

There are several risk factors associated with heart disease that have been widely documented. These include high blood pressure, elevated cholesterol levels, diabetes, obesity, an unhealthy diet, physical inactivity, smoking, excessive alcohol consumption, a family history of heart disease, and advancing age. These factors have been extensively studied and are recognized by the medical community as significant contributors to the development of heart disease. However, many people have no symptoms at all in the early stages of heart disease. The only way to confirm a diagnosis is through diagnostic tests such as an electrocardiogram, stress test, cardiac catheterization, or coronary angiogram.

Early detection of heart disease and preventative screenings can help save lives. There are common signs that are widely recognized. These include chest pain or discomfort, shortness of breath, cold sweat, feelings of nausea, dizziness, and lightheadedness. It is crucial to note that while these symptoms are frequently observed, individual experiences may vary.

Machine learning has shown great promise in predicting the risk of heart disease, enabling early intervention and personalized treatment plans. However, there are challenges in developing accurate heart disease prediction models, such as the complexity and heterogeneity of the data. The current predictive models used by healthcare professionals often rely on traditional approaches such as logistic regression or decision trees [12].

However, these approaches are limited in their accuracy, as they tend to rely on a single set of features and metrics that can be used to accurately predict heart disease risk. As such, there has been an increased interest in hybrid machine learning models that can combine features from multiple sources such as medical history, lifestyle behaviors, and physiological data for better prediction accuracy. In recent years, there has been a notable emphasis on the development of hybrid classification models in the field of heart disease prediction. These models aim to enhance accuracy and performance by leveraging the strengths of multiple machine learning algorithms. Through the integration of various algorithms, these hybrid models can effectively handle complex patterns and

extract valuable insights from large and diverse datasets. The objective of such studies is to advance the accuracy and overall performance of heart disease prediction, leading to improved diagnostic and prognostic capabilities. [13].

To address the limitations associated with individual algorithms and achieve improved outcomes, researchers have been focusing on integrating various techniques within hybrid models for heart disease prediction. By combining the strengths of decision trees, artificial neural networks, and support vector machines, these models aim to provide more accurate and reliable results. The demand for enhanced predictive models in the field of heart disease has been steadily increasing. The integration of multiple predictive models into a hybrid approach has the potential to create a robust and accurate prediction system.

In heart disease prediction, the hybridization of feature selection and classification techniques involves combining two essential steps. The first step is the identification of relevant features from a pool of potential predictors, while the second step entails the utilization of a classification algorithm to predict the presence or absence of heart disease based on the selected features. The primary objective of feature selection is to reduce dimensionality, eliminate redundant or irrelevant features, and retain the most informative variables that significantly contribute to the prediction task. On the other hand, classification algorithms utilize the selected features to construct a model capable of effectively discriminating between individuals with and without heart disease. The integration of feature selection and classification techniques offers numerous advantages in heart disease prediction. Most notably, it enhances the accuracy of prediction models by focusing solely on the most informative features. By doing so, these hybrid approaches can improve the efficiency and effectiveness of heart disease prediction systems.

Integrating feature selection and classification techniques in heart disease prediction offers several advantages that contribute to improved model performance and interpretability. By eliminating irrelevant or redundant variables, the selected features create a more concise representation of the data, reducing noise and enhancing the model's ability to capture relevant patterns and relationships. As a result, the prediction performance is significantly improved, with higher sensitivity, specificity, and overall accuracy. Moreover, hybrid approaches incorporating feature selection and classification techniques enhance the interpretability of heart disease prediction models. Feature selection

helps identify the most important clinical, demographic, and physiological factors associated with heart disease. This information provides valuable insights into the underlying risk factors and mechanisms, enabling clinicians and researchers to gain a better understanding of the disease. It can also guide targeted interventions and interventions tailored to individual patients. Additionally, the selected features contribute to the development of more interpretable models that offer explanations for their predictions. This promotes transparency and fosters trust in the decision-making process.

The hybridization of feature selection and classification techniques holds significant promise in improving the accuracy and interpretability of heart disease prediction models. By identifying relevant features and utilizing effective classification algorithms, these hybrid approaches contribute to more accurate risk assessment, personalized treatment plans, and improved patient outcomes. Further advancements in this field will continue to refine and optimize the hybrid techniques, leading to more robust and clinically applicable models for heart disease prediction. Hybrid classification models are a promising approach for improving the accuracy and robustness of heart disease prediction. By combining the strengths of different machine learning algorithms and integrating other techniques, such as feature selection and data augmentation, we can develop more accurate and effective heart disease prediction models, enabling early intervention and personalized treatment plans.

The development of advanced hybrid machine learning models for heart disease prediction holds significant potential for healthcare providers. By leveraging these models, providers can enhance their understanding of the contributing factors to this condition and effectively identify individuals at high risk. This knowledge opens up new opportunities for earlier intervention and treatment of this life-threatening disease. As heart disease remains one of the leading causes of mortality globally, the integration of artificial intelligence techniques in the form of hybrid machine learning models presents an exciting avenue for researchers. These models can aid in predicting and preventing heart disease, ultimately improving patient outcomes.

In our study, we have put forward a novel hybrid technique that combines a genetic algorithm for feature selection and a hybrid approach of SVM-CNN for heart disease prediction. The primary goal of this research is to identify the most relevant features from the dataset using a genetic algorithm and subsequently enhance the overall perfor-



mance of the prediction model. The experimental results demonstrate that our proposed model achieves higher accuracy by leveraging the most crucial and informative features for heart disease prediction. This study contributes to the field by showcasing the effectiveness of our hybrid technique in improving the accuracy and performance of heart disease prediction models.

## 1.1 Motivation for the Research

The motivation behind this research work is to prevent extremity of it. And other reasons are listed as follows:

- To reduce the possibility of death by detecting a patient's conditions at an early stage. One of the most deadly disease around the globe is heart disease. By developing accurate prediction models, we aim to identify people at high risk due to the cause of heart disease, allowing for early interventions and preventive measures. Early detection and treatment can significantly reduce death rates associated with heart disease.
- Lower the cost of treatment. By focusing on prediction research, strive to develop effective risk assessment tools that can be incorporated into routine medical practice. This proactive approach has the potential to improve healthcare resource allocation, optimize patient care, and reduce healthcare costs associated with heart disease.
- Heart disease prediction can help to inform public health policy, as well as inform doctors and patients about their risk in order to make more informed decisions about preventative care.
- Heart disease often develops gradually over time, and early stages may be asymptomatic. Prediction models play a crucial role in identifying individuals at an elevated risk of developing heart disease even before symptoms manifest. This enables timely interventions, such as lifestyle changes (e.g., diet, exercise), medication management, and counseling, which can significantly reduce the progression of the disease and improve overall outcomes.
- The availability of large-scale health data sets, advancements in computational

power, and sophisticated machine-learning algorithms have opened up new opportunities for heart disease prediction research. By leveraging these technologies, researchers can analyze vast amounts of patient data, identify hidden patterns, and develop more accurate prediction models, thereby enhancing our understanding of the disease and refining risk assessment strategies.

In summary, the motivation behind research on heart disease prediction lies in the potential to improve public health, reduce mortality rates, facilitate personalized medicine, enable early interventions, and leverage technological advancements for more accurate risk assessment and preventive strategies.

## 1.2 Problem Statement

In the prediction and diagnostics of heart diseases, prediction is subject to a number of restrictions such as accurate parameters, feature selection, scales, and conditions, these factors can directly impact the accuracy and efficiency of the heart disease prediction process [1]. While considering the importance of the prediction process of cardiac diseases, in the past few years, researchers around the world have presented numerous methods and solutions. These methods mainly rely on feature selection techniques and on hybrid approaches [14]. However, the intended models and approaches lack consideration of the optimal and fittest solution for the prediction process. Furthermore, the feature selection algorithm in [15] ignores the important attributes of the dataset. The attributes of data and prediction algorithms have a direct impact on the overall performance of the system [15]. Moreover, the existing prediction approaches have the problem of analyzing the accuracy and efficiency of the system and of the feature selection technique which increase the complexity of the system and shows biasness.

## 1.3 Research Objectives

Research objectives set forth the precise benchmarks that must be met in order to accomplish a given research goal. This study's major goal is to highlight the following things.

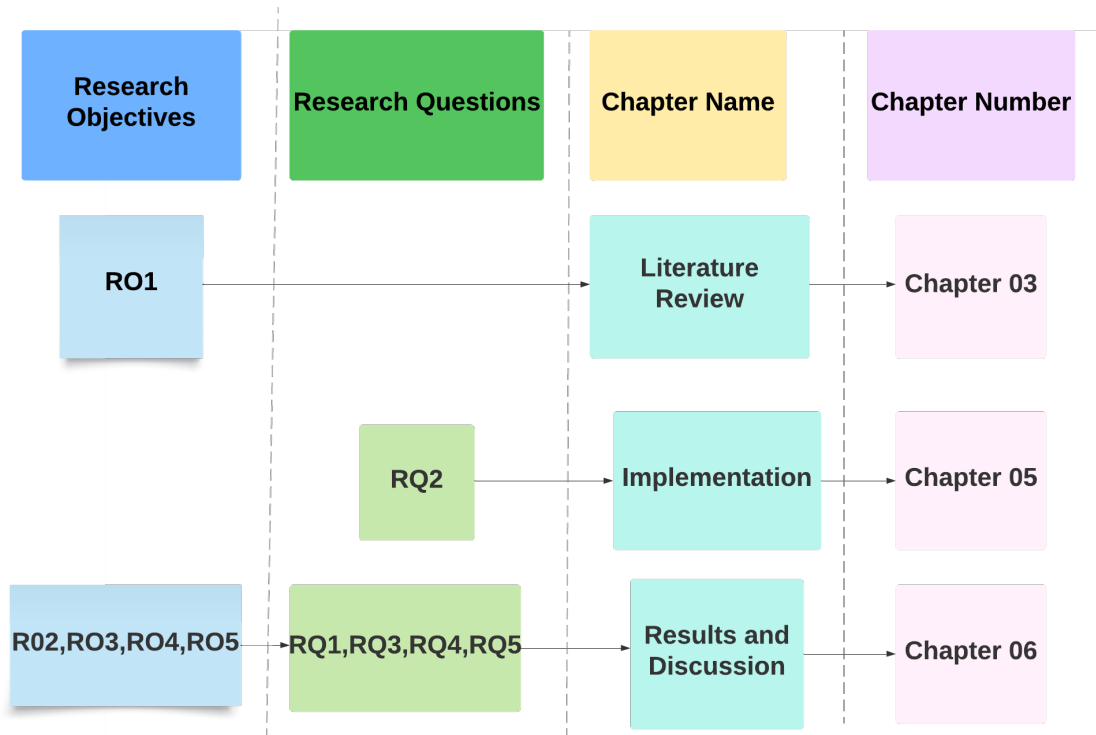
1. To identify the techniques being used for heart disease prediction from literature

- through a detailed literature review. (Chapter 3)
2. To recognize the most informative features for predicting heart disease by applying a genetic algorithm for feature selection on the three datasets (UCI, Z-Alizadeh Sani, and Cardiovascular Disease Datasets). (Chapter 6)
  3. Evaluate the performance of the "SVM-CNN" model with genetic algorithm-based feature selection in predicting heart disease across the three datasets in terms of accuracy, precision, recall, and AUC-ROC. (Chapter 6)
  4. Compare the performance of the SVM-CNN model with genetic algorithm-based feature selection to other commonly used machine learning algorithms for heart disease prediction on the three datasets. (Chapter 6)
  5. Assess the impact of varying sample sizes on the performance of the SVM-CNN model with genetic algorithm-based feature selection. (Chapter 6)

## 1.4 Research Questions

A list of research questions addressing the problems has been developed based on the study objectives. For the research goal to be accomplished, these questions must be addressed. Mapping of Research objectives and research questions along with their chapters are illustrated in 1.1. The following research questions are planned:

1. Which features selected by the genetic algorithm have the highest predictive power for heart disease across the UCI dataset, Z-Alizadeh Sani dataset, and cardiovascular dataset? (Chapter 6)
2. How does the performance of the heart disease prediction model utilizing SVM-CNN vary when different subdivisions of features selected by the GA are used? (Chapter 5)
3. Can the combined SVM-CNN model use features selected by the genetic algorithm to outperform other machine-learning algorithms commonly used for heart disease prediction on the UCI dataset, Z- Alizadeh Sani dataset, and cardiovascular dataset? (Chapter 6)



**Figure 1.1:** Mapping among Research Objectives, Research Questions, and Chapters

4. What is the comparative performance of the SVM-CNN model with feature selection using the genetic algorithm on each dataset in terms of accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC)? (Chapter 6)
5. How does the performance of the SVM-CNN model with genetic algorithm-based feature selection vary when considering different sample sizes within the UCI dataset, Z-Alizadeh Sani dataset, and cardiovascular dataset? (Chapter 6)

## 1.5 Solution Statement

The proposed solution has three main phases. "Input phase", "Feature Selection phase", and "classification and prediction phase". In the "input phase", three datasets from the repository are chosen. The input phase focuses on data preprocessing. The datasets in the input phase are passed to the second phase which is the process of FS. In the FS phase, a genetic algorithm is used. The FS process will return the set of selected features based on multiple factors and criteria. Now, the set of the selected feature are passed

to the last phase known as classification and prediction. In the "classification phase", the selected features are passed to the classification model. In the classification and prediction process, a hybrid model SVM-CNN is used to classify the selected features and data based on factors and criteria and then predict the results. CNN will improve the prediction and accuracy of the results which are obtained from the classification phase. In last our model is compared with other previous models. We can formulate the solution statement as: The solution is divided into three main phases, the first two phases focus on data preparation, preprocessing, and feature selection such that the biological implausibility of the dataset is not harmed and then training hybrid ML models (SVM-CNN) with optimal performance.

## 1.6 Major Contributions

This research aims to provide an efficient solution. Critical analysis has been performed on the existing literature on heart disease prediction and classification. An effective heart disease prediction model is proposed based on DL and ML models. The disease classification will be based on the features selected from the dataset. Classification algorithms will be used for classifications of heart diseases. The proposed model in this study offers several advantages when compared to current and previous techniques for heart disease prediction. The contributions of this research can be categorized into three main steps:

- In this study, I used three different state-of-the-art datasets namely, the UCI, Statlog, and Z-Alizadeh Sani. Different preprocessing techniques have been performed such as data cleaning, standardization, normalization, etc.
- A feature selection technique called the genetic algorithm is employed in this study to identify and select important features while preserving the biological plausibility and originality of the dataset. Genetic algorithms are optimization methods inspired by natural selection and genetics. When utilized for feature selection in heart disease prediction, the genetic algorithm aids in the identification of the most informative and relevant features from a vast pool of potential predictors. By selecting these important features, the algorithm enhances the efficiency and effectiveness of the prediction model, mitigating computational complexity and

potential overfitting concerns. This approach ensures that the dataset's integrity and biological significance are maintained.

- The processed dataset is then used to train a hybrid model SVM-CNN to detect whether heart disease is present in the patient or not. The combination of SVM and CNN leverages the strengths of both algorithms to enhance prediction accuracy for heart disease. SVM is a potent machine learning algorithm that excels at handling high-dimensional data and nonlinear relationships. CNN, on the other hand, is particularly effective in capturing spatial dependencies and patterns in data, making it suitable for analyzing complex data. By combining the two, we can leverage the strengths of both algorithms and achieve improved prediction accuracy for heart disease.

The utilization of a genetic algorithm for feature selection enhances the generalization capability of the prediction model. By selecting the most informative features, the model becomes more robust and less prone to overfitting, a phenomenon where the model performs well on training data but struggles to generalize to new data. Overfitting can be mitigated by reducing the number of features and focusing on the most relevant ones. This approach ensures that the model captures the underlying patterns present in heart disease data, enabling better generalization to new, unseen samples. Moreover, the combination of genetic algorithm-based feature selection with SVM-CNN models offers valuable insights into the important features that contribute to heart disease prediction. By identifying and selecting relevant features, researchers gain a better understanding of the underlying factors and risk factors associated with heart disease. This interpretability can aid in medical decision-making, as clinicians can prioritize and focus on the most influential factors in diagnosing and treating heart disease. The combination of feature selection and SVM-CNN prediction models allows for personalized heart disease risk assessment. By identifying the most relevant features for each individual, the model can provide tailored risk predictions, taking into account individual risk factors, genetics, and other relevant data. This personalized approach can guide targeted interventions, lifestyle modifications, and treatment strategies, optimizing patient outcomes and improving the efficacy of preventive measures.

## 1.7 Research Significance

The research conducted in this study holds significant importance and impact in both academic and industrial domains. The utilization of a genetic algorithm for feature selection in heart disease prediction, along with the integration of SVM-CNN for the prediction process, carries implications that extend to both academia and industry.

- Industry:
  - Improved Risk Assessment: Enhanced risk assessment is a notable outcome of employing accurate prediction models that integrate genetic algorithms, Support Vector Machines (SVM), and Convolutional Neural Networks (CNN). Such models have the potential to significantly improve risk assessment capabilities within the healthcare industry. By effectively combining these techniques, healthcare providers can identify individuals who are at a heightened risk of developing heart disease. This enables the implementation of proactive interventions and the formulation of personalized treatment plans. The utilization of genetic algorithms facilitates the selection of the most relevant and informative features, optimizing the prediction model's performance. SVM and CNN, on the other hand, provide powerful classification algorithms that can effectively analyze and interpret the selected features. Through this combined approach, the prediction model can accurately assess an individual's risk of heart disease. The integration of these advanced techniques in risk assessment holds immense value for the healthcare industry. By identifying individuals at high risk, healthcare providers can initiate preventive measures such as lifestyle interventions, targeted screenings, or early medical interventions. This can lead to a reduction in the incidence and severity of heart disease, resulting in improved patient outcomes and a more efficient allocation of healthcare resources.
  - Personalized Healthcare: The integration of genetic information and personalized risk assessment can support the development of personalized healthcare strategies. This approach allows for tailored interventions, lifestyle modifications, and medication management, leading to improved patient outcomes and reduced healthcare costs.

- Decision Support Tools: The proposed approach can serve as a decision-support tool for clinicians and healthcare practitioners. By providing interpretable insights into the relevant features contributing to heart disease prediction, the approach can guide medical professionals in making informed decisions regarding diagnosis, treatment, and preventive measures.
- Technological Advancement: The research leverages advanced algorithms and techniques, including genetic algorithms, SVM, and CNN, demonstrating the potential of incorporating cutting-edge technologies in the field of heart disease prediction. Industry stakeholders can benefit from the technological advancements highlighted in the research by exploring their integration into healthcare systems and platforms.
- Academia:
  - Advancing Knowledge: This research contributes to the academic understanding of heart disease prediction by exploring the effectiveness of combining advanced algorithms, such as genetic algorithms, SVM, and CNN, to improve prediction accuracy. It provides insights into the optimal use of these techniques and their application to heart disease datasets.
  - Methodological Contributions: The research introduces a novel approach that combines feature selection using GA with the power of SVM-CNN models. This methodology can serve as a reference for future studies in heart disease prediction and inspire further advancements in feature selection and model combination techniques.
  - Comparative Analysis: By comparing the performance of the proposed approach to other commonly used algorithms, the research offers a benchmark for evaluating and selecting appropriate algorithms for heart disease prediction. This comparison helps in identifying the strengths and weaknesses of different techniques and provides a basis for future algorithm selection in related studies.
  - Dataset Evaluation: The use of diverse datasets, such as the UCI dataset, Z-Alizadeh Sani dataset, and the cardiovascular dataset, allows for comprehensive evaluation and comparison of the proposed approach across different data characteristics. This evaluation helps understand the generalizability



and robustness of the method in various scenarios.

## 1.8 Thesis Structure

The remaining sections of the thesis are organized as follows. Chapter 2 provides a comprehensive literature review, presenting an overview of existing research and studies relevant to the topic of heart disease prediction using genetic algorithms and SVM-CNN models. This chapter serves to establish the theoretical and empirical foundation for the research.

Chapter 3 focuses on the methodology employed in the study. It delves into the specific components and techniques utilized, including the genetic algorithm for feature selection and the integration of SVM-CNN models for prediction. The chapter elucidates the rationale behind the chosen methodologies and explains their implementation in detail.

The implementation of the proposed model is presented in Chapter 4. This chapter outlines the practical aspects of translating the theoretical framework into a functional system. It discusses the software or tools employed, data preprocessing procedures, and the technical aspects of model implementation.

Chapter 5 is dedicated to presenting the results and conducting a thorough discussion of the proposed model. It analyzes the outcomes obtained from the experiments or simulations, providing insights into the model's performance, accuracy, and effectiveness. The chapter also includes a detailed discussion of the findings in relation to existing literature and research objectives.

Finally, Chapter 6 concludes the thesis by summarizing the key findings and contributions of the study. It discusses the implications of the research, highlights limitations, and suggests avenues for future work. This chapter provides closure to the thesis and offers recommendations for further research and advancements in the field.

A detailed breakdown of the thesis structure is depicted in [1.2](#).

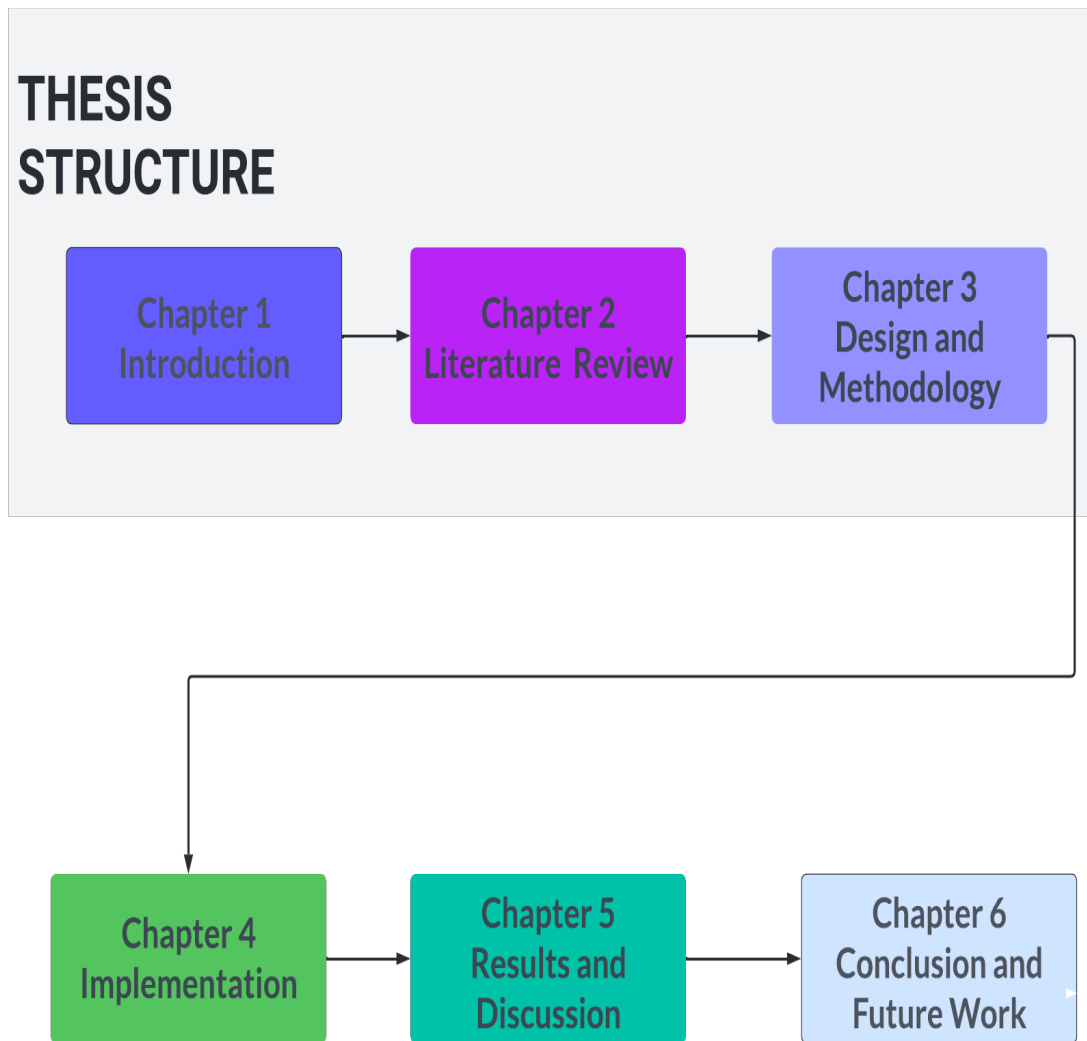


Figure 1.2: Thesis Structure

# Literature Review

Major organ that pumps blood throughout the body is the heart, which is situated right beneath the breastbone. The heart is essential to our health because it pumps blood throughout the body. [16]. Heart disease is a complex and widespread condition that poses significant challenges in healthcare, particularly in cardiology, detecting cardiac disease accurately and early is of utmost importance [17]. Researchers have developed various intelligent systems that can improve heart disease prediction [18]. The prediction of heart disease is a crucial undertaking in the healthcare field, with the goal of dividing individuals into distinct categories based on their likelihood of developing cardiovascular issues. The use of classification models for predicting heart disease can aid in the early identification of potential problems, risk assessment, and the creation of intervention plans, ultimately leading to positive patient outcomes. Within the realm of heart disease prediction, a variety of machine learning algorithms and approaches have been utilized, including decision tree, logistic regression, support vector machines (SVM), random forests, and artificial neural networks. These algorithms are valuable tools for classifying individuals into different heart disease categories. The classification process typically involves two key steps: the selection of relevant features and the training of the model.

## 2.1 Feature Selection in Heart disease prediction

In the context of heart disease prediction, feature selection plays a vital role in identifying the most relevant attributes or risk factors that contribute to the development of the disease. Several common characteristics are often considered in this process, including

age, sex, blood pressure, cholesterol levels, smoking habits, family history, and past medical problems.

The utilization of feature selection methods to enhance the accuracy of heart disease prediction models is addressed in a research paper [19]. The paper provides valuable insights into the evaluation of various feature selection techniques and their impact on the performance of machine learning classification algorithms [19]. A comprehensive analysis is conducted on a diverse set of filter, wrapper, and embedded feature selection methods, encompassing the Chi-squared test, information gain, relief-F, correlation-based feature selection, recursive feature elimination, and principal component analysis. The study [19] explores the effectiveness and comparative performance of these feature selection techniques, shedding light on their potential benefits in the context of heart disease prediction. By evaluating the impact of each method on the performance of machine learning algorithms, the research contributes to understanding the importance of feature selection and its role in improving the accuracy of heart disease prediction models. This allows for a thorough comparison of the performance of different techniques [19]. The experimental results in [19] show that feature selection techniques can significantly gain accuracy and reduce the dimensionality of heart disease prediction models. [19] The random forest classifier combined with principal component analysis achieved the highest accuracy of 88%. With the right feature selection, machine learning models can be made more effective at predicting cardiac disease, according to the study cited in [19]. To validate and further develop the findings, additional research using larger and more varied datasets is required. In Table, current feature selection methods are contrasted.

### 2.1.

The paper [20] provides valuable insights into suitable machine-learning approaches for heart disease prediction and got up to 95 percent accuracy. However, expanding the scope of the study to include larger datasets and more modern classifiers could further improve the findings, overall, the research demonstrates the potential of machine learning to assist in medical diagnosis and highlights directions for future work in this area [20]. In a study conducted by the researchers [21], the objective was to assess the performance of different feature selection techniques and classification algorithms for heart disease prediction. The paper focused on evaluating the effectiveness of these techniques in accurately predicting the presence or absence of heart disease. By employing a range of feature selection methods and classification algorithms, the researchers

aimed to identify the most effective combination for heart disease prediction. The study involved analyzing the performance metrics of the selected techniques, such as accuracy, precision, recall, and F1-score, to determine their predictive capabilities. The findings of this research provide valuable insights into the performance of various feature selection techniques and classification algorithms specifically applied to heart disease prediction. By evaluating and comparing the performance of these techniques, the study contributes to the development of more accurate and reliable models for predicting heart disease. The author [21] found that the feature subset selected by the backward feature selection achieves the highest accuracy of 88.52% with the decision tree classifier. The accuracy of computer-aided diagnosis systems has been successfully improved through the application of machine learning. Using pertinent characteristics selected using a variety of feature-selection techniques. Experimentally evaluates the performance of models created using machine learning techniques [22].

PCA, Chi-squared testing, Relief-F, and symmetrical uncertainty have all been used to examine four widely used heart disease datasets in order to produce different feature sets. Then, a variety of classification algorithms have been used to create models that are then evaluated in order to determine the optimum feature combinations, in order to improve the accuracy of heart condition predictions [22]. A machine learning algorithm is put out in [23] to forecast cardiac disease based on a patient's medical history. Through the use of a genetic algorithm for feature selection prior to classification, the scientists hope to increase the precision and effectiveness of heart disease prediction. The proposed model,[23], was able to predict cardiac disease with a fair amount of accuracy. The prediction accuracy increased from 72.9% to 91.9% by using the evolutionary algorithm for feature selection prior to classification [23].

This shows the effectiveness of the genetic algorithm in selecting the most relevant features for the classification model. The authors [24] used feature selection methods like information gain and chi-square to select the most relevant features from the original dataset that contribute the most to the prediction accuracy. Ensemble learning techniques like Bagging, Ada-Boost, and Random Forest were applied to the selected features. This led to improved prediction accuracy compared to using a single classification model. This study [12] used 6 different machine learning classification algorithms on 2 different data sets to conduct an experimental investigation of the model's effectiveness. The relevant attributes are chosen using the integrated feature selection approach.

Two different heart disease datasets from Kaggle were used to test six machine learning classification algorithms, including Random-Forest, Decision-Tree, K-Nearest-Neighbor, Support-Vector-Machine, Gaussian-Naive-Bayes, and Logistic-Regression, as well as two feature selection methods, Ridge-Regression and Lasso-Regression. The Random-Forest classification algorithm performed best in both datasets, with a 93.25 percent accuracy in the second [12] dataset and a 77.25 percent accuracy in the first.

Multiple heart disease datasets are utilized for experimentation analysis in this paper [25], which focuses on feature-selection strategies and algorithms and demonstrates the accuracy increase. Decision-Tree, Logistic-Regression, Logistic-Regression SVM, Nave-Bayes, and Random-Forest-algorithms are employed as feature selection strategies utilizing the Rapid Miner tool, and improvement is demonstrated in the results by demonstrating the accuracy [25]. In [26], the authors suggest a hybrid feature selection strategy for diagnosing heart disease combining Support Vector Machines (SVM) and Grey Wolf Optimisation (GWO). The goal is to narrow down the field of potential predictors to the most pertinent aspects in order to increase the precision of heart disease diagnosis [26]. The GWO algorithm and SVM are combined in the suggested approach. The social behaviour of grey wolves served as the basis for the nature-inspired optimisation algorithm known as GWO [26]. It is employed to look for the best possible subset of features to use in order to improve classification accuracy. The chosen features are used using SVM, a well-known machine learning technique, to identify heart disease.

**Table 2.1:** Existing Feature Selection Techniques

Authors	Techniques	Datasets	Accuracy	Limitations
M. G. Javed et.al [26]	Combination of GWO algorithm with SVM. SVM performs feature selection. GWO is used to find an optimal subset of features.	Cleveland and Stat-log	89.83%	Small datasets are used. when large datasets will be used accuracy decreases.

Bashir et.al[25]	Information Gain (IG), Chi-Squared (CHI2), and Relief-F. The selected features are then used as input for the Random Forest classification algorithm.	Cleveland Dataset	84.85%.	It only evaluates the model on a single dataset, making it difficult to determine how well it would generalize to other scenarios.
Pemmaraju,et al.2022[12]	Lasso and Ridge for FS and 6 classification algorithms	BRFSS source of heart disease, Kaggle by ALEX TEBoul	dataset-1 77.25% dataset-2 93.25%	Can improve accuracy by using hybrid models.
Lakshmanarao et al. 2021 [24]	information gain and chi-square are used. Ensemble learning techniques Bagging, AdaBoost, and Random Forest.	UCI, Kaggle	Dataset-1 87% Dataset-2 91%	small dataset used. Using a larger dataset could potentially improve the prediction performance further.
Reddy, et al. 2023 [27]	correlation-based feature selection method.	Cleveland+Statlog	Classifier performance has improved.accuracy 97%.	The complexity of the model is higher than before.

## 2.2 Classification in Heart disease prediction

After the features have been chosen, a classification model is trained using a labelled dataset in which each occurrence is linked to a class label indicating whether or not cardiac disease is present. Using the UCI heart illness dataset, one such work [13] suggested a hybrid classification model for heart disease prediction. ANN is trained using swarm optimisation, which optimises its weight and biases to increase performance by up to 89.83%. The system can effectively learn complex patterns and dependencies in the data, leading to accurate prediction of heart disease. The computational complexity of the swarm optimization algorithm can be high. This may limit the scalability and

real-time applicability of the system. The interpretability of the model and the ability to explain the underlying reasoning for predictions may be challenging with complex neural network architecture. The study showed that the hybrid model could improve the accuracy and robustness of heart disease prediction.

Another work [28] suggested a hybrid classification model for heart disease prediction employing LR, SVM, KNN, naive Bayes, and DT classifiers, as well as ANN to forecast the development of heart disease. The SVM 85 percent UCI Cleveland heart disease dataset produced the highest accuracy. The study demonstrated that the hybrid model could predict cardiac illness at the cutting edge of technology.. However, they fail to address the impact of different hyperparameter settings on each supervised learning model and the interplay between them, also, the paper does not explore the effects of combining multiple supervised classification models together for better performance. A study [29] used a filter algorithm for reducing low-value features and GA is used to select high-value features. A stacked-Genetic algorithm using the Ada-boost technique. Statlog dataset used from UCI Repository. This model can offer an improved approach to diagnosing heart diseases with greater accuracy and higher efficiency. Potential inaccuracies due to unequal feature importance and the lack of model evaluation and data assumptions. The author [21] found that the feature subset selected by the backward feature selection achieves the highest accuracy of 88.52% with the decision tree classifier. Existing Techniques of classification are shown in table 2.2.

This study [12] used 6 different machine learning classification algorithms on 2 different data sets to conduct an experimental investigation of the model's effectiveness. The integrated feature selection method is used to choose the relevant characteristics. Two different heart disease datasets from Kaggle were used to test two feature-selection techniques, Ridge-Regression and Lasso-Regression, as well as six machine learning classification algorithms, Random-Forest, Decision-Tree, K-Nearest-Neighbour, Support-Vector-Machine, Gaussian-Naive-Bayes, and Logistic-Regression. In both datasets, the Random Forest classification method had the highest accuracy, with a 77.25 percent accuracy on the first dataset and a 93.25 percent accuracy on the second [12]. The main achievements of this paper [30] are its accuracy and effectiveness in predicting heart disease. The researchers tested the system on three datasets of heart disease patients and found that the random forest classifier achieved an average accuracy of around 90% when optimized using the evolutionary approach [30].



**Table 2.2:** Existing Techniques of Classification

Authors	Techniques	Datasets	Accuracy	Limitations
Nandyet al. [13]	Swarm optimization is used to train ANN, optimizing its weight and biases to improve performance.	UCI Dataset	89.83%	limits the scalability and real-time applicability of the system.
Mohan, et al. 2019 [31]	hybrid random forest with a linear model (HRFLM)	UCI datasets	88.7%.	Only small datasets are used.
Kolukisa et al. 2023 [16]	Heart disease is predicted by using 10 different classifiers, six are single classifiers and one is an ensemble classifier with four differ	Cleveland, Statlog, and Z-Alizadeh Sani	91.7%	more complex and computationally expensive.
Kolukisa et al. 2023 [28]	LR, SVM, KNN, naive Bayes, and DT classifiers, ANN is used to predict the occurrence of heart disease.	Cleveland Dataset	85%	fails to address the impact of different hyper parameter.

### 2.3 A hybrid of Feature Selection and Classification in Heart Disease Prediction

The hybridization of feature selection and classification techniques in heart disease prediction involves combining two distinct steps: identifying the most relevant features from a pool of potential predictors and employing a classification algorithm to predict the presence or absence of heart disease based on the selected features. Feature selection aims to reduce dimensionality, eliminate redundant or irrelevant features, and retain the most informative variables that contribute significantly to the prediction task [32, 8]. Classification algorithms, on the other hand, utilize the selected features to build a model that can effectively discriminate between individuals with and without heart disease.

The integration of feature selection and classification techniques offers several advantages in heart disease prediction. First, it enhances the accuracy of prediction models by focusing on the most informative features. By eliminating irrelevant or redundant variables, the selected features provide a more concise representation of the data, reducing noise and improving the model's ability to capture relevant patterns and relationships. This leads to improved prediction performance, including higher sensitivity, specificity, and overall accuracy. Furthermore, hybrid approaches incorporating feature selection and classification techniques enhance the interpretability of heart disease prediction models [33, 34]. Feature selection helps identify the most important clinical, demographic, and physiological factors associated with heart disease. This enables clinicians and researchers to gain insights into the underlying risk factors and mechanisms, facilitating a better understanding of the disease and potentially guiding targeted interventions. The selected features can also aid in developing more interpretable models that provide explanations for their predictions, promoting transparency and trust in the decision-making process [35, 36, 37].

In [16] heart disease is predicted by using 10 different classifiers, six are single classifiers and one is an ensemble classifier with four different variations. In [16] there are 3 different types of datasets are used i.e. Cleveland, Statlog, and Z-Alizadeh Sani, for feature selection two ensemble approaches, are proposed, the Exhaustive ensemble feature selection approach and the Probabilistic ensemble feature selection approach, the best accuracy scores are obtained as 91.78%, 85.47%, for the Z-Alizadeh Sani and Cleveland using MLP classifiers and 86.66% for the Statlog dataset using KNN classifier. In [16] there are 10 different types of classifiers and two different types of feature selection methods are used that make the system more complex and computationally expensive. In [38] Bhanu Prakash Doppala et al proposed a hybrid machine learning approach for heart disease prediction using Genetic Algorithm with radial basis function (GA-RBF), using dataset Cleveland was taken from UCI Repository and this model achieve an accuracy of 94.20% while considering 9 features. In [38], the proposed technique only a single dataset is used, this may limit the generalizability of the findings to other datasets, however, there are several limitations that should be taken into consideration, including the results being based on a single dataset, further research should explore the performance of the proposed system on other datasets and evaluate the effectiveness of the feature selection methods to identify other relevant features for coronary diseases.

In [39], a dimensional reduction strategy for feature selection is put out. Chi-square with PCA is also utilised to enhance the performance and prediction of the suggested model in addition to dimensional reduction. From the subset of 72 features, the dimensional reduction approach chooses three groupings of features. The proposed technique in [39] is tested on small sample size and the performance may be compromised due to the high computational complexity of algorithms. In [32], a comparative analysis is performed among 3 proposed algorithms to measure the effectiveness and accuracy of the system. The methods which are used in [32], are confusion matrix, precision, specificity, sensitivity, and F1 score. The UCI datasets are used for the prediction of disease.

Zhang, Dengqing, et al. explore the use of an embedded feature selection method [40], combined with a deep neural network, to predict the presence of heart disease, the main contribution of this paper is in demonstrating the effectiveness of an embedded feature selection method using linear-SVC and deep neural network for heart disease prediction, the authors concluded that the model was successful, with a prediction accuracy of 97%, and that it was superior to other classification algorithms, despite the success of the model, there are some limitations, firstly, the paper only focused on one particular dataset, and therefore the accuracy of the model in other datasets is not known, additionally, the authors do not discuss how sensitive the model is to imbalanced data.

A hybrid machine learning model is proposed in [41] to predict heart disease, the primary contribution of this paper is to analyze two predictive models, Decision Tree (DT) and Random Forest (RF), and combine the models to create a hybrid model. The hybrid model (RF-DT) in [41] is then applied to clinical data taken from the UCI repository named as Cleveland dataset to assess its accuracy and performance, using a hybrid approach 88.6% accuracy was obtained which is higher than DT and DF. While considering [41], there is no pre-processing methods are mentioned that can be used for data normalization, feature selection, and handling missing values, therefore, the performance of this model is compromised, we can say that this model does not accurately predict heart disease. Hybrid approaches for disease prediction and classification are very popular nowadays because of their performance and accuracy [42]. An advantage is taken in [14], by proposing a hybrid approach using an artificial bee algorithm and a genetic algorithm. Results show that employing and combing these algorithms can effectively improve the accuracy and performance of the feature selection process. Table 2.3 demonstrates the overview of existing techniques with limitations [26].

**Table 2.3:** Existing Hybrid Techniques

Authors	Techniques	Datasets	Accuracy	Limitations
V. K. Sudha et al. 2023[43]	weight by SVM for feature selection CNN+LSTM for prediction	Cleveland dataset from UCI Repository	89%	It only evaluates the model on a single dataset.
Girish S. Bhavakar et al. 2022 [44]	Hybrid RNN and LSTM for prediction	Cleveland dataset from UCI Repository	95%.	small dataset, No pre-processing techniques are defined.
Jafar Abdollahi et al. 2022 [29]	Reduce and eliminate low-value features with the filter algorithm. Select high-value features with the genetic algorithm.	Statlog dataset from UCI Repository	97%	Potential inaccuracies due to unequal feature importance
Senthil Murugan Nagarajan et al. 2022 [31]	genetic-based crow search algorithm for feature selection and DNN for classification	Statlog dataset from UCI Repository	88%	Accuracy is limited by relying on a small sample set of data.
BP Doppala et al. 2021 [38]	GA-RBF	Cleveland dataset from UCI Repository	94.20%	single dataset.
Gárate-Escamila et al. 2020 [39]	Dimensionality reduction method and principal component analysis	Cleveland dataset from UCI Repository	98%.	Raw data computed lower results and would require greater.
Karunakaran Velswamy et al. 2021 [15]	Modified bee algorithm for feature selection and SVM, Naïve Bayes, and KNN for classification	UCI Repository	85%	With fewer features of the dataset

Dr. M. Kavitha et al. 2021 [41]	Hybrid DT-RF for prediction	Cleveland dataset from UCI Repository	88%	no pre-processing
Bindu M G et al. 2020 [14]	Artificial bee colony and Genetic algorithm for feature selection	Cleveland dataset from UCI Repository	91.8%	Limited in its data used to evaluate the effectiveness.

Using structured data, the CNN-UDRP algorithm predicts the risk of diseases. KNN algorithm with the Naive Bayes method to forecast cardiac disease. Compare the outcomes of the KNN and Naive Bayes algorithms, NB has a higher accuracy of 82% than KNN [45]. By providing the input of patient records, which aid in understanding the level of illness risk prediction, and able to provide an accurate disease risk prediction. The likelihood of developing heart disease is estimated to be low, high, or medium. Due to this technique, disease risk prediction may be accomplished with little effort and expense [45]. In [31], Machine Learning techniques are used for the prediction and classification of heart diseases. UCI dataset is pre-processed and then machine learning techniques are applied to process data, which learn from the data and predict the heart disease based on values on data [5, 8]. In [15], a modified bee algorithm with classifiers is proposed.

The modified bee algorithm finds the optimum features from the subset of features. Results reveal that the overall performance and accuracy of the system are improved. However, the complexity of the model is higher than before. The correlation-based feature selection method is used in [27] which selects the best principle components and hyperparameters for the classification models.

Machine learning has been effectively used to increase the accuracy of computer-aided diagnosis systems. Using relevant traits chosen utilising a number of feature-selection methods. examines through experimentation how well models built using machine learning methods operate. Relief9, symmetrical uncertainty, Chi-squared testing, and principal component analysis have all been applied to analyse four frequently used heart disease datasets and yield various feature sets. Then, a variety of classification algorithms have been used to create models that are then evaluated in order to determine the optimum feature combinations, in order to improve the accuracy of heart condition

predictions [22]. A general disease prediction model proposed in [46], is based on the patient's symptoms. K-Nearest Neighbour (KNN) and Convolutional Neural Network (CNN) machine learning algorithms are used to accurately forecast disease. Disease symptoms dataset needed for disease prediction. To handle the problem of data consistency and irregularity a new convolutional neural network (CNN) based on a multimodal disease risk prediction algorithm is proposed in [12].

A floating window with adaptive size for feature elimination (FWAFE) is used for feature selection in [18], while for prediction two types of the neural network are used i.e. ANN and DNN, two types of the hybrid model are proposed FWAFE-ANN and FWAFE-DNN that uses data from UCI repository named as Cleveland dataset, FWAFE-ANN achieved 91.11% and FWAFE-DNN achieved 93.33% accuracy. However, this research paper [18] does not discuss the performance of the method, or any evaluation results, it also does not include any discussion of how the proposed method may be used in a clinical setting, further research is needed to evaluate the accuracy of the proposed method, additionally, more research is needed to identify additional risk factors that could further improve the accuracy of the model and to explore other potential applications of the proposed method.

A unique feature selection method called random forest-feature sensitivity and feature correlation (RF-FSFC), which integrates feature sensitivity and feature correlation to increase prediction accuracy, was proposed in [47]. The [47] paper's major contributions are a thorough analysis of the feature selection procedure, a discussion of the significance of feature selection in the prediction of heart disease, and the introduction of a novel algorithm that combines sample-based sensitivity and correlation-based feature selection techniques. The paper [47] also highlights some of the limitations of the proposed technique, first, it relies heavily on existing knowledge regarding heart-disease-related features which may not be available in all contexts, Secondly, the limited availability of data used in the experiments may not be sufficient to capture the full spectrum of correlations between features and heart disease outcomes, Finally, the evaluation of the algorithm relies mostly on benchmark datasets and does not reflect the potential performance of the technique in a real-world setting, despite the limitations this paper obtained 86.141% accuracy that is higher than other models.

This paper focuses on the role of feature selection in the prediction of heart disease,

especially in the context of machine learning algorithms as it discusses the following. Feature selection is an important part of predictive systems, particularly when it comes to predicting the onset of heart disease [48]. This is especially relevant for predictive systems utilising machine learning algorithms, as the quality of the predictions depends largely on the quality of the features.

Specifically, the paper [48] discusses methods for identifying the most relevant features for prediction, as well as methods for eliminating redundant features, the Cleveland Heart dataset from the UCI repository was used in the study, and the classification algorithms Naive-Bayes, Random-Forest, Extra-Trees, and Logistic-regression that were given with chosen features using LASSO and Ridge-regression were examined, the accuracy achieved 94.92% using Lasso-regression and ridge-regression, the paper does not discuss the potential for false-positive results when using predictive systems with feature selection, it is also not clear as to how the feature selection process is implemented in the predictive systems, or what methods are used to ensure effective selection. Finally, the authors [48] do not discuss the potential for overfitting when using predictive systems with feature selection.

The paper [28] presents an overview of the performance of several supervised classification models, specifically logistic-regression, support-vector-machine, k-nearest-neighbor, naive-Bayes and decision-tree classifiers, and ANN on predicting the occurrence of heart disease, it provided a comprehensive performance analysis of each model with respect to accuracy, precision, recall, and other metrics. It concluded that the support vector machine achieved the highest performance at 85%, and 87% in terms of accuracy and precision, followed by a k-nearest neighbor and the naive Bayes classifier, additionally, the paper found that the decision tree classifier had the highest recall, but drew the conclusion that the model's performance was not as accurate when compared to the other supervised classification models.

Despite its noteworthy findings, this paper [28] has several limitations, For example, the paper fails to address the impact of different hyperparameter settings on each supervised learning model and the interplay between them, also, the paper does not explore the effects of combining multiple supervised classification models together for better performance.

In [43] examines the potential for a hybrid CNN and LSTM network to predict heart

disease, this hybrid approach combines the CNN's powerful feature extraction capabilities and the LSTM's memory-based information processing to produce a predictive model that performs better than either approach alone, the authors evaluated their model on a medium dataset, using binary classification, and showed it was capable of achieving significant improvement compared to other state-of-the-art models. The main contributions of this paper [43] are the exploration of a hybrid CNN-LSTM network for heart disease prediction, which provides an improved prediction with 89% accuracy compared to traditional approaches; and the use of different visualization techniques to understand the model's decision-making process. Also, this paper [43] has some limitations, for example, it only evaluates the model on a single dataset, making it difficult to determine how well it would generalize to other scenarios, additionally, the model could be improved by exploring different hyperparameters, network architectures, and other strategies to improve results.

The objective of Chaurasia, Vikas, et al.'s study [49], is to create an efficient method for predicting heart disease using an ensemble technique based on sequential feature selection. The authors use the gradient-boosting-based sequential feature selection technique to choose the most appropriate features from the training dataset, and they also suggest an effective method for creating an ensemble of base classifiers. The evaluation is restricted to a specific dataset, which makes generalising the results challenging, and the sensitivity of the model has been found to be lower than other existing methods, which may limit its practical applications. Additionally, the authors do not evaluate the performance of the model in other scenarios or other datasets, which may limit its practical applications.

An innovative way of predicting heart-disease risk using a hybrid GA and PSO approach is presented in this paper [50]. It combines the strengths of both approaches to produce more accurate predictions than either method alone, it does this by utilizing the strengths of each method, such as GA's ability to search larger spaces and PSO's ability to reduce unwanted oscillation in small spaces, also there are some potential limitations and gaps of this approach include potential costs associated with its use and the comparison to other methods of prediction [50].

The Multilayer Perceptron (MLP) and Enhanced Brownian Motion based on Dragonfly Algorithm (EBMDA) techniques, as well as optimised unsupervised techniques for fea-



ture selection, are combined to propose a novel approach for predicting heart disease. Based on chosen features, the proposed approach uses EBMDA to optimise the MLP model parameters. The findings demonstrate that the MLP-EBMDA method surpasses a number of other cutting-edge techniques in terms of precision, sensitivity, specificity, and AUC-ROC [51]. For precise cardiac disease prediction, a Hybrid Truncate Swarm Algorithm and Ensemble Deep Learning (TSA-EDL) technique is used [17]. The suggested hybrid TSA-EDL technique displays promising results in accurately predicting heart illness, which is a vital task for early identification and appropriate treatment [17]. The tasks involved in heart disease prediction are pre-processing, clustering, and classification.

In conclusion, the hybrid machine learning model is an efficient approach to heart disease prediction. The experimental results demonstrate the potential of integrating diverse machine learning algorithms for improved predictive performance. The model can assist doctors in early diagnosis and treatment planning for heart disease patients. Hybrid classification models are a promising approach for improving the accuracy and robustness of heart disease prediction. By combining the strengths of different machine learning algorithms and integrating other techniques, such as feature selection, we can develop more accurate and effective heart disease prediction models, enabling early intervention and personalized treatment plans. These experiments have shown the promise of machine learning, yet predicted performance can still be enhanced. Most studies have relied on a single machine-learning algorithm or a simple hybrid of two algorithms. Integrating diverse algorithms that learn different patterns can provide a more robust heart disease prediction model.

## CHAPTER 3

# Design and Methodology

This chapter provides information about the research process, the research methods used, and the data processing methods. The approach used in scientific research is composed of a few basic steps. Each study differs in some way due to the period of time, atmosphere, conditions, and location in which it is being conducted. The research process is a common and well-known phenomenon. This study serves as a guide for carrying out important research. Throughout this study process, the investigation passed through numerous significant phases for the researcher to draw a conclusion. A modification in one phase of the research process will also affect the other stages because each phase is interconnected. Researchers must evaluate the other stages after introducing a modification in one phase to ensure that changes are reflected in all phases.

### 3.1 Dataset Description

For experimentation and implementation, publically available benchmark datasets are used. There are three types of datasets used for the evaluation and validation process. Table 3.1 demonstrates the name of datasets, number of instances, number of features, and labels. Dataset 1 named "UCI Heart Disease Dataset" is a combination of 4 databases named as follows: Cleveland, Hungarian, Switzerland, and Long Beach heart disease datasets containing 76 features and 920 instances related to various clinical and demographical factors. The dataset focuses on predicting the presence or absence of heart disease in patients. It includes features such as age, sex, cholesterol levels, chest pain type, resting electrocardiographic results, the presence of major blood vessels, etc.

Dataset 1 is available on the UCI repository. This heart disease dataset is widely used in research and provides valuable resources for developing and evaluating predictive models for heart disease diagnosis and risk assessment. In 920 instances, 463 are normal people and 457 are abnormal people i.e. suffered from heart disease.

Dataset 2 named "Z-Alizadeh Sani Dataset" contains 56 attributes and 303 instances. The "Z-Alizadeh Sani" dataset contains clinical and laboratory attributes of patients who experienced coronary artery disease (CAD) and underwent coronary artery bypass graft (CABG) surgery. The data were collected from patients referred to the Tehran Heart Center in Iran. The dataset includes 303 instances (patients) and 55 attributes. These attributes encompass clinical features such as age, gender, and various risk factors associated with heart disease, as well as laboratory measurements such as blood tests, echocardiographic parameters, etc. In 303 instances, 88 are normal people and 216 are heart disease patients.

Dataset 3 named Cardiovascular Disease Dataset contains 11 attributes and 70,000 instances, making it the largest Heart Disease Dataset available for the research purpose. For researching and making predictions on the prevalence of cardiovascular illness, researchers frequently utilize the "Cardiovascular Disease" dataset available on Kaggle. The "Cardiovascular Disease" dataset includes a collection of clinical and demographic attributes of individuals, with the goal of predicting the presence or absence of cardiovascular disease.

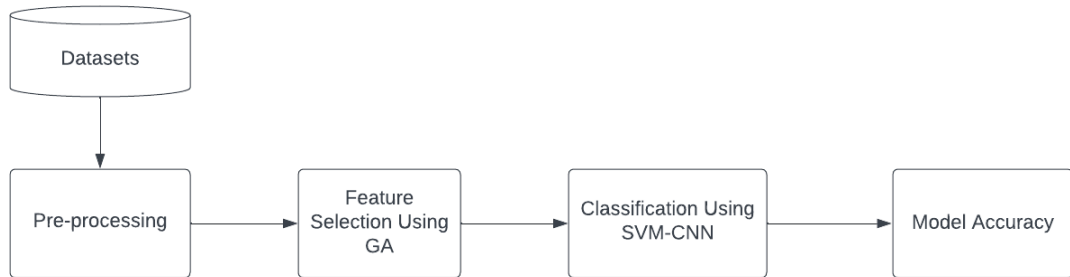
**Table 3.1:** Datasets Description

Datasets	Instances	Features	Labels
UCI Heart Disease Dataset	920	76	4
Z-Alizadeh Sani Dataset	303	56	2
Cardiovascular Disease Dataset	70,000	11	2

## 3.2 Framework of Proposed Model

The proposed model mainly focuses on identifying heart disease with better accuracy using selected features. The performance of the ML algorithm has been tested on three datasets UCI Dataset, Z-Alizadeh Sani Dataset, and Cardiovascular Dataset. The work-

flow of the model has been implemented in three stages shown in 3.1. Pre-processing of data, Feature Selection, and the last step is the classification and performance evaluation of heart disease.



**Figure 3.1:** Workflow of Methodology

### 3.2.1 Data Pre-processing

Various pre-processing techniques are applied to each dataset to enhance the proposed model's capabilities and performance. Following Data pre-processing techniques are applied.

**Removing Null Values** These null values have a detrimental effect on the accuracy and performance of any machine learning method. Therefore, it is essential to eliminate null values from the dataset before utilising any machine learning approach.

**Handling Outliers** Identify and handle outliers in the dataset. Outliers can be addressed by removing them if they are data entry errors, by applying statistical techniques such as truncation, or by replacing outliers with more representative values.

**Label Encoding** Label encoding is a common encoding technique for categorical information. This approach assigns a unique number to every label based on its alphabetical order. The label encoding gives each occurrence in the chosen datasets a special identification number.

**Handling Imbalanced Classes** Dataset 1 suffers from imbalanced class distribution, we use SMOTE technique to balance the classes and improve model performance.

**Data splitting** Data is split into the ratio of 70:30. 70% is used for model training while the remaining 30% is used for testing purposes. The test set is used to evaluate the model performance on unseen data.

**Feature scaling** Feature Scaling is applied to the selected datasets. The Feature scaling is carried out by using standard scalers. In feature scaling the trained data is transformed into values of -1 to 1.

### 3.2.2 Feature Selection

After the pre-processing, pre-processed datasets are passed towards the genetic algorithm which is one of the evolutionary algorithms. The genetic algorithm with CNN finds the optimal solution based on the criteria of the AUC-ROC curve and F1 score.

**Genetic Algorithm** A genetic algorithm is a heuristic optimization technique that mimics the process of natural selection to solve optimization problems. In the context of feature selection, the genetic algorithm aims to find the subset of features that yields the best performance on a given task (e.g., classification or regression).

The genetic algorithm starts with a population of randomly generated feature subsets. Each subset is represented as a binary string where each bit corresponds to a feature, and the value of the bit indicates whether the feature is selected (1) or not (0). The fitness of each subset is evaluated using a fitness function that measures the performance of the corresponding feature subset on the task at hand. In the above code, the fitness function is based on the F1-score.

The genetic algorithm then proceeds through a series of iterations or generations. In each generation, the algorithm selects the fittest individuals (i.e., feature subsets with the highest fitness scores) and applies genetic operators to create new offspring. The genetic operators include crossover (which combines two parent feature subsets to create a new offspring) and mutation (which randomly flips some bits in a feature subset). The offspring are then added to the population and the process repeats until a stopping criterion is met (e.g., a maximum number of generations is reached).

Through the repeated application of genetic operators, the algorithm explores the space of possible feature subsets and gradually converges to a subset that optimizes the fitness

function. In the context of feature selection, this subset corresponds to the subset of features that yields the best performance on the given task.

### 3.2.3 Classification Using Hybrid Model SVM-CNN

Set up the SVM model, including the choice of kernel function (e.g., linear, polynomial, or radial basis function) and hyperparameter selection. ii. Training: Train the SVM model using the training dataset. Define the architecture of the CNN model, including the arrangement and number of convolutional, pooling, and fully connected layers. Train the CNN model using the training dataset, optimizing the model's weights through forward and backward propagation. Determine how the output features or representations from the SVM and CNN models are combined or integrated to create a hybrid model.

Specify the method of integrating the outputs of the SVM and CNN models, such as concatenation or weighted fusion. Train the hybrid model using the combined features from the SVM and CNN models, optimizing the model's parameters. Perform hyperparameter tuning for the hybrid model to find the optimal settings, including learning rate, regularization parameters, or fusion weights. Model Evaluation: Evaluate the hybrid model's performance using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score, or (AUC-ROC) on the testing dataset.

## 3.3 Proposed Methodology

The proposed methodology is divided into four phase's problem formulation, Data collection, proposed solution, and Validation. Figure 3.2 illustrates the proposed methodology of the current study.

### 3.3.1 Problem Formulation Phase

The problem formulation phase in heart disease prediction involves defining the problem, specifying the objectives, and determining the scope and requirements of the predictive modeling task. A problem in the classification and prediction of cardiac disorders is formulated at this level. It was challenging to pinpoint a problem. First, topics in computer science and artificial intelligence are picked. The machine learning field is where the data is gathered. Then choose the heart disease feature selection and prediction.

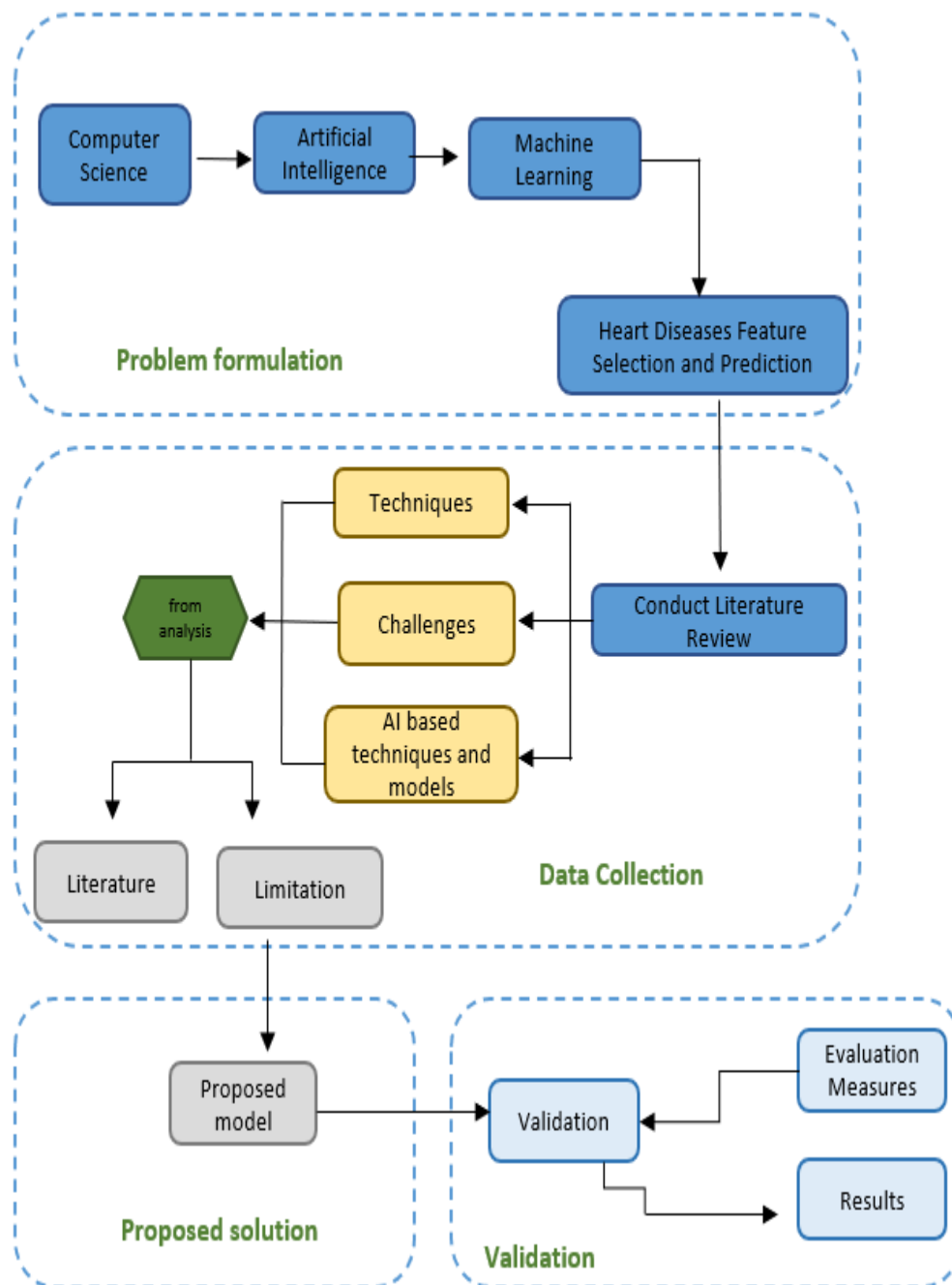


Figure 3.2: proposed methodology

Here are the key aspects to consider during this phase.

**Problem Definition** Clearly define the problem at hand, which is the prediction of heart disease in individuals. Determine the specific task type, which is binary classification (presence or absence of heart disease). State the ultimate goal of the prediction task, which is to accurately identify individuals at risk of heart disease for early intervention and treatment.

**Data Understanding** : Identify the sources from which the heart disease dataset is obtained, such as the UCI repository, Kaggle, or a proprietary database. We obtained data from three resources 2 are from UCI Repository and one is from Kaggle. Determine the availability of relevant features and attributes necessary for heart disease prediction, such as demographic information, medical history, and diagnostic test results. Explore the dataset's size, structure, format, and any limitations or biases that may be present.

**Objectives and Evaluation** Specify the evaluation metrics that will be used to assess the predictive models, such as accuracy, precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC). Set performance goals for the heart disease prediction task, aiming for high accuracy and reliable classification results. Consider the ethical implications of the predictive models, ensuring that the models are fair, unbiased, and avoid discrimination.

**Domain Knowledge and Expertise** Seek guidance from domain experts or medical professionals to gain a deeper understanding of heart disease risk factors, symptoms, and relevant medical knowledge. Collaborate with experts to identify the most informative features that have a significant impact on heart disease prediction, helping to guide the feature selection process.

**Scope and Limitations** Determine the number of instances (patients) and features available in the heart disease dataset, considering its adequacy for model training and evaluation. Consider the availability of computational resources, time limitations, and any other constraints that may impact the development and deployment of the predictive models. Ensure compliance with legal and privacy regulations when handling



sensitive medical data, ensuring the appropriate anonymization and protection of patient information. The problem formulation phase sets the foundation for the heart disease prediction project, providing a clear understanding of the problem, objectives, data requirements, and limitations. This phase helps to align the project goals with the available resources and expertise, ensuring a well-defined and focused approach to the subsequent stages of data pre-processing, model selection, and evaluation.

### 3.3.2 Data Collection Phase

Data gathering followed an approach based on a review of the literature. To start, a review of the literature was done to collect information from it. Then picked out several methods and fixes from the literature. Different limitations were identified following the literature research. The most popular and reliable way for gathering and analyzing data is the literature review process. Prior to reporting, the pertinent data is evaluated and analyzed during the data-gathering phase. The process of collecting data is continuous; it begins with the gathering and finishes with reporting after evaluation and analysis.

Explore publicly available databases, such as the UCI Machine Learning Repository, Kaggle, or government health repositories, that provide heart disease-related datasets. Identify the key variables and features required for heart disease prediction, including demographic data, medical history, lifestyle factors, and diagnostic test results (e.g., blood pressure, cholesterol levels, and ECG readings). Determine the specific target variable, such as whether heart disease is present or not, to be predicted based on the available data. Decide on the level of granularity needed for the data, such as individual patient records or aggregated data at a certain time interval. Ensure compliance with data licensing agreements, terms of use, and legal requirements associated with accessing and using the chosen data sources. Establish agreements or partnerships with data providers, ensuring the ethical and secure handling of patient data while adhering to privacy regulations.

Extract the relevant data from the identified sources, using appropriate techniques such as SQL queries, web scraping, or data download from public repositories. Merge and integrate data from multiple sources, ensuring that the data is consistent, accurate, and compatible for subsequent analysis. Perform data cleaning tasks such as handling missing values, correcting data inconsistencies, removing duplicates, and addressing

outliers or noisy data points. Verify the integrity and quality of the collected data through checks, validations, and cross-referencing with other sources or expert opinions. Anonymize or de-identify the collected data to protect patient privacy and comply with relevant regulations. Document the data sources, variables, feature descriptions, data collection process, and any transformations or pre-processing steps applied to the collected data.

Create a data dictionary or documentation specifying the meaning, format, and possible values of each variable and feature. The data collection phase is crucial for acquiring the necessary data to develop accurate heart disease prediction models. It involves identifying relevant data sources, defining data requirements, obtaining necessary permissions, integrating data, ensuring data quality, and properly documenting the collected data. A well-executed data collection phase lays the foundation for subsequent stages of data pre-processing, feature engineering, model development, and evaluation.

### 3.3.3 Proposed Solution Phase

In this stage, a solution is put up to overcome the constraints that the literature research method revealed. In the proposed solution phase, solutions to the problem or problems identified throughout the research process are to be highlighted. The whole scope of the issue is addressed in the suggested solution. Before arriving at the final decision, there are various requirements for the solutions that must be satisfied. The proposed solution for heart disease classification involves utilizing a Genetic Algorithm (GA) for feature selection and a combination of Support Vector Machine (SVM) and Convolutional Neural Network (CNN) models (referred to as SVM-CNN) for prediction. This hybrid approach aims to enhance the accuracy and robustness of heart disease classification by selecting relevant features and leveraging the strengths of both SVM and CNN algorithms.

**Genetic Algorithm for Feature Selection** A genetic algorithm is a search and optimization technique inspired by the process of natural selection. It mimics the evolution process to find an optimal solution. Feature selection using a genetic algorithm involves representing each feature subset as a chromosome and applying genetic operators such as crossover and mutation to generate new feature subsets in each generation. Fitness

evaluation is performed using the SVM-CNN model's performance as the fitness function, measuring the classification accuracy or other appropriate metrics. The genetic algorithm iteratively evolves the population of feature subsets over multiple generations, selecting the fittest individuals and promoting the survival of high-performing feature subsets.

**SVM-CNN Hybrid Model for Prediction** Support Vector Machine (SVM) is a popular machine-learning algorithm used for classification tasks. It seeks to find an optimal hyperplane that maximally separates different classes in the feature space. Convolutional Neural Network (CNN) is a deep learning architecture particularly effective for sequence data analysis, capturing spatial and temporal patterns through convolutional and pooling layers.

The proposed SVM-CNN hybrid model combines the strengths of SVM and CNN to leverage both the discriminative power of SVM in high-dimensional feature spaces and the ability of CNN to learn complex representations from raw or transformed data. The selected feature subset from the genetic algorithm is integrated into the SVM-CNN hybrid model as input, enhancing the model's ability to capture informative and discriminative features. The proposed solution involves training the SVM-CNN hybrid model using the selected feature subset and labeled heart disease data. The hybrid model is trained using appropriate optimization techniques to minimize the classification error and maximize predictive performance.

Model hyperparameters, such as kernel type and regularization parameters for SVM, as well as CNN architecture, learning rate, and the dropout rate for CNN, are tuned using techniques like grid search or random search to optimize the model's performance. The trained SVM-CNN hybrid model is evaluated using a separate test dataset, assessing its classification performance using metrics like accuracy, precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC). The proposed solution should be evaluated on multiple datasets, including external datasets, to validate its effectiveness across different populations and ensure its robustness. The proposed solution can be implemented using programming languages such as Python, utilizing libraries/frameworks such as sci-kit-learn for SVM, TensorFlow, or PyTorch for CNN, and genetic algorithm libraries for feature selection.

### 3.3.4 Results and Validation Phase

The proposed solution's findings were computed, and the process of validation was carried out using current state-of-the-art techniques. This stage will enable us to evaluate the system's efficiency and precision. This stage will reveal how well and efficiently we comprehend the issue and our solution proposal. Validating the heart disease prediction model is crucial to ensure its accuracy, reliability, and generalization to unseen data. Here are the key validation steps for heart disease prediction. Split the available labeled data into training and testing datasets. The split is 70% for training and 30% for testing. Randomize the data during the split to ensure an unbiased representation of the heart disease classes in both sets. Train the heart disease prediction model using the training dataset. Use machine learning and deep learning algorithms, SVM-CNN on the chosen approach. Set the hyperparameters based on domain knowledge.

Evaluate the trained model on the testing dataset to assess its performance. Calculate various evaluation metrics, such as accuracy, precision, recall, F1-score, or AUC-ROC, to measure the model's predictive capabilities. Ensure that the evaluation metrics are relevant to the problem's specific requirements and consider the class imbalance, if present. Conduct hyper-parameter tuning to optimize the model's performance. Analyze the model's predictions and its limitations. Interpret the learned patterns and feature importance to gain insights into the factors contributing to heart disease prediction. Report the performance metrics achieved by the heart disease prediction model on the testing and validation set. Provide a comprehensive analysis of the model's strengths, weaknesses, and limitations.

### 3.3.5 Evaluation Method and Criteria

The proposed model is compared against state of art methods from the literature as part of the evaluation process. The same dataset is used to test and train both the existing model and our model, and graphs are then created to show trends. To prevent biases, the result plotting follows the same procedures and standards as the previous solution. When evaluating the performance of a heart disease prediction model, several evaluation measures and criteria can be used to assess its effectiveness. Here are some employed evaluation measures and criteria for heart disease prediction:

**Accuracy:** Accuracy is a measure of the overall correctness of a model's predictions by calculating the ratio of successfully predicted cases to all instances. In the case of imbalanced datasets, accuracy may be biased but still provides a general understanding of the model's performance.

**Precision:** Precision quantifies the percentage of accurate positive predictions relative to all instances where a positive outcome was expected. It assesses the model's ability to reliably predict positive cases, particularly in identifying individuals with heart disease.

**Recall (Sensitivity):** Recall, also known as sensitivity, measures the percentage of correctly predicted positive examples from the dataset. It evaluates the model's accuracy in detecting individuals with heart disease while minimizing false negatives.

**F1-Score:** The F1 score is a balanced evaluation metric that considers both false positives and false negatives. It is figured out as the harmonic mean of recall and accuracy. Better performance is indicated by higher F1 scores, which offer a single value that combines recall and accuracy.

**Area-Under the Receiver Operating Characteristic Curve (AUC-ROC):** The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the model's capacity to differentiate between positive and negative instances at various classification thresholds. By plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity), the AUC-ROC determines the area under the curve. A higher AUC-ROC value indicates a better-discriminating performance of the model, approaching 1.

**Confusion Matrix:** The confusion matrix tabulates the model's predictions in relation to the actual classes. It includes metrics such as true positives, true negatives, false positives, and false negatives, enabling a comprehensive evaluation of the model's performance.

**Receiver Operating Characteristic (ROC) Curve:** The Receiver Operating Characteristic (ROC) curve visually illustrates the trade-off between sensitivity (true positive

rate) and specificity (false positive rate) at different classification thresholds. It aids in visualizing the model's discrimination ability and can assist in selecting an appropriate threshold based on the desired balance between sensitivity and specificity.

# Implementation

## 4.1 Proposed Solution

A Detailed description of the proposed model along with the details of its components is discussed in this section. The proposed Model consists of three phases. Input Phase, Feature Selection Phase, and Classification phase. Input phases consist of the data sets. The selected datasets are pre-processed to ensure the accuracy and efficiency of the proposed model. The input phase consists of three datasets having different numbers of instances and attributes. To check the performance of the proposed model it is trained and tested on different sizes of datasets. A complete description of the proposed model is presented in Figure 5.1.

In the second phase which is the feature selection phase, the datasets are given input to the evolutionary algorithm (Genetic Algorithm). The genetic algorithm selects the list of features from the inputted data set. In this process there is a list of selected features that are used for the classification phase is used. In last there is a classification phase there is a support vector machine that classifies the input based on the conditions of the selected features.

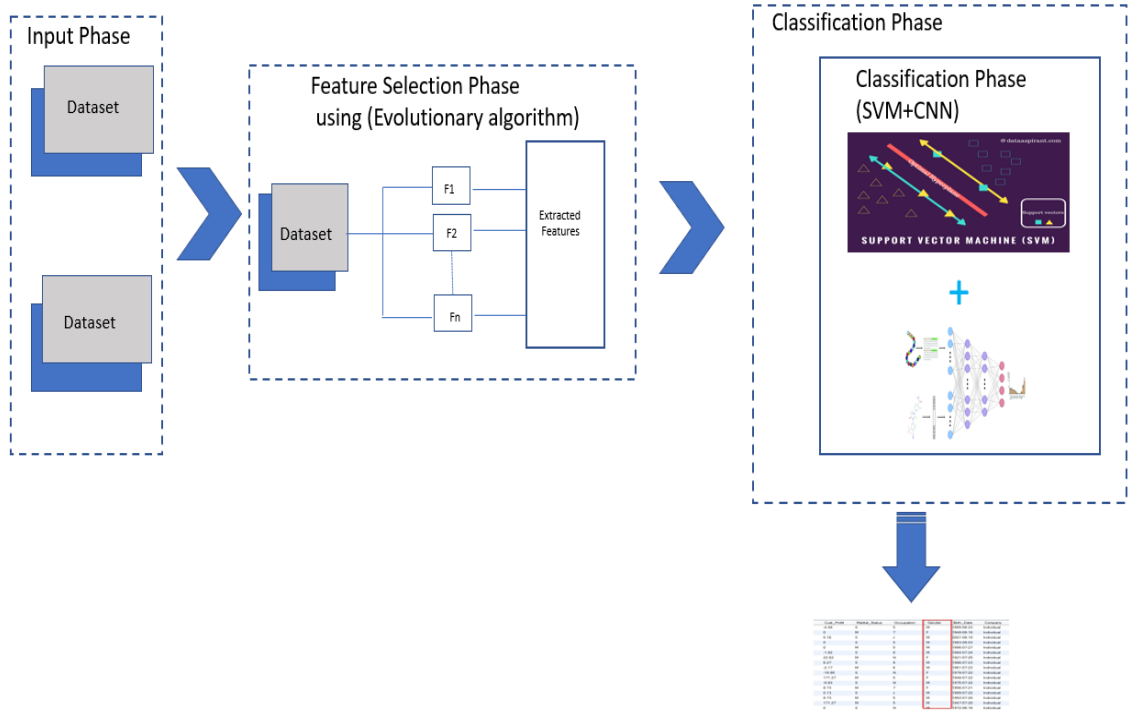


Figure 4.1: Architecture of Proposed Model

## 4.2 Components of proposed Solution

### 4.2.1 Input Phase

In the input phase datasets are selected from the publically available repositories. Pre-processing techniques are applied to the selected datasets in this phase. The applied techniques are removing null values, Label Encoding, Data splitting, and Feature scaling.

### 4.2.2 Feature selection Phase

In this phase, pre-processed datasets are passed towards the genetic algorithm which is one of the evolutionary algorithms. The genetic algorithm with CNN finds the optimal solution based on the criteria of the AUC-ROC curve and F1 score. After finding the best optimal features the list is passed to the classification phase.

**Genetic Algorithm** GA is a useful, reliable strategy and search approach. The natural selection of the test individuals used by nature to generate species served as the basis for the GA search algorithm. In that it determines the most suitable and



effective solution to a given computer issue, a genetic algorithm is comparable to natural evolution. The biological ideas of evolution and the survival of the fittest serve as the foundation of the genetic algorithm.

Because it offers a solid and reliable answer that is graded against fitness criteria, this algorithm is significantly more effective and powerful than an exhaustive search algorithm. The fitness function measures how closely a particular solution comes to being optimum. Feature selection aims to reduce the dimensionality of the data by identifying the subset of features that are most predictive or influential for the problem at hand. The genetic algorithm, inspired by the process of natural selection, iteratively evaluates and evolves a population of potential feature subsets based on their fitness or performance in solving the problem. This component helps improve the efficiency and accuracy of the solution by focusing on the most significant features and reducing noise or irrelevant information.

This method offers a decent and resilient solution that is graded against fitness criteria, making it far more powerful and efficient than an exhaustive search algorithm. A solution's closeness to optimality is measured using the fitness function. The possible solutions to the problem are represented by a set of chromosomes. A chromosome is a string of binary digits. Each digit is called a gene. The initial population can be created randomly. The Genetic Algorithm (GA) is a search and optimization technique inspired by the principles of natural evolution. It iteratively evolves a population of candidate solutions to a problem by mimicking the processes of selection, reproduction, and mutation.

**Initialization:** Start by creating an initial population of potential solutions to the problem. Each solution is represented as a set of parameters called chromosomes or genes. The population is typically generated randomly or using some heuristic.

**Evaluation:** Evaluate the fitness of each individual in the population. The fitness function measures how well each solution performs with respect to the problem at hand. The fitness function can be problem-specific and aims to quantify the quality or suitability of a solution.

**Selection:** Select individuals from the population to serve as parents for the next generation. The selection process is usually based on the fitness values of the individuals. Solutions with higher fitness have a higher probability of being selected, mimicking the concept of survival of the fittest.

**Reproduction:** Create offspring by combining the genetic material of the selected parents. This is typically done through genetic operators such as crossover and mutation. Crossover involves exchanging genetic information between two parents to produce new individuals, while mutation introduces small random changes to the genetic material to promote exploration of the search space.

**Replacement:** Replace some individuals in the current population with newly created offspring. The replacement strategy can vary, but commonly the offspring replace the least fit individuals, ensuring that the population maintains diversity and potential improvements.

**Termination:** Repeat steps 2-5 for a certain number of generations or until a termination criterion is met. Termination criteria can be based on the number of generations, reaching a satisfactory fitness level, or a predefined time limit. Or termination is also met when we analyze that the next step will overfit the data then we stop the process.

### 4.2.3 Classification Phase

In the classification phase there is a SVM which will classify the data. The classified data is helps to predict the heart dieses based on the features which are selected by the genetic algorithm.

**Support Vector Machine:** SVM, or support vector machines, is another name for a classification and regression prediction tool that makes use of machine learning theory for the best accuracy and to prevent model overfitting. Locating the hyperplane in N-dimensional space is the support vector machine's goal. For the separation of two classes, there are numerous potential hyperplanes from which to choose. Drawing the hyperplane with the greatest possible margin distance is the goal in order to classify upcoming data points. Decision boundaries called hyperplanes aid in categorizing the

data points. Support vectors are data points that relate to the sides of the hyperplane and which indicate the characteristics of the classes. The hyperplane will be a line that the attributes that belong to separate classes if the input features are two. A hyperplane is a two-dimensional plane if the input features are three. The hyperplanes become challenging when the input feature reaches three. Support vectors that are near the hyperplane have an impact on the hyperplane's position and orientation. The sigmoid function compresses the input in logistic regression into the range  $[0, 1]$ . Assign the input value a label of 1 if it exceeds the threshold value (0.5). The output of the linear function is taken into consideration while determining the class label in SVM.

A class label is assigned if the output is 1, and a different class label is assigned if the output is -1. There is a margin between threshold values, which range from 1 to -1. The margin value between the data points and the hyperplane should be maximized so that future data points can be classified more confidently. The kernel's goal is to make it possible for operations to be carried out in the input space. The data is transformed using the kernel functions, and based on these transformations, an ideal boundary between the potential outputs is discovered. We categorize the binary class data for the linearly separable data using the linear kernel. The non-linear SVM is utilized in the scenario of non-linear separable data though.

The border determined by utilizing the non-linear kernels is not a straight line in a non-linear SVM. Depending on the nature of the dataset, several kernels such as Gaussian, polynomial, sigmoid, or others, are employed in non-linear SVM.

**Convolution Neural Network** A synthetic neural network with numerous concealed layers. CNN can extract greater order of connection by using more hidden layers. The incoming and outgoing layers of a CNN are split at different levels. Whatever their size or structure, neural networks must have neurons, connections, values, bias, and functionalities. Here's a detailed explanation of how CNNs work in our model and their components:

**Input Layer:** The input layer of a CNN receives the pre-processed data.

**Convolutional Layer:** The convolutional layer is the core component of a CNN. It consists of a set of learnable filters (also called kernels) that convolve over the input

data. Each filter detects specific features, by performing element-wise multiplication and aggregation. The outcome is a feature map that draws attention to the fact that certain features were present in the input.

**Activation Function:** After each convolutional operation, an activation function is applied element-wise to introduce non-linearity. Common activation functions used in CNNs include Rectified Linear Unit (ReLU), sigmoid, or hyperbolic tangent. ReLU is the most popular choice due to its simplicity and effectiveness in preventing the vanishing gradient problem. We use ReLU in our model.

**Pooling Layer:** The pooling layer reduces the spatial dimensions of the feature maps while retaining the most salient information. It achieves this by applying operations like max pooling or average pooling within a localized region. Pooling helps in reducing computational complexity and making the learned features more invariant to small spatial variations.

**Fully Connected Layer:** The fully connected layer, also known as the dense layer, takes the output from the previous layers and connects every neuron to every neuron in the subsequent layer. It learns high-level representations by combining the learned features from previous layers. These layers are similar to those in a traditional neural network and perform classification or regression tasks.

**Dropout:** Dropout is a regularization technique commonly used in CNNs to prevent overfitting. During training, a certain percentage of neurons in the fully connected layer are randomly ignored or "dropped out." This helps in reducing interdependent learning between neurons and forces the network to learn more robust features.

**Output Layer:** The output layer of the CNN provides the final predictions or outputs. The activation function in the output layer depends on the nature of the problem being solved. For binary classification, a sigmoid function is typically used, while for multi-class classification, a softmax function is commonly used.

**A hybrid of SVM-CNN:** A hybrid approach that combines Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) for classification tasks. SVM

is a machine learning algorithm used for supervised classification, while CNN is a deep learning technique commonly used for classification. By combining the strengths of both SVM and CNN, the hybrid approach aims to leverage the discriminative power of SVM and the ability of CNN to learn hierarchical features automatically from the input data. These components give an effective means of classifying or categorizing the input data based on the selected features, utilizing the power of both traditional machine learning and deep learning techniques.

### 4.3 Experimental setup:

Python is used to implement the proposed work. With Python 3.8 installed on the Windows operating system, the Anaconda environment is used. The hardware specifications are an Intel Core i7 6th generation CPU and 8 GB of RAM. The Python code is implemented using the Pycharm tool, and the Pycharm project makes use of Anaconda as an environment. Each environment and tool used in research work is explored in detail in this section.

#### 4.3.1 Experiment NO.1 Using UCI Dataset 1:

**Pre-processing:** In experiment 1, we use dataset 1 named as UCI Dataset with 920 instances and 76 features. First of all pre-processing techniques are applied to clean the data. Start by preparing the UCI dataset, ensuring it is labeled with the corresponding class labels for each instance. Here we did labeling on UCI heart disease Dataset and it has 4 labels. We apply standardization from -1 to 1 to transform data. Transform categorical variables into numerical data using a label encoder. Then we split the data into training and testing in the ratio of 70:30. UCI data was so imbalanced, for balancing the data we use SMOTE technique to balance the data. After balancing the data feature scaling is used to transform the data, we use a standard scaler to transform data from -1 to 1.

**Feature Selection and Classification Using F1 Score:** We use a Genetic algorithm for feature selection using a population size of 50, number of generation 5, crossover probability 0.8, and probability of mutation 0.2. The fitness function is defined to calculate the fitness score of an individual. It takes an individual as an input

and returns the fitness score, which is the F1 score in this case. It selects the individual based on the individual's genes, trains a CNN model using selected features, and calculates the F1 score using the trained model and testing data. Crossover performs by randomly selecting a gene and changing its value. The main loop starts, which iterates for a specific number of generations. After the evaluation process, the best individual is selected based on its fitness score. The selected features are selected from the best individual's genes. Finally, the selected features are printed. The next step is defining a hyper-parameter to tune. We use different hyper-parameters to train our model to achieve better accuracy. We use the best hyperparameter with SVM to train the final model. We find the best hyperparameters and achieve higher accuracy. To evaluate this model we test our model and get the higher results on the basis of accuracy, precision, recall, and F1 score.

### 4.3.2 Experiment NO.2 Using Z-Alizadeh Sani Dataset 2:

**Pre-processing:** With 303 instances and 56 attributes, dataset 2—also known as the Z-Alizadeh Sani Dataset—is used in experiment 2. Data cleaning procedures are used first to prepare the data. Prepare the Z-Alizadeh Sani dataset first, making sure that each instance is labelled with the appropriate class label. The Z-Alizadeh Sani Dataset, which comprises binary class data and 2 labels, was labelled in this instance. Data is transformed via standardisation from -1 to 1. Utilise a label encoder to convert categorical variables into numerical information. Then, we divided the data 70:30 between training and testing. We utilise a normal scaler to change the data from -1 to 1 after splitting it using the data feature scaling.

**Feature Selection and Classification Using F1 Score:** For feature selection, we employ a genetic method using population size, generational length, crossover probability, and mutation probability. The fitness function is intended to determine a person's overall fitness level. It accepts a person as an input and returns the fitness score, in this case the F1 score. The individual is chosen based on the individual's genes, a CNN model is trained using a subset of characteristics, and the F1 score is computed using the trained model and test data. Crossover operates by picking a gene at random and altering its value. The main loop begins and runs for a predetermined number of generations. The top candidate is chosen following the examination procedure based

on their fitness score. The best genes of each individual are used to pick the desired qualities. The chosen features are printed off at the end. Choosing a hyper-parameter to tweak is the next step. To improve the accuracy of our model, we train it with a variety of hyper-parameters. To train the final model with SVM, we select the best hyperparameter. We increase accuracy by locating the optimum hyperparameters. We test our model and obtain the better results in terms of accuracy, precision, recall, and F1 score in order to evaluate our model.

### 4.3.3 Experiment NO.3 Using Cardiovascular Disease Dataset 3:

We use the Cardiovascular Disease Dataset, a dataset containing 70,000 cases and 11 characteristics, in experiment 3. The data is first cleaned using pre-processing procedures. Make sure the UCI dataset is ready by labelling each instance with the appropriate class label in the beginning. We labelled the Cardiovascular Disease Dataset in this instance, and it has two labels. To change the data, we standardise it from -1 to 1. Utilise a label encoder to convert category variables into numerical data. The data was then divided 70:30 between training and testing. After separating the data, we utilise feature scaling to turn the data from -1 to 1. We do this by using a conventional scaler.

**Feature Selection and Classification Using Accuracy Score:** The population size, number of generations, crossover frequency, and mutation probability are all factors we consider when choosing features using a genetic algorithm. In order to determine a person's fitness score, the fitness function is defined. It accepts an individual as input and outputs the fitness score, which in this case is the F1 score. It chooses the person based on their genes, trains a CNN model with a subset of the features, and then computes the F1 score using the trained model and test data. By picking a gene at random and altering its value, crossover operates. Start of the main loop, which repeats for a predetermined number of generations. Based on their fitness score, the top candidate is chosen following the evaluation procedure. The best genes from each person are used to choose the traits. The chosen features are then printed. Determining a hyper-parameter to tweak is the following step. To train our model more accurately, we use various hyper-parameters. To train the final model with SVM, the best hyperparameter is used. Our efforts result in greater accuracy and the best hyperparameters. We test our model to evaluate it, and the greater results on the basis of accuracy, precision, recall, and F1

score are obtained.



# Results and Discussions

The experiments are conducted in the setting and environment that were previously stated. The data sets are in numeric format. Data cleaning techniques are applied to the data before the experimentation process. The main steps which are performed during the experiment are as follows:

- Data Pre-processing techniques are applied to the existing data.
- The processed data is then used as input for the genetic algorithm, which will process data and return the list of optimal features.
- The list of optimal features is selected by using a genetic algorithm and CNN.
- The features which are selected by the genetic algorithm have the baseline of the AUC-ROC curve and F1 score.
- For all possible lists, the optimal listed feature is selected by CNN.
- A classification Mechanism using SVM is applied to the existing list, and the behaviours and trends are recorded in terms of plots.

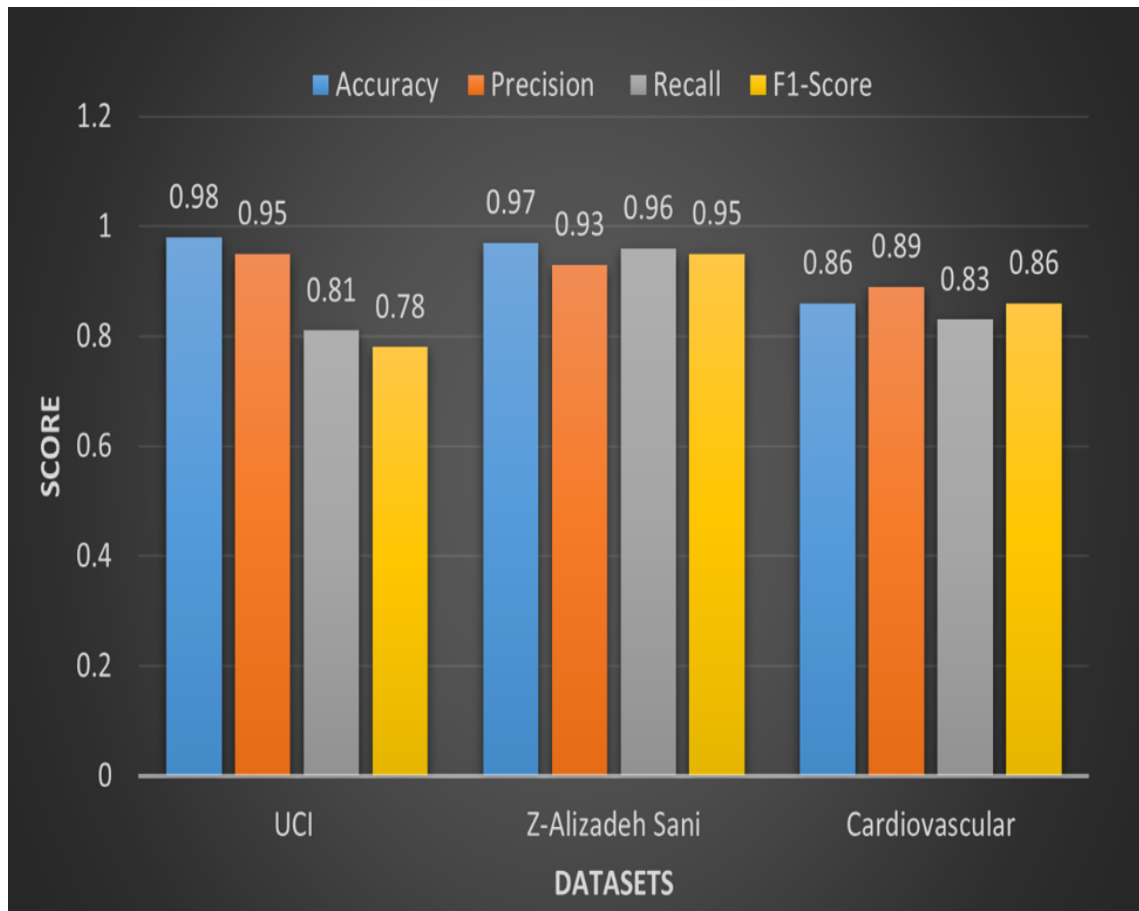
## 5.1 Feature Selection Using GA on Datasets:

When we apply Genetic Algorithms to three datasets, it successfully identifies optimal features that improve the performance of the machine learning model. Table 5.2 demonstrates the total number of features, selected features, and names of features among different datasets.

**Table 5.1:** Feature Selection Using GA: Total Features, Selected Features, and their names

Dataset Name	Total Number of Features	Selected Features by GA	Features Names
UCI Dataset	76	06	'age', 'sex', 'fbs', 'restecg', 'thalch', 'slope'
Z-Alizadeh Sani	56	19	'Age', 'Length', 'Sex', 'BMI', 'HTN', 'EX-Smoker', 'BP', 'PR', 'Typical Chest Pain', 'Atypical', 'Nonanginal', 'Q Wave', 'St Elevation', 'Poor R Progression', 'HDL', 'BUN', 'Neut', 'EF-TTE', 'Region RWMA'
Cardiovascular Disease Dataset	11	09	'age', 'gender', 'height', 'weight', 'ap <sub>h</sub> i', 'ap <sub>l</sub> o', 'cholesterol', 'gluc', 'alco'

## 5.2 Comparison of datasets:



**Figure 5.1:** Comparison among datasets

Here's a comparison of the three datasets based on their accuracy, precision, recall, and F1-score shown in figure 5.1:

UCI Dataset: Accuracy: 0.98 Precision: 0.95 Recall: 0.81 F1-score: 0.78

Z-Alizadeh Sani Dataset: Accuracy: 0.97 Precision: 0.93 Recall: 0.96 F1-score: 0.95

Cardiovascular Dataset: Accuracy: 0.86 Precision: 0.89 Recall: 0.83 F1-score: 0.86

Comparing the datasets based on these performance metrics, we can draw the following observations:

Accuracy:

The UCI dataset has the highest accuracy of 0.98, followed closely by the Z-Alizadeh Sani dataset with 0.97 accuracy. The Cardiovascular dataset has the lowest accuracy at 0.86.

Precision:

The UCI dataset achieves the highest precision of 0.95, indicating a relatively low false positive rate. The Z-Alizadeh Sani dataset has a slightly lower precision of 0.93. The Cardiovascular dataset has the highest precision at 0.89.

Recall:

The Z-Alizadeh Sani dataset has the highest recall rate of 0.96, indicating a low false negative rate. The Cardiovascular dataset follows closely with a recall rate of 0.83, while the UCI dataset has the lowest recall at 0.81.

F1-score:

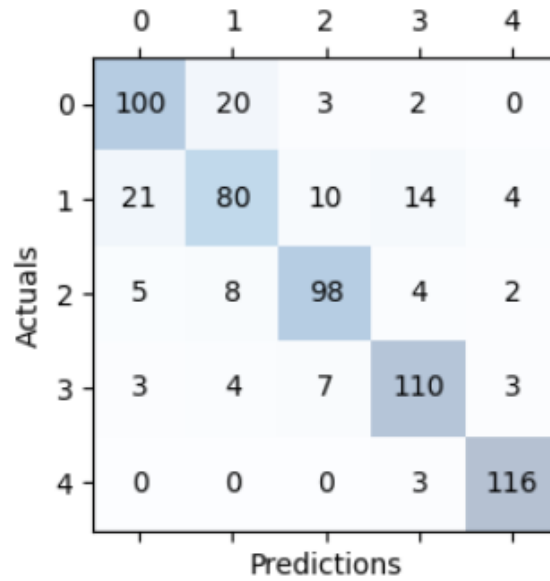
The Z-Alizadeh Sani dataset achieves the highest F1-score of 0.95, indicating a good balance between precision and recall. The Cardiovascular dataset has the next highest F1-score at 0.86, and the UCI dataset has the lowest F1-score of 0.78.

Overall, the UCI dataset performs well in terms of accuracy and precision but has relatively lower recall and F1-score compared to the other datasets. The Z-Alizadeh Sani dataset shows high performance across all metrics, indicating a well-balanced classification performance. The Cardiovascular dataset, although having lower accuracy compared to the others, still achieves a good F1-score and has the highest precision among the three datasets.

### 5.3 Confusion Matrix and AUR-ROC Curve:

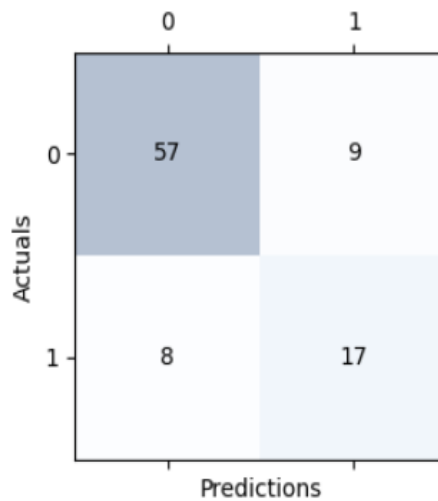
The Confusion Matrix of the dataset UCI heart Disease Dataset is depicted in figure 5.2. The dataset is trained and tested on the proposed model and predictions, and patterns according to positive and negative trends are recorded in the form of a matrix. It has been shown that 80 of the predictions provided by our model are in fact true positives. These individuals actually exhibit illness symptoms. In the dataset, 14 predictions are positive, however, our suggested algorithm gets them wrong.

Even though 4 of the forecasts are incorrect, they are actually negative. The method accurately predicts 110 predictions that have undesirable symptoms.



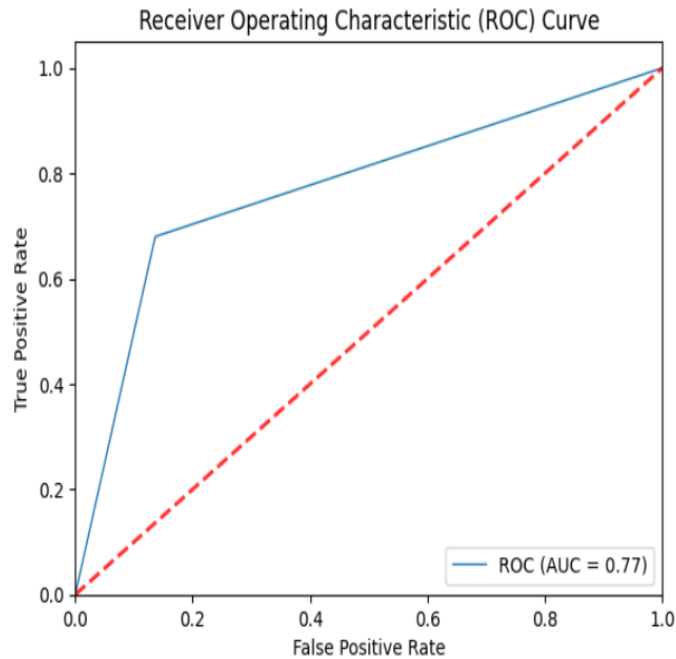
**Figure 5.2:** Confusion Matrix of Heart Disease uci dataset

The matrix of the Z-Alizadeh Sani Dataset is shown in figure 5.3. It is observed that 57 predictions which are made by our model are actually true positive. These patients have truly the symptoms of diseases. 9 predictions are positive in the dataset but our proposed system predicts them wrongly. 8 number of the predictions are actually negative but the proposed system predicts them wrongly. 17 predictions actually have negative symptoms and the system predicts them correctly.



**Figure 5.3:** Confusion Matrix of Heart Disease Z-Alizadeh Sani dataset

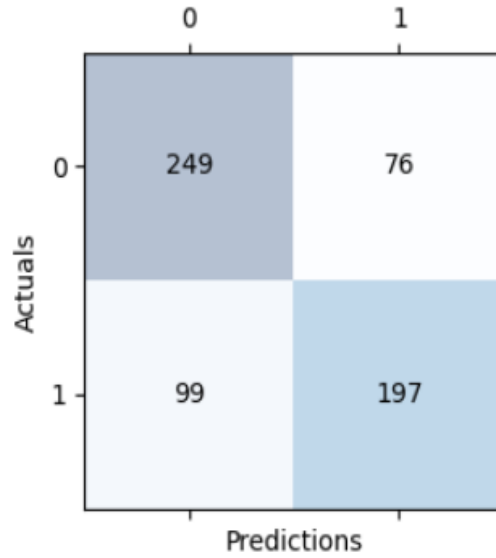
The AUC-ROC Curve is a straight line between the True positive rate and the false negative rate. Plotting the true positive Rate on the y-axis against the false positive rate on the x-axis results in the AUC-ROC curve. The model's classification threshold is changed, and the true Positive rate and false Positive Rate are calculated at each threshold to create the curve. The performance of the model improves as the curve approaches the top-left corner because it denotes a higher True positive rate and a lower false positive rate. A curve that reaches the top-left corner would describe the perfect classifier. The ROC Curve of the dataset is 0.77. The curve is illustrated in figure 5.4.



**Figure 5.4:** AUC-ROC Cruve

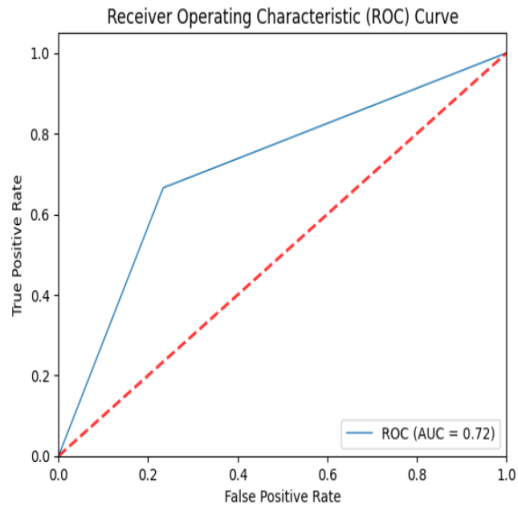
Figure 5.5 illustrates the Cardio dataset matrix. It has been shown that 249 of the predictions provided by our model are in fact true positives. These folks actually exhibit illness symptoms. 76 predictions in the dataset are positive, however, our suggested approach predicted them incorrectly. 99 percent of forecasts are false, yet the proposed system predicted them incorrectly. 197 forecasts were actually negative, and the system properly predicted them.

Between the True positive rate and the false negative rate, the AUC-ROC Curve draws a straight line. The AUC-ROC curve is produced by plotting the true positive rate on the y-axis against the false positive rate on the x-axis. The curve is produced by altering the classification threshold of the model and calculating the true Positive rate and the false



**Figure 5.5:** Confusion Matrix of Cardio Dataset

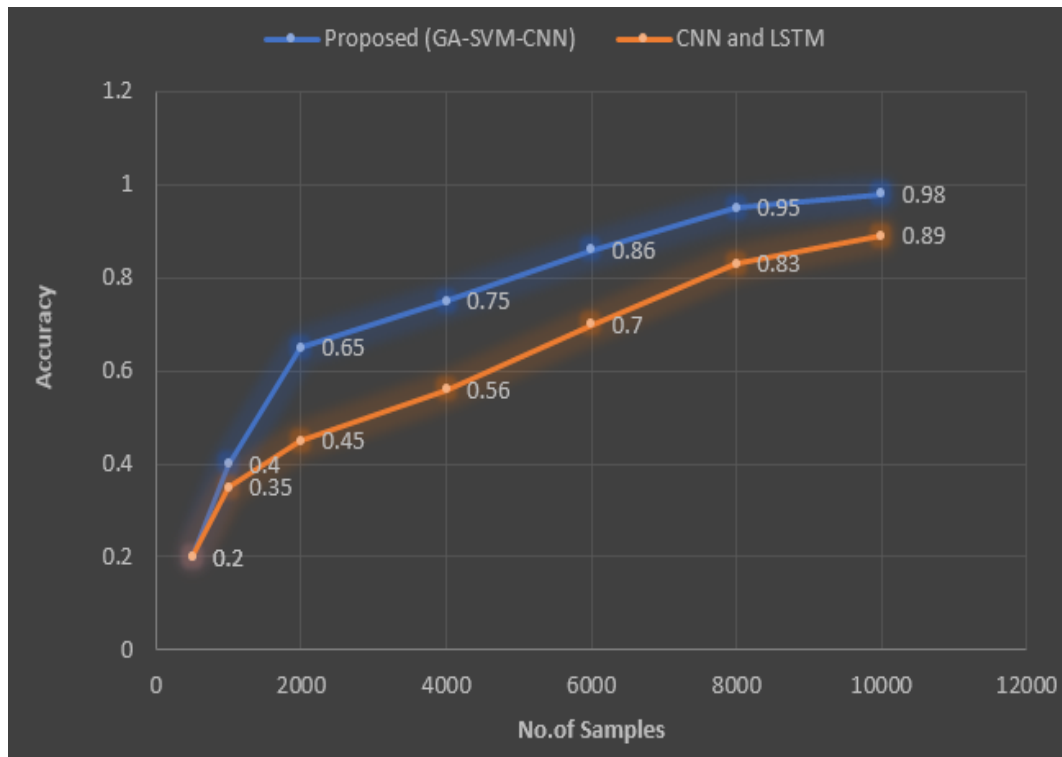
positive rate at each threshold. As the curve moves closer to the top-left corner, which indicates a greater True positive rate and a lower false positive rate, the performance of the model gets better. The ideal classifier would have a curve that reaches the top-left corner. The dataset’s ROC Curve is 0.72. The curve is depicted in figure 5.6.



**Figure 5.6:** AUC-ROC Curve Cardiovascular

## 5.4 Accuracy Comparison:

For the results and evaluation of the proposed model, three datasets are used. In figure 5.7 the accuracy comparison of the dataset UCI Heart Disease Dataset is plotted. Plotted accuracy of the dataset is the combination of training and testing. The proposed model and existing model are executed on the same dataset. The existing model is a combination of CNN and LSTM. At the start accuracy of the proposed model and the existing one is the same which is 0.2. After continuous training and testing proposed model outperform the existing model in accuracy. The model reaches an accuracy of 0.98 while the existing have only 0.89.

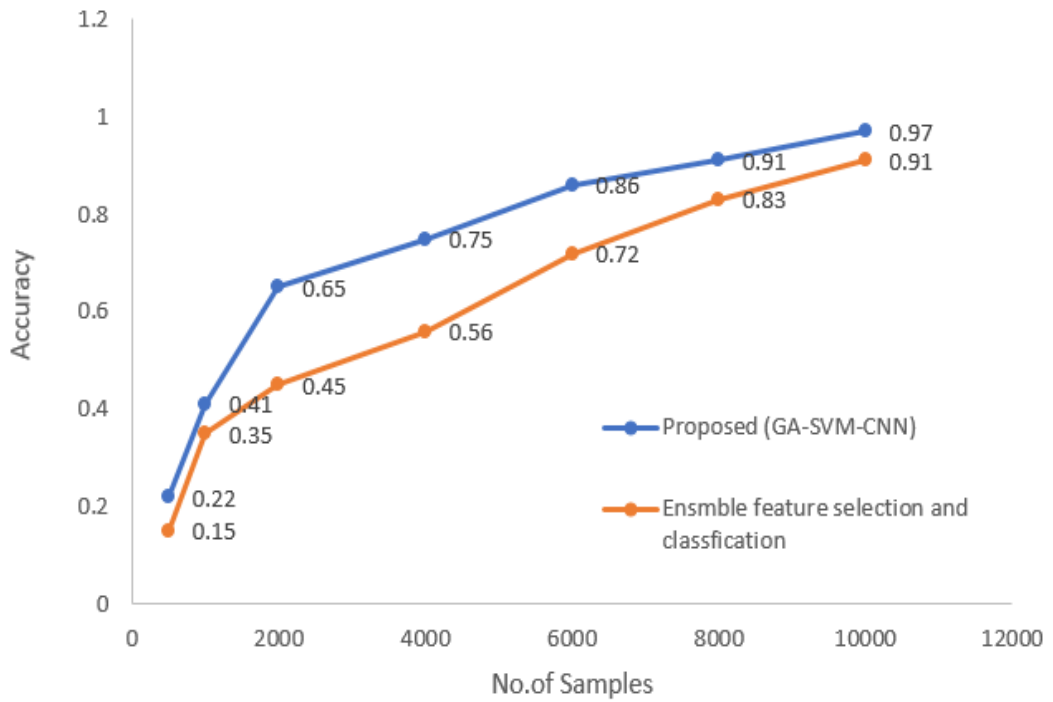


**Figure 5.7:** Accuracy Comparison of heart disease uci

In Figure 5.8, the accuracy comparison of Z-Alizadeh Sani Dataset is plotted. Plotted accuracy of the dataset is the reflection of training and testing. The standing model uses the technique of ensemble feature selection and classification on the same dataset, the suggested model and the current model are both run in the same environment.



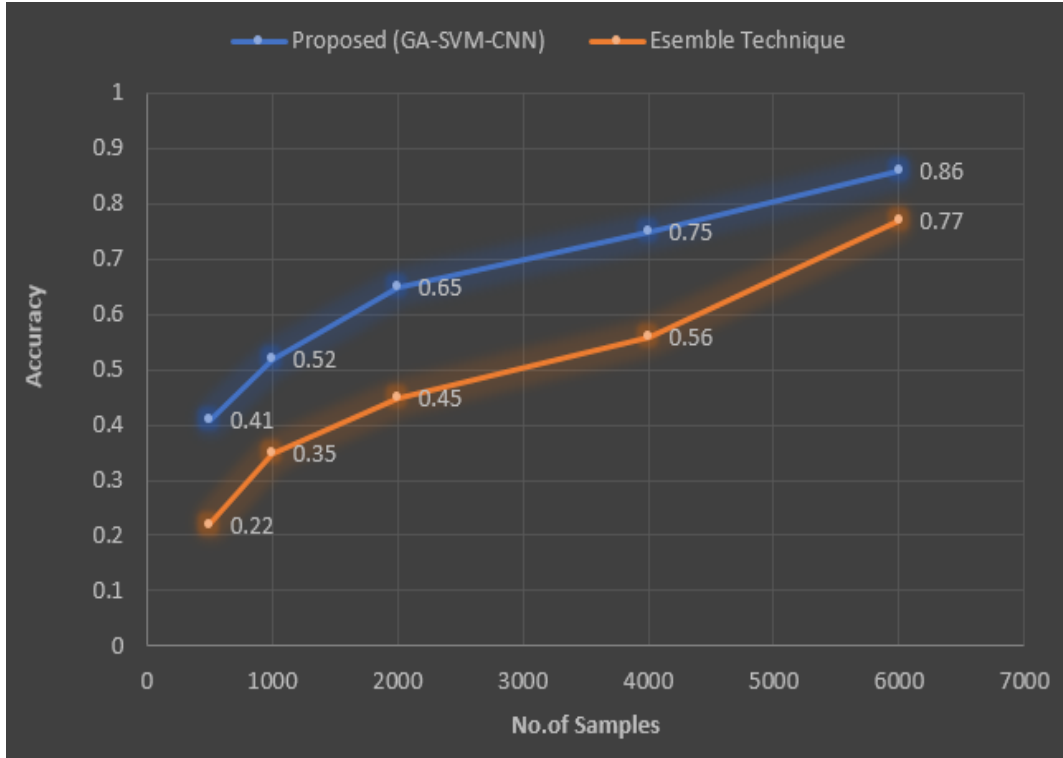
The proposed model's initial accuracy is 0.2, compared to the present model's initial accuracy of 0.15. The proposed solution outperforms the existing model in accuracy after repeated training and testing. The proposed model's accuracy is 0.97 as opposed to the old model's accuracy of 0.91.



**Figure 5.8:** Accuracy Comparison of Z-Alizadeh Sani

The accuracy comparison of the cardiovascular Heart Disease Dataset is presented in Figure 5.9. The dataset's plotted accuracy is the result of both training and testing. On the same dataset, the suggested model and the current model both are executed on the same dataset.

The current one uses the ensemble technique. The proposed model outperforms the existing model at the beginning of the process. The suggested model outperforms the existing model in accuracy after repeated training and testing. The proposed model's accuracy is 0.86 as opposed to the old model's accuracy of 0.71.



**Figure 5.9:** Accuracy Comparison Cardiovascular Disease Dataset

## 5.5 Comparison of Precision, Recall, and F1-score:

To evaluate the performance of the existing different parameters and scoring criteria are kept in consideration. Along with accuracy the precision, recall, and F1-score of the proposed model is calculated and compared with existing models 5.3. The precision of the proposed models is calculated by using the following equation. 5.5.1.

$$Precision = TP / TP + FN \quad (5.5.1)$$

The precision of the proposed model is recorded as 0.95 while the existing one has 0.93 for the data set heart diseases UCI. For the second and third datasets, the precision is 0.93 and 0.89 respectively. While the existing ones have an accuracy of 0.91 and 0.86. It is recorded that the proposed approach knocks the existing ones in terms of precision. Table 5.2 depicts the comparison between different performance metrics of the existing and proposed model.

Recall is another performance metric for the proposed model. The Recall is calculated by using the following equation 5.5.2.

**Table 5.2:** Comparison of Proposed and Existing Technique

Proposed Hybrid Model			Existing Model			
Datasets	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Heart-Disease-uci	0.95	0.81	0.78	0.93	0.77	0.52
Z-Alizadeh Sani	0.93	0.96	0.95	0.91	0.92	0.91
Cardiovascular	0.89	0.83	0.86	0.86	0.80	0.81

$$Recall = TP / (TP + FP) \quad (5.5.2)$$

Recalls of the proposed model on different datasets are recorded as 0.81, 0.96, and 0.83. while the existing have an accuracy of 0.77, 0.92, and 0.80. F1-Score is formulated by using the following equation. The F1-Score of the proposed model is 0.5, 0.95, and 0.86 while the existing models have 0.52, 0.91, and 0.81. for recall the following equation is used [5.5.3](#)

$$F1 - Score = 2 * Recall * Precision / (Recall + Precision) \quad (5.5.3)$$

## Conclusion and Future Work

The combination of a genetic algorithm for feature selection and a hybrid SVM-CNN model for heart disease prediction has shown promising results. This approach capitalizes on the strengths of both techniques, leveraging the genetic algorithm's ability to identify relevant features and the SVM-CNN hybrid model's capacity for accurate prediction. The genetic algorithm plays a crucial role in selecting the most informative features from a large pool of potential predictors. By iteratively evaluating the fitness of different feature subsets, it optimizes the selection process and reduces dimensionality. This feature selection step is critical for enhancing the model's performance, as it helps eliminate irrelevant or redundant features that may introduce noise or overfitting. The SVM-CNN hybrid model takes advantage of the power of both support vector machines (SVM) and convolutional neural networks (CNN).

SVMs excel at binary classification tasks by creating an optimal hyperplane to separate data points, while CNNs are well-suited for classification. By combining the two, we obtain a model that can effectively handle both structured and unstructured features for heart disease prediction. The hybrid model's performance is enhanced by the selected features, which are fed into the SVM for classification and then passed to the CNN for further analysis. This integration allows the model to capture complex patterns and relationships present in the data, leading to improved accuracy and robustness. The results obtained from applying this approach to heart disease prediction are highly encouraging.

By leveraging the genetic algorithm's feature selection, we achieve a more efficient and accurate model, reducing computation time and improving predictive performance. The

hybrid SVM-CNN model demonstrates its ability to handle diverse data types, capturing numerical features to make accurate predictions. However, it is important to note that this approach is not without limitations. The effectiveness of the genetic algorithm heavily relies on the quality and diversity of the initial feature pool. Additionally, the performance of the SVM-CNN hybrid model is influenced by factors such as dataset size, class imbalance, and the complexity of the heart disease patterns present in the data. Further research and validation are necessary to assess the generalizability and robustness of this approach across different datasets and populations. Nonetheless, the combination of a genetic algorithm for feature selection and a hybrid SVM-CNN model holds great promise for enhancing heart disease prediction accuracy and contributing to the development of personalized healthcare solutions in the future.

The computational complexity of the model can be high. The genetic algorithm involves an iterative search process to find the optimal feature subset, which is time-consuming and computationally demanding, especially with large datasets. Additionally, the hybrid SVM-CNN model may require significant "computational resources" and training time due to the complexity of CNNs. The result of the GA-SVM-CNN approach heavily relies on the quality and relevance of the input data. If the dataset is incomplete, contains outliers, or exhibits class imbalance, it may affect the accuracy and reliability of the predictions. Preprocessing and data cleaning techniques should be employed to ensure data quality before applying the approach.

# Bibliography

- [1] *Cardiovascular diseases (CVDs)*. [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)). Accessed: 2023-005-22.
- [2] *European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice*. <https://pubmed.ncbi.nlm.nih.gov/27222591/>. Accessed: 2023-005-22.
- [3] piepoli Piepoli MF;Hoes AW;Agewall S;Albus C;Brotons C;Catapano AL;Cooney MT;Corrà U;Cosyns B;Deaton C;Graham I;Hall MS;Hobbs FDR;Løchen ML;Löllgen H;Marques-Vidal P;Perk J;Prescott E;Redon J;Richter DJ;Sattar N;Smulders Y;Tiberi M;van der Worp HB;van Dis I;Versc. *2016 European Guidelines on Cardiovascular Disease Prevention in Clinical Practice: The Sixth joint task force of the European Society of Cardiology and other societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)developed with the special contribution of the European Association for Cardiovascular Prevention amp; Rehabilitation (EACPR)*. URL: <https://pubmed.ncbi.nlm.nih.gov/27222591/>.
- [4] Chowdhury R Chowdhury R;Khan H;Heydon E;Shroufi A;Fahimi S;Moore C;Stricker B;Mendis S;Hofman A;Mant J;Franco OH; *Adherence to cardiovascular therapy: A meta-analysis of prevalence and clinical consequences*. URL: <https://pubmed.ncbi.nlm.nih.gov/23907142/>.
- [5] ;Mozaffarian D;Benjamin EJ;Go AS;Arnett DK;Blaha MJ;Cushman M;Das SR;de Ferranti S;Després JP;Fullerton HJ;Howard VJ;Huffman MD;Isasi CR;Jiménez MC;Judd SE;Kissela BM;Lichtman JH;Lisabeth LD;Liu S;Mackey RH;Magid DJ;McGuire DK;Mohler ER;Moy CS;Muntner P;M. *Heart disease and stroke statistics-2016 up-*

- date: A report from the American Heart Association.* URL: <https://pubmed.ncbi.nlm.nih.gov/26673558/>.
- [6] Jha P;Ramasundarahettige C;Landsman V;Rostron B;Thun M;Anderson RN;McAfee T;Peto R; *21st-century hazards of smoking and benefits of cessation in the United States.* URL: <https://pubmed.ncbi.nlm.nih.gov/23343063/>.
- [7] Whelton PK;Carey RM;Aronow WS;Casey DE;Collins KJ;Dennison Himmelfarb C;DePalma SM;Gidding S;Jamerson KA;Jones DW;MacLaughlin EJ;Muntner P;Ovbiagele B;Smith SC;Spencer CC;Stafford RS;Taler SJ;Thomas RJ;Williams KA;Williamson JD;Wright JT; *2017 ACC/AHA/AAPA/ABC/ACPM/AGS/apha/ash/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: Executive summary: A report of the american college of cardiology/american heart association task force on clinical practice guidelines.* URL: <https://pubmed.ncbi.nlm.nih.gov/29133354/>.
- [8] Attia ZI;Kapa S;Lopez-Jimenez F;McKie PM;Ladewig DJ;Satam G;Pellikka PA;Enriquez-Sarano M;Noseworthy PA;Munger TM;Asirvatham SJ;Scott CG;Carter RE;Friedman PA; *Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram.* URL: <https://pubmed.ncbi.nlm.nih.gov/30617318/>.
- [9] Krittanawong C;Virk HUH;Bangalore S;Wang Z;Johnson KW;Pinotti R;Zhang H;Kaplan S;Narasimhan B;Kitai T;Baber U;Halperin JL;Tang WHW; *Machine learning prediction in cardiovascular diseases: A meta-analysis.* URL: <https://pubmed.ncbi.nlm.nih.gov/32994452/>.
- [10] Dey D;Slomka PJ;Leeson P;Comaniciu D;Shrestha S;Sengupta PP;Marwick TH; *Artificial Intelligence in cardiovascular imaging: JACC state-of-the-art review.* URL: <https://pubmed.ncbi.nlm.nih.gov/30898208/>.
- [11] Maini E;Venkateswarlu B;Maini B;Marwaha D; *Machine learning-based heart disease prediction system for Indian population: An exploratory study done in South India.* URL: <https://pubmed.ncbi.nlm.nih.gov/34305284/>.
- [12] Anish Gopal Pemmaraju, A Asish, and Subhalaxmi Das. “Heart Disease Prediction Using Feature Selection and Machine Learning Techniques”. In: *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)*. IEEE. 2022, pp. 28–33.

## BIBLIOGRAPHY

- [13] Sudarshan Nandy et al. “An intelligent heart disease prediction system based on swarm-artificial neural network”. In: *Neural Computing and Applications* (2021), pp. 1–15.
- [14] MG Bindu and MK Sabu. “A hybrid feature selection approach using artificial bee colony and genetic algorithm”. In: *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*. IEEE, 2020, pp. 211–216.
- [15] Karunakaran Velswamy et al. “Classification model for heart disease prediction with feature selection through modified bee algorithm”. In: *Soft Computing* (2021), pp. 1–9.
- [16] Burak Kolukisa and Burcu Bakir-Gungor. “Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis”. In: *Computer Standards & Interfaces* 84 (2023), p. 103706.
- [17] Jaishri Wankhede, Palaniappan Sambandam, and Magesh Kumar. “Effective prediction of heart disease using hybrid ensemble deep learning and tunicate swarm algorithm”. In: *Journal of Biomolecular Structure and Dynamics* 40.23 (2022), pp. 13334–13345.
- [18] Ashir Javeed et al. “Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification”. In: *Mobile Information Systems* 2020 (2020), pp. 1–11.
- [19] Muhammad Salman Pathan et al. “Analyzing the impact of feature selection on the accuracy of heart disease prediction”. In: *Healthcare Analytics* 2 (2022), p. 100060.
- [20] N Satish Chandra Reddy et al. “Classification and feature selection approaches by machine learning techniques: Heart disease prediction”. In: *International Journal of Innovative Computing* 9.1 (2019).
- [21] Kaushalya Dissanayake and Md Gapar Md Johar. “Comparative study on heart disease prediction using feature selection techniques on classification algorithms”. In: *Applied Computational Intelligence and Soft Computing* 2021 (2021), pp. 1–17.
- [22] Robinson Spencer et al. “Exploring feature selection and classification methods for predicting heart disease”. In: *Digital health* 6 (2020), p. 2055207620914777.



- [23] Samina Kanwal et al. “An effective classification algorithm for heart disease prediction with genetic algorithm for feature selection”. In: *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. IEEE. 2021, pp. 1–6.
- [24] A Lakshmanarao, A Srisaila, and T Srinivasa Ravi Kiran. “Heart disease prediction using feature selection and ensemble learning techniques”. In: *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE. 2021, pp. 994–998.
- [25] Saba Bashir et al. “Improving heart disease prediction using feature selection approaches”. In: *2019 16th international bhurban conference on applied sciences and technology (IBCAST)*. IEEE. 2019, pp. 619–623.
- [26] Qasem Al-Tashi, Helmi Rais, and Said Jadid. “Feature selection method based on grey wolf optimization for coronary artery disease classification”. In: *Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018)*. Springer. 2019, pp. 257–266.
- [27] Karna Vishnu Vardhana Reddy et al. “An Efficient Prediction System for Coronary Heart Disease Risk Using Selected Principal Components and Hyperparameter Optimization”. In: *Applied Sciences* 13.1 (2023), p. 118.
- [28] Ezekiel Adebayo Ogundepo and Waheed Babatunde Yahya. “Performance analysis of supervised classification models on heart disease prediction”. In: *Innovations in Systems and Software Engineering* (2023), pp. 1–16.
- [29] Jafar Abdollahi and Babak Nouri-Moghaddam. “A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation”. In: *Iran Journal of Computer Science* 5.3 (2022), pp. 229–246.
- [30] MA Jabbar, BL Deekshatulu, and Priti Chandra. “Intelligent heart disease prediction system using random forest and evolutionary approach”. In: *Journal of network and innovative computing* 4.2016 (2016), pp. 175–184.
- [31] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. “Effective heart disease prediction using hybrid machine learning techniques”. In: *IEEE access* 7 (2019), pp. 81542–81554.

## BIBLIOGRAPHY

- [32] Rohit Bharti et al. “Prediction of heart disease using a combination of machine learning and deep learning”. In: *Computational intelligence and neuroscience 2021* (2021).
- [33] Senthil Murugan Nagarajan et al. “Innovative feature selection and classification model for heart disease prediction”. In: *Journal of Reliable Intelligent Environments* 8.4 (2022), pp. 333–343.
- [34] P Hamsagayathri and S Vigneshwaran. “Symptoms based disease prediction using machine learning techniques”. In: *2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*. IEEE, 2021, pp. 747–752.
- [35] Robinson Spencer et al. “Exploring feature selection and classification methods for predicting heart disease”. In: *Digital health* 6 (2020), p. 2055207620914777.
- [36] Min Chen et al. “Disease prediction by machine learning over big data from health-care communities”. In: *Ieee Access* 5 (2017), pp. 8869–8879.
- [37] Burak Kolukisa and Burcu Bakir-Gungor. “Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis”. In: *Computer Standards & Interfaces* 84 (2023), p. 103706.
- [38] Bhanu Prakash Doppala et al. “A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset”. In: *Distributed and Parallel Databases* (2021), pp. 1–20.
- [39] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, and Emmanuel Andrés. “Classification models for heart disease prediction using feature selection and PCA”. In: *Informatics in Medicine Unlocked* 19 (2020), p. 100330. ISSN: 2352-9148. DOI: <https://doi.org/10.1016/j.imu.2020.100330>. URL: <https://www.sciencedirect.com/science/article/pii/S2352914820300125>.
- [40] Dengqing Zhang et al. “Heart disease prediction based on the embedded feature selection method and deep neural network”. In: *Journal of healthcare engineering* 2021 (2021), pp. 1–9.
- [41] M Kavitha et al. “Heart disease prediction using hybrid machine learning model”. In: *2021 6th international conference on inventive computation technologies (ICICT)*. IEEE, 2021, pp. 1329–1333.

## BIBLIOGRAPHY

- [42] B Geluvaraj et al. “A Hybrid Approach for Predicting Diseases using Clustering and Classification Techniques”. In: *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE. 2022, pp. 1–6.
- [43] VK Sudha and D Kumar. “Hybrid CNN and LSTM Network For Heart Disease Prediction”. In: *SN Computer Science* 4.2 (2023), p. 172.
- [44] Girish S Bhavekar and Agam Das Goswami. “A hybrid model for heart disease prediction using recurrent neural network and long short term memory”. In: *International Journal of Information Technology* 14.4 (2022), pp. 1781–1789.
- [45] Sayali Ambekar and Rashmi Phalnikar. “Disease risk prediction by using convolutional neural network”. In: *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE. 2018, pp. 1–5.
- [46] Dhiraaj Dahiwade, Gajanan Patle, and Ektaa Meshram. “Designing disease prediction model using machine learning approach”. In: *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE. 2019, pp. 1211–1215.
- [47] G Saranya and A Pravin. “A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease”. In: *Journal of Ambient Intelligence and Humanized Computing* (2022), pp. 1–15.
- [48] Debjani Panda et al. “Predictive systems: Role of feature selection in prediction of heart disease”. In: *Journal of Physics: Conference Series*. Vol. 1372. 1. IOP Publishing. 2019, p. 012074.
- [49] Vikas Chaurasia and Aparna Chaurasia. “Novel Method of Characterization of Heart Disease Prediction Using Sequential Feature Selection-Based Ensemble Technique”. In: *Biomedical Materials & Devices* (2023), pp. 1–10.
- [50] Mohamed G El-Shafiey et al. “A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest”. In: *Multimedia Tools and Applications* 81.13 (2022), pp. 18155–18179.
- [51] D Deepika and N Balaji. “Effective heart disease prediction using novel MLP-EBMDA approach”. In: *Biomedical Signal Processing and Control* 72 (2022), p. 103318.