# Early Risk Prediction in Acute Aortic Syndrome On Clinical Data Using Machine Learning

by

Mehdi Tavafi

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (MSc) in Computational Science

The Faculty of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

© Mehdi Tavafi, 2024

## THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE
**Laurentian Université/Université Laurentienne**
Faculty of Graduate Studies/Faculté des études supérieures

| | | |
|---|---|---|
| Title of Thesis<br>Titre de la thèse | Early Risk Prediction in Acute Aortic Syndrome On Clinical Data Using Machine Learning | |
| Name of Candidate<br>Nom du candidat | Tafavi, Mehdi | |
| Degree<br>Diplôme | Master of Science | |
| Department/Program<br>Département/Programme | Computational Sciences | Date of Defence<br>Date de la soutenance April, 2024 |

**APPROVED/APPROUVÉ**

Thesis Examiners/Examinateurs de thèse:

Dr. Kalpdrum Passi
(Co-Supervisor/Co-directeur(trice) de thèse)

Dr. Robert Ohle
(Co-Supervisor/Co-directeur(trice) de thèse)

Dr. Ratvinder Grewal
(Committee member/Membre du comité)

Dr. Vasu Appanna
(Committee member/Membre du comité)

Dr. R. Kanchana
(External Examiner/Examinateur externe)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Dr. Tammy Eger
Madame Tammy Eger
VP Research (Graduate Studies)
VR, Recherche (Études supérieures)

**ACCESSIBILITY CLAUSE AND PERMISSION TO USE**

I, **Mehdi Tafavi**, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.

# Abstract

Advancements in machine learning present novel opportunities for early prediction of Acute Aortic Syndrome (AAS) as a critical and life-threatening clinical condition and the identification of critical features influencing this prediction. This study concentrates on integrating, cleaning, and handling missing data from extensive clinical datasets sourced from 150 emergency departments across Canada and the USA. Covering medical histories of nearly 150,000 patients from 2021 to 2022, the dataset comprises categorical clinical variables. Additionally, the research focuses on constructing predictive machine learning models utilizing various data-splitting strategies and classifiers to optimize AAS prediction. Methodologically, the study encompasses data identification, acquisition, exploration, processing, and feature extraction, followed by dimensionality reduction using Principal Component Analysis (PCA) and other feature selection methods such as Correlation-based (CFS) and Relief. The multiple imputations method and the SMOTE method are utilized for handling missing and imbalanced data, respectively. The findings demonstrate that employing the Relief-feature method with an 80-10-10 split ratio alongside the Random Forest classifier yields an exceptional accuracy of 99.3%, surpassing alternative models.. Furthermore, this research addresses a prevalent challenge encountered by many researchers regarding dataset size limitations, thereby facilitating the utilization of the integrated and prepared dataset for research on AAS and other cardiovascular diseases.

Keywords: Machine learning, Acute Aortic Syndrome (AAS), Clinical data, Data Integration, Data cleaning, SMOTE method, Feature Extraction, Principal Component Analysis (PCA), Relief method, Correlation-based feature selection (CFS)

# Acknowledgments

I extend my heartfelt appreciation to those whose unwavering support has been instrumental in the successful completion of this thesis. At the forefront of my gratitude is Dr. Kalpdrum Passi, my thesis supervisor, whose profound insights and Steadfast encouragement played a pivotal role in shaping the trajectory of this project. Dr. Passi's guidance, especially during the writing phase, proved invaluable to this thesis, and his consistent support throughout my master's program has been a cornerstone of my academic journey.

I am equally honored to express my gratitude to Dr. Robert Ohle, my supervisor in the medical field. His extensive medical knowledge and continuous support were integral to the success of this project. Dr. Ohle's mentorship has provided me with a deeper understanding of the medical intricacies related to my research, contributing significantly to its overall quality.

To my beloved parents, words cannot capture the depth of my appreciation for your enduring support in all my endeavors. Your unconditional love, constant encouragement, and the sacrifices you've made have been the bedrock upon which I've built my dreams. The values of hard work and perseverance you instilled in me have been guiding principles throughout my academic journey. For this, I am eternally grateful.

A special acknowledgment goes to my sister, confidante, and friend. Your unwavering support has been a constant source of strength in my life. Thank you for believing in me, even during moments of self-doubt. Your faith has been a guiding light, propelling me forward with confidence.

I am acutely aware that this research would not have been possible without the collective support of these remarkable individuals and many more. I deeply appreciate their unwavering belief in me and their invaluable contributions to my academic and personal goals.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1
# Introduction

Acute Aortic Syndrome (AAS) represents a critical and life-threatening clinical condition encompassing three distinct diagnoses: acute aortic dissection, penetrating atherosclerotic ulcer, and intramural hematoma [1]. AAS involves a flaw in the aorta's wall, potentially causing obstruction in the vessel or its branches and posing the risk of aortic rupture, which can lead to severe blood loss. Symptoms may manifest in various forms, such as chest pain [2]. Such a tear can extend along the entire length of the vessel, obstructing blood supply to various organs, and potentially leading to vessel rupture [3], [4].

The general population's frequency of AAS varies from 3.5 to 6.0 cases per 100,000 patient-years, with a higher incidence in individuals aged 64 to 74 years (27 cases per 100,000) and those aged 75 years and older (35 cases per 100,000) [3], [5]. However, a recent study by Robert Ohle and team reported a lower incidence of 0.61 per 100,000, with a notable decrease in one-year mortality from 47.4% to 29.1%, with 14.9% occurring in the emergency department, and a 12.5% rate of misdiagnosis [6]. As a result, the range of reported incidences now extends from 0.61 to 6.0 cases per 100,000 patient-years.

The diagnostic landscape of AAS poses a complex challenge due to its diverse clinical manifestations and the absence of standardized guidelines for suspicion [2],[7]. The absence of widely accepted guidelines, ensuring both safety and efficiency, poses a significant challenge for clinicians in determining when to suspect AAS [2]. A considerable number of medical practitioners overlook the consideration of AAS in their initial diagnostic assessments or lack

precision. Taking into account the limited availability of comprehensive medical information for these patients, the scope of the data, and the impact of various influencing factors, diagnosing AAS becomes considerably more challenging [8].

The primary factor leading to ED mortality in AAS is diagnostic delay, often due to hesitancy to consider AAS initially. Implementing high-quality clinical guidelines and innovative alert systems can significantly reduce the time to imaging, surgical consultation, and treatment, crucially improving patient outcomes [6].

Concurrently, the rise of machine learning, a facet of artificial intelligence facilitating computers in learning from data, has sparked a revolution across numerous sectors, with healthcare being a notable beneficiary [9].

Acknowledging the evident gap in clinical guidance, this thesis takes a proactive approach by leveraging a substantial dataset that integrates clinical and medical records from a significant patient cohort. This dataset is gathered from 150 emergency departments located in both Canada and the USA. It includes the medical histories of close to 150,000 patients, covering the timeframe from 2021 to 2022. The data includes categorical clinical variables and spans the period from 2021 to 2022.

Subsequently, this thesis explores the potential of using machine learning methods to achieve optimal accuracy in predicting patients who might be at risk of suffering from AAS. The overarching objective is to evaluate the necessity for a clinical decision rule capable of enhancing the accuracy of AAS diagnosis. To accomplish this, we conducted a comprehensive systematic review of diagnostic accuracy, aiming to explore and evaluate existing methodologies in the diagnosis of AAS.

The main goal of this approach is not only to enhance our comprehension of AAS but also to establish the foundation for creating predictive machine-learning techniques. These techniques are designed to aid clinicians in making prompt and precise AAS diagnoses. Through a thorough systematic review of diagnostic accuracy, our objective is to construct a predictive model for anticipating the likelihood of an AAS diagnosis. Additionally, we aim to identify critical medical factors that play a pivotal role in expediting the AAS diagnostic process when patients present at the Emergency Department.

## 1.1 Acute Aortic Syndrome (AAS)

Cardiovascular diseases continue to be a leading cause of morbidity and mortality globally, posing a significant challenge to public health. Within this broad spectrum of cardiovascular disorders, acute aortic syndrome (AAS) emerges as a critical and often life-threatening condition that demands immediate attention and intervention [2],[3].

Acute aortic syndrome is a term used to talk about serious aorta problems that have similar symptoms and difficulties. It includes different life-threatening issues affecting the aorta, our body's biggest artery. These problems can be aortic dissection (AD), intramural hematoma (IMH), penetrating atherosclerotic ulcer (PAU), aortic rupture, and sometimes, aorta injury with an intimal laceration (see Figure 1.1) [3],[6]. The common factor in AAS is the disruption of the middle layer of the aorta, leading to bleeding within the wall (intramural hematoma), causing the layers to separate (dissection), or going through the wall (ruptured PAU or trauma) [3].

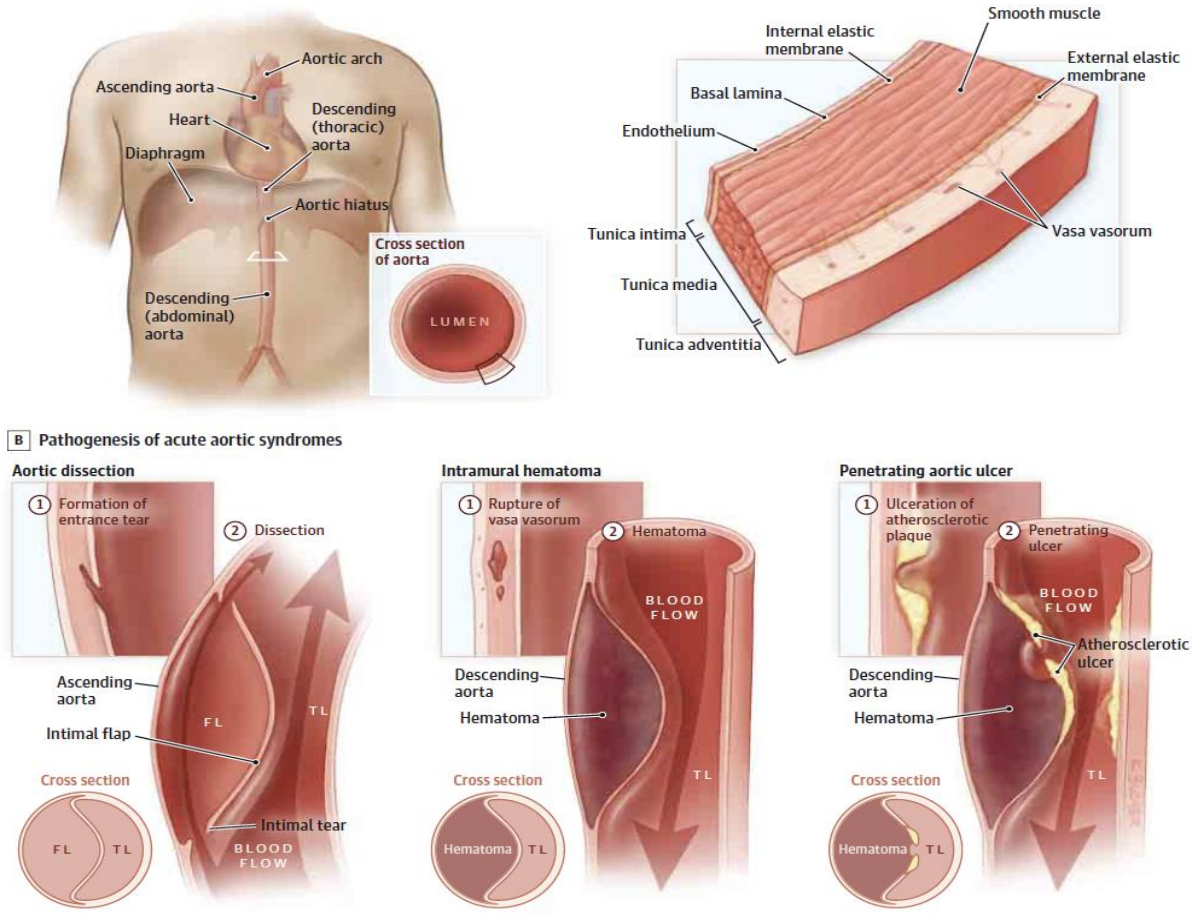


**Figure 1.1 Anatomy of the Aorta and Pathogenesis of Acute Aortic Syndrome [3]**

High blood pressure and various genetic disorders that affect connective tissues are the most observed risk factors. Patients experiencing AAS often show similar symptoms, regardless of whether it involves aortic dissection, intramural hematoma (IMH), penetrating atherosclerotic ulcer (PAU), or a contained aortic rupture. Pain is the primary symptom when someone has acute aortic dissection, and it should immediately trigger attention, including the use of diagnostic imaging methods like multi-slice computed tomography or magnetic resonance imaging [1], [11].

The aorta serves as the main artery responsible for supplying blood throughout the entire body. In AAS, there is a structural flaw in the aorta's wall, which has the potential to obstruct the vessel or its branches, leading to the possibility of rupture and substantial blood loss. This condition can

present diverse symptoms, including chest pain or neurological issues, and these manifestations may resemble those commonly associated with conditions like a heart attack or stroke [1].

As the body's largest artery, the aorta plays a pivotal role in distributing blood from the heart to the entire circulatory system (see Figure 1.2). When the integrity of the aortic wall is compromised, whether due to dissection, intramural hematoma, or penetrating ulcers, it can result in severe consequences, potentially giving rise to catastrophic complications and a significantly heightened risk of sudden death [2].

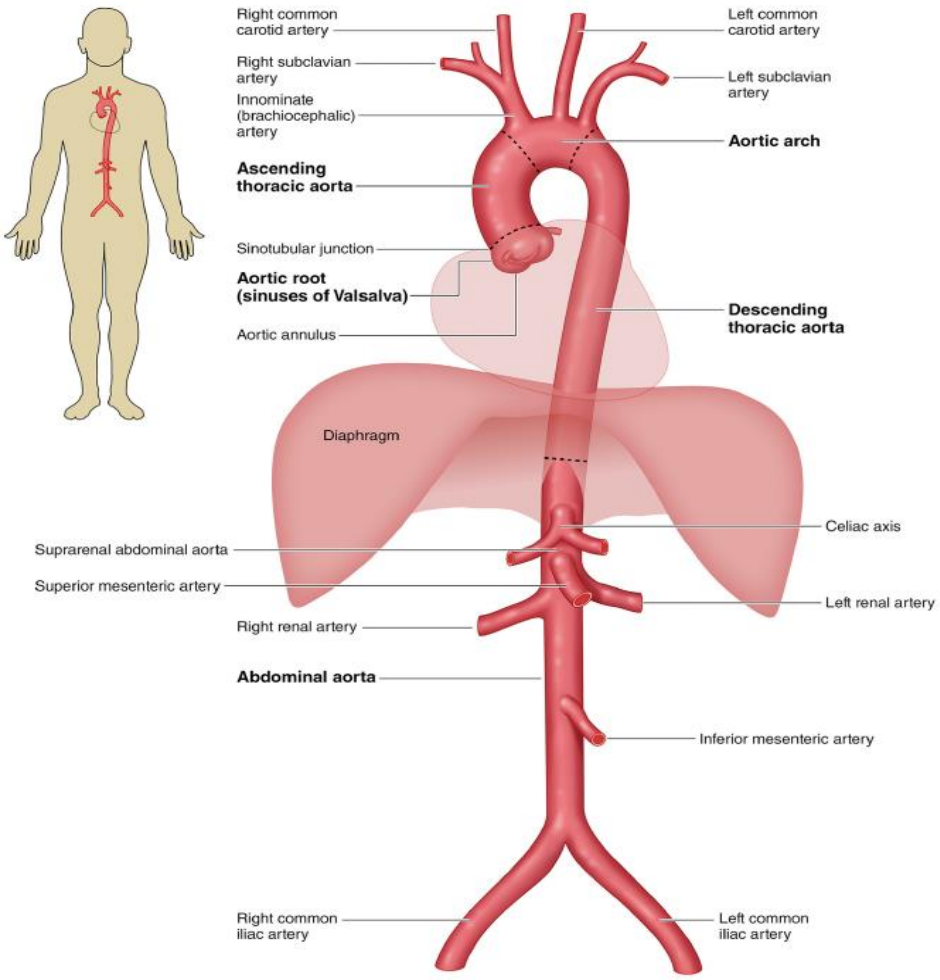


**Figure 1.2 The Anatomy of the Aorta and Its Main Branches [2]**

### 1.1.1 Subtypes of AAS

As mentioned earlier, AAS includes three primary types Acute Aortic Dissection (AAD), Intramural Hematoma (IMH), and Penetrating Aortic Ulcer (PAU) (see Figure 1.3).

**Acute Aortic Dissection (AAD):** A tear occurs in the inner layer of the aortic wall, enabling blood to enter the layers of the wall.

**Intramural Hematoma (IMH):** Bleeding happens within the aortic wall without a clear tear in the inner layer.

**Penetrating Aortic Ulcer (PAU):** A defect or ulceration in the aortic wall penetrates into the media or deeper layers.

Various risk factors, such as hypertension, atherosclerosis, connective tissue disorders, and genetic predispositions are associated with these different subtypes [2], [12].

In aortic dissection, a tear in the inner layer allows blood to seep into the middle part, separating the true and false lumens. In intramural hematoma, blood goes into the middle layer, forming a clot that pushes the outer wall outward while keeping the aortic passage looking somewhat normal. With a penetrating atherosclerotic ulcer, blood enters the middle layer, but the scarring usually keeps it contained, often causing a localized dissection or pseudoaneurysm [2].

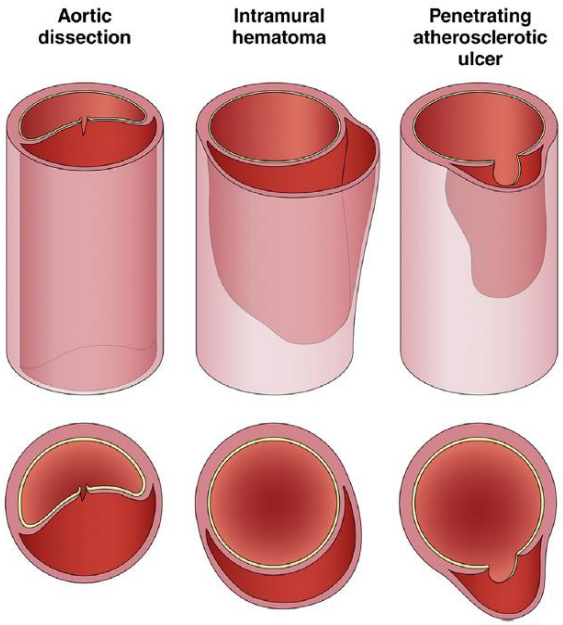


**Figure 1.3 Anatomy of the Aorta and Pathogenesis of Acute Aortic Syndrome [2]**

## 1.2 Machine Learning

Machine learning is a specialized field within artificial intelligence (AI) that focuses on creating algorithms and statistical models. These models allow computer systems to enhance their performance on specific tasks by learning from experience. Unlike traditional programming, where tasks are explicitly defined, machine learning systems learn and improve from data, recognizing patterns and making decisions without direct programming [13],[9].

Machine learning finds applications in various domains, such as data analysis, image and speech recognition, natural language processing, recommendation systems, and autonomous vehicles. It has become an integral part of technological advancements and continues to play a crucial role in the development of intelligent systems [13].

### 1.2.1 Types of Machine Learning

1. **Supervised Learning:** This method involves training an algorithm on a labeled dataset, where input data is paired with desired outputs. The model generalizes from this training data to make predictions or decisions on new, unseen data.

2. **Unsupervised Learning:** In unsupervised learning, algorithms explore data without specific guidance, identifying patterns, relationships, or structures within the data. Common tasks include clustering and dimensionality reduction.

3. **Reinforcement Learning:** Algorithms in reinforcement learning learn by interacting with an environment and receiving feedback through rewards or penalties. The goal is to develop a strategy that maximizes cumulative rewards over time.

## 1.3 Machine Learning in Acute Aortic Syndrome

Machine learning algorithms have shown significant promise in the realm of medical imaging and clinical data analysis for the detection and classification of Acute Aortic Syndrome (AAS). Here's a breakdown of how these technologies are utilized in this context.

### 1.3.1 Medical Imaging Interpretation

1. **Computed Tomography (CT) Scans**: Machine learning algorithms can analyze CT scans to identify subtle patterns, anomalies, and specific characteristics associated with AAS. These algorithms learn from a large dataset of CT images to enhance their ability to accurately detect and classify AAS [14].

2. **Magnetic Resonance Imaging (MRI):** Similar to CT scans, machine learning algorithms can analyze MRI images to identify features indicative of AAS. The algorithms learn from a diverse set of MRI data to improve their ability to recognize different subtypes of AAS [15].

### 1.3.2 Clinical Variable and Results Analysis

1. **Risk Prediction Models**: Machine learning models can be trained on a combination of clinical variables (e.g., age, blood pressure, medical history) and imaging results to predict the risk of AAS in individuals. These models leverage patterns and relationships within the data to provide a quantitative estimate of a patient's risk [16].

2. **Integration of Multimodal Data**: Machine learning allows for the integration of information from various sources, including imaging data and clinical variables. This holistic approach provides a more comprehensive view for accurate detection and classification of AAS [16].

### 1.3.3 Benefits and Impact

1. **Early Detection**: Machine learning algorithms enable the early detection of AAS by identifying subtle signs that may not be easily discernible through traditional methods and preventing mortality.

2. **Precision and Efficiency**: The ability of machine learning to analyze large datasets swiftly enhances the precision and efficiency of AAS diagnosis, which is crucial in emergencies.

## 1.4 Objectives

In the early prediction of acute aortic syndrome, the dataset's size significantly impacts how well machine learning methods can work [17]. A large and detailed dataset is essential for picking up subtle patterns and variations that are important for early detection [18]. This helps machine learning algorithms become more accurate in foreseeing and predicting the beginning of acute aortic syndrome. The size of the dataset becomes even more crucial when concentrating on early prediction. This highlights the importance of having a thorough and well-organized dataset to improve how well predictive models perform in this specific medical area [19]. Given this context, here's an overview of what this thesis aims to achieve.

**1. Integration and Preprocessing of a big dataset:** One of the primary objectives of this study is to integrate and preprocess an extensive dataset, comprising the clinical and medical records of nearly 150,000 cardiovascular patients from 150 emergency departments in Canada and the USA. This initial step holds paramount importance in creating a consolidated and well-organized dataset. Through meticulous integration and preprocessing, the aim is to enhance the quality and uniformity of the data, ensuring it is optimally structured for subsequent machine learning analysis. The integrity of this dataset is foundational to the success of the study, laying the groundwork for accurate predictions and meaningful insights into Acute Aortic Syndrome (AAS).

**2. Exploring Feature Extraction Methods and Data Splitting Strategies for Enhanced Classifier Performance:** This step involves experimenting with various feature extraction methods and data splitting strategies to prepare the data effectively for combination with different classifiers in the subsequent step. The goal is to optimize the dataset by selecting the most relevant features and determining the best way to divide the data for training and testing. By exploring

different techniques, the aim is to improve the performance of the classifiers and ultimately enhance the accuracy of predictions for Acute Aortic Syndrome (AAS).

**3. Machine Learning Model Development:** The second objective revolves around the development of a specialized machine learning model tailored explicitly for predicting AAS. Leveraging the integrated dataset, we seek to harness the potential of machine learning algorithms to discern intricate patterns and relationships within the extensive cardiovascular patient data. The goal is to create a robust and accurate predictive model that goes beyond traditional diagnostic methods, offering a cutting-edge approach to identifying individuals at risk of AAS based on comprehensive clinical and medical information.

**4. Performance Evaluation:** Following the development of the machine learning model, the third objective involves a comprehensive evaluation of its performance in predicting AAS. Rigorous testing and validation protocols will be employed to assess the model's Area Under the Curve (AUC) values of the Receiver Operating Characteristic (ROC), sensitivity, and specificity. This step is critical in ensuring the reliability of the predictive capabilities, providing insights into the model's strengths and potential areas of improvement. The outcomes of this evaluation, particularly sensitivity, specificity, and AUC values, will contribute to the credibility and applicability of the developed machine-learning model.

**5. Comparative Analysis of Machine Learning Models and Prediction Features:** This objective involves a thorough comparison of the results obtained from various machine learning models to determine their effectiveness and accuracy in predicting AAS. The primary focus is on evaluating metrics like AUC, sensitivity, and specificity to identify the most reliable methods. The

goal is to pinpoint important features that can aid doctors in accurately diagnosing this disease in the Emergency department.

## 1.5 Contributions

### 1.5.1 Novelty of this Research

**1. Unprecedented Scale and Impact of a Comprehensive Patient Dataset:** This study distinguishes itself by leveraging a comprehensive and recent dataset that includes information from around 150,000 patients across the years 2021-2022. The dataset is sourced from 150 emergency departments in Canada and the USA, making it one of the most extensive and up-to-date datasets in current research. Encompassing a broad spectrum, the dataset comprises detailed information in four key categories: demographics, laboratory results, vital values, and procedures administered to the patients.

2. **Enhanced Machine Learning Capability:** This research introduces a pioneering approach by applying machine learning techniques to clinical datasets for predictive purposes, specifically targeting Acute Aortic Syndrome (AAS). Previous studies have primarily focused on statistical analysis, with only a small portion employing machine learning techniques, often focusing on subsets of AAS such as AAD and PAU. Therefore, this study represents the first instance of utilizing machine learning for AAS prediction, enhancing the field's predictive capabilities.

### 1.5.2 Accuracy and Results

In this study, the machine learning models demonstrated an outstanding accuracy rate of 99.3%. This achievement signifies a remarkable improvement compared to prior research endeavors. The substantial increase in accuracy is a testament to the efficacy of the methodologies employed and the significance of utilizing a comprehensive dataset.

The primary focus of this heightened accuracy was directed towards predicting patients who might be at risk of suffering from Acute Aortic Syndrome (AAS). The models, trained on the extensive dataset, showcased an unparalleled ability to discern patterns and associations within the data, enabling more precise identification of individuals susceptible to AAS based on their experimental results.

The exceptional accuracy achieved in this study not only marks a substantial improvement in predictive capabilities but also holds profound implications for advancing the field of cardiovascular medicine and enhancing the accuracy of AAS detection and prognosis.

## 1.6 Thesis Outline

This thesis is organized into five main chapters, each serving a specific purpose in the overall narrative of the research. Here is a brief overview of what each chapter entails:

**Chapter 2: Literature Review** - This chapter provides a comprehensive review of the existing literature in the fields of Acute Aortic Syndrome and some subsets in this disease such as Aortic Dissection, and machine learning. It explores previous research on the use of clinical datasets with machine learning in diagnosing AAS, highlighting the current state of knowledge, and identifying gaps that this study aims to address.

**Chapter 3: Data and Pre-processing -** This chapter extensively explores the initial phases of data integration, collection, and pre-processing. It outlines the methodologies and factors utilized in acquiring the clinical dataset sourced from 150 Emergency Departments in the USA and Canada. The chapter delves into the pre-processing procedures implemented to guarantee the data's quality and pertinence for subsequent analysis, encompassing techniques for handling missing data.

**Chapter 4: Methods Strategies –** This chapter provides a detailed exploration of the methodology applied to the dataset prepared in the preceding chapter of the study. Focusing on feature extraction, the section outlines specific techniques employed in this process. Additionally, it elucidates the procedures involved in selecting and training machine learning models, along with the metrics used for evaluating model performance. This chapter offers an in-depth examination of the analytical methods employed throughout the study.

**Chapter 5: Results and Discussion** - This chapter presents the results obtained from the machine learning model. It provides a detailed analysis of these results, discussing them in the context of the study's objectives and the existing literature. It also explores the implications of these results for the diagnosis of depression.

**Chapter 6: Conclusion and Future Work -** The final chapter concludes the thesis by summarizing the key findings of the study. It discusses the implications of these findings and their contribution to the field. It also suggests potential avenues for future research, building on the work done in this study.

By following this structure, the thesis aims to provide a comprehensive exploration of the potential of machine learning in diagnosing AAS in the Emergency department using a clinical data set.

# Chapter 2
# Literature Review

## 2.1 Introduction

Acute Aortic Syndrome (AAS) stands out as a critical medical emergency necessitating prompt and precise diagnosis to mitigate potentially devastating consequences. Over the past decade, a significant body of literature has emerged, reflecting collaborative efforts to deepen our understanding of AAS, refine diagnostic strategies, and improve patient management. This literature review endeavors to comprehensively survey and synthesize the extensive research conducted on AAS, encompassing diagnostic models, predictive tools, machine learning applications, and evolving methodologies.

AAS, often referred to as the "great masquerader" due to its varied clinical presentation, demands innovative approaches for identification, prediction, and management. This review meticulously examines seminal studies, navigating the landscape of AAS research to unveil the evolution of diagnostic criteria, the integration of advanced predictive models, and the adoption of cutting-edge machine learning techniques.

A multidisciplinary approach involving clinical, radiological, and computational methodologies emerges as essential in unraveling the complexities of AAS. In subsequent sections, this exploration delves into diverse studies shaping the contemporary landscape of AAS research, addressing critical aspects like risk prediction, diagnostic accuracy, and the integration of machine learning. Through this synthesis, the aim is to consolidate existing knowledge, identify gaps, and pave the way for future research endeavors aimed at refining the approach to diagnosing and

managing AAS. Table 2.1 succinctly summarizes key studies in AAS research, showcasing diverse methodologies and advancements in diagnostic and predictive approaches.

## 2.2 Literature Review of AAS

In studies related to AAS and its subsets, such as Acute Aortic Dissection (AAD) criteria, the key measure of success is primarily assessed through the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), along with sensitivity, and specificity. Analyzing these parameters allows for a comprehensive statistical evaluation of the results in the below-mentioned research studies and facilitates meaningful comparisons between them.

In some research, the choice of methodology for predicting AAS is influenced by the type of dataset sources available. Some studies utilize Computed Tomography (CT) Scans and Magnetic Resonance Imaging (MRI), while others combine Clinical Variables and CT images. As a result, researchers adopt a hybrid approach, incorporating both clinical and imaging data. Additionally, some studies concentrate specifically on clinical measurements and demographic data, employing machine learning techniques. Various researchers utilize a diverse array of machine-learning strategies to analyze these datasets.

In the study by Ohle et al. [6], the overall annual incidence of AAS was found to be 0.61 per 100,000, which was lower than other population-based studies. In the 14 days before diagnosis, 12.5% of patients were seen in the Emergency Department (ED) with a presentation consistent with AAS. The research encompassed a total of 1,299 incident cases of AAS spanning from 2003 to 2018, following the diagnostic phase and the exclusion of numerous cases from a merged dataset of 5,204 patients. The burden of mortality was notable, with one-year mortality decreasing from

47.4% to 29.1%, and ED mortality reported at 14.9%. The study demographics revealed 1,299 cases of AAS, with a mean age of 68.03 and 38.5% female representation, while rural areas accounted for 8.6% of cases. It's noteworthy that the majority of AAS cases in this study originated from urban areas.

Chua et al. [20] conducted a 10-year single-center retrospective review from January 1998 to December 2008, examining records with an "aorta dissection" diagnosis. The study, involving 133 records, found that neither age, gender, nor a history of hypertension were significant risk factors for the missed diagnosis of AAD. The likelihood of a missed diagnosis was significantly higher in the absence of a pulse deficit. Surprisingly, well-known risk factors for AAD, such as age, male sex, and hypertension, were not associated with a higher risk of missed diagnosis when presenting in the Emergency Medical Department. The absence of a pulse deficit or a widened mediastinum did not exclude the diagnosis of AAD. Out of 68 patients included in the analysis, 38.2% had a missed diagnosis during the 10-year study period.

Otani et al. [21] employed a retrospective, single-center, observational design for a post hoc analysis of consecutive AAS patients aged ≥15 years between April 2011 and March 2021. Their study, utilizing CT results, revealed elevated D-dimer levels from the early stages of AAS. The study included 344 patients and aimed to investigate a strategy that combines risk score and D-Dimer levels. This comprehensive approach highlights the continuous effort to refine risk assessment strategies in AAS through the integration of risk scores, D-Dimer levels, and advanced imaging techniques. The clinical utility of D-dimer was observed to be unaffected by the time interval from AAS onset to D-dimer measurement but was influenced by AAS characteristics. Data were collected from patients who underwent both contrast-enhanced CT and unenhanced CT, with exclusions for traumatic cases or those with cardiac arrest before CT.

In their study, Kodolitsch et al. [22] ensured meticulous data collection using custom study forms, entering information twice into an Oracle database to maintain transcription accuracy. The development of the prediction rule involved the use of S-Plus and multivariable logistic regression, a statistical method commonly employed in various studies for purposes such as assessing AAS pretest probability, validating diagnostic accuracy, and evaluating the risk of aortic dissection based on individual signs or symptoms. Their logistic regression model, with only 0.6% missing data, was utilized to assess the risk of aortic dissection. The study integrated significant factors into a multivariate model through stepwise procedures, including a simplified "aortic pain" variable. The study identified aortic pain with immediate onset, a tearing or ripping character, or both; mediastinal widening, aortic widening, or both on chest radiography; and pulse differentials, blood pressure differentials, or both as independent predictors of acute aortic dissection. The probability of dissection was stratified into low, intermediate, and high groups based on the presence of these variables, allowing for improved patient care through better selection for prompt diagnostic imaging.

In Zaschke et al.'s [23] study, continuous variables were reported either as means and standard deviation or as median and first and third quartile, based on appropriateness. The study utilized both univariable and multivariable logistic regression, employing Sigma Stat and SPSS Statistics for statistical analyses and setting a significance threshold at $p < 0.05$. The early suspicion of aortic dissection significantly shortened the median time from pain to surgical correction from 8.6 h in patients with an initial misdiagnosis to 5.5 h in patients with the correct initial diagnosis ($p<0.001$). Of all acute type A aortic dissection patients, 49% had a positive Aortic Dissection Detection Risk Score.

In a prospective study by Goldschmiedt et al. [24], a non-restrictive clinical decision rule for suspected AAS in Emergency Department (ED) patients was assessed. The study, conducted from July 2013 to August 2014, stratified patients into low and high-risk groups based on decision rule results, excluding those with acute trauma, prior AAS, or aortic surgery. Concurrently, CT dose reduction protocols were implemented as a quality improvement measure. The study cohort showed a lower CT utilization rate compared to the historic cohort, with a trend toward higher CT diagnostic yield. AAS incidence remained similar, while the mean effective radiation dose significantly decreased. However, the clinical decision rule exhibited poor performance in risk-stratifying patients for AAS, correctly identifying only 56% of patients with AAS as high-risk. Despite this, its implementation was associated with a significant and safe decrease in CT utilization.

Xu et al. [25] conducted a national retrospective study from 2010 to 2020, identifying 1,585 patients with AAS among 4,064 cases. The overall incidence of aortic dissection, including out-of-hospital cases, was 3.13 per 100,000 persons-year, with an annual increase of 3%, primarily driven by Type A cases. The study revealed consistent management strategies and 30-day mortality rates of 31.9% for type A and 9.7% for type B AAS. Despite advancements, mortality after AAS remains high, and the disease burden may rise with an aging population, highlighting the importance of disease prevention and addressing ethnic disparities.

In patients with low clinical probability of acute aortic syndromes (AASs), deciding on advanced aortic imaging poses challenges. Bima et al. [26] systematically reviewed cross-sectional studies assessing the integration of the Aortic Dissection Detection Risk Score (ADD-RS) with D-Dimer (DD) for AAS diagnosis. In this review of studies after screening 680 studies, 3,804 patients were included. Despite methodological limitations, the integration of ADD-RS with DD showed high

sensitivity in diagnosing AASs. Pooled sensitivity for ADD-RS = 0 and DD < 500 ng/mL was recorded as 99.9%, for ADD-RS ≤ 1 and DD < 500 ng/mL the sensitivity was recorded as 98.9%, and for ADD-RS ≤ 1 and DD < DD the sensitivity was recorded as 97.6%.

Yellapragada et al. [27] proposed an end-to-end automatic approach for detecting AAS in computed tomography (CT) images. Using convolutional neural networks (CNNs), they extracted N cross-sections along the segmented aorta centerline for each CT scan, forming a new volume, which was then classified using 3D CNN and multiple instance learning (MIL) models. Testing on 2291 contrast CT volumes yielded AAS detection with AUC values of 0.965 and 0.985 using 3DCNN and MIL, respectively. This approach offers potential for expediting diagnosis and treatment of AAS, addressing limitations of previous methods such as labor-intensive data annotation and the underutilization of deep learning techniques.

## 2.3 Statistical Analysis of AAS and Its Subsets

In the realm of AAS research, Morello et al.'s [28] study in 2023 stands out with the smallest patient cohort, involving 128 individuals as highlighted in Table 2.1. This investigation focused on pre-test probability assessment and D-Dimer-based evaluation in patients with prior AAS. The researchers independently applied the Aortic Dissection Detection (ADD) risk score, utilizing SPSS software for statistical analyses and incorporating a prospective collection of clinical information and advanced imaging procedures. In the comparative analysis of the ADD score, Aorta Simplified (AORTAs) score, and D-dimer, the AUC for ADD and AORTAs scores on ROC analysis were reported as 0.74 and 0.72, respectively. Notably, the inclusion of D-dimer significantly enhanced sensitivity to 97%, albeit with a trade-off of lower specificity at 13%. In contrast, the ADD score and AORTAs score maintained higher specificities at 83% and 88%,

respectively. Additionally, the evaluation of copeptin at an optimal threshold of 14 pmol/L demonstrated 78.8% sensitivity and 74.6% specificity in detecting AAS. A noteworthy finding of this study was the prevalence of non-AAS (nAAS) in patients with prior AAS evaluated in the ED for truncal pain and other red-flag symptoms. The rate of nAAS was substantial, reaching 28%. This contrasts with previous ED studies from the same group conducted on unselected patients with similar symptoms, where the prevalence of AASs ranged from 13% to 22%. This suggests that in patients with prior AAS presenting with specific symptoms in the ED, the likelihood of encountering non-AAS conditions is higher than in unselected patient populations with similar complaints.

In the study focused on AAS, McLatchie's [29] research conducted in 2023 initially centered on the identification and analysis of cases in three UK Emergency Departments. Later, the study significantly expanded, encompassing a total of 5,548 patients. This extensive multicenter investigation observed adults attending 27 EDs across the UK, providing a detailed analysis of patients. The findings revealed that only 0.3% of patients with potential AAS symptoms were diagnosed with AAS, while 7% underwent computed tomography angiography. Clinical decision tools that incorporated clinician gestalt emerged as the most promising, but the research underscored the necessity for further prospective work, including the evaluation of the role of D-dimer. The study meticulously considered 44 variables and patient characteristics, comprising history, past medical history, and physical examination findings.

Wang et al. [30] conducted a comprehensive study on the Early Prediction Model for AAS, involving 1,298 patients in both model development and validation phases. Demographic and clinical factors collected upon emergency department admission were utilized to establish predictive variables, employing logistic regression. Independent predictors were identified through

multivariable analysis, resulting in a model with robust accuracy, showcasing an AUC of 0.838 in the development cohort and 0.821 in the validation cohort. The study culminated in the creation of a practical nomogram, highlighting the sequential progression from model development to prediction, thus contributing to a comprehensive understanding of AAS.

In another significant research effort, Ohle et al. [31] and their team conducted a comprehensive study to evaluate the Canadian Clinical Practice Guidelines Risk using a dataset of 1,340 individual cases in a historical case-control study. Recruited from three academic emergency departments spanning from 2002 to 2020, the study focused on 379 cases of AAS with specific exclusion criteria. They employed 17 variables in population-based studies to develop risk prediction tools for AAS, encompassing factors such as abrupt onset pain, severe pain, migrating/radiating pain, hypotension, and clinical suspicion levels, with a dichotomization of three-level risk scores into high and low risk. The recent examination of clinical decision tools revealed a slight improvement in the Risk Prediction score's (RIPP) AUC, reaching 0.8284. Notably, the RIPP score demonstrated outstanding sensitivity at 99.7%, despite a lower specificity of 53%. It excelled in accurately classifying high-risk cases of AAS, supported by a higher C statistic.

In a diagnostic strategy study for AAS conducted by Gorla et al. [32], 17 variables were utilized. The study assessed diagnostic efficacy by applying the ADD risk score and combining it with D-dimer, using measures such as sensitivity, specificity, predictive values, and ROC analysis. The findings revealed that individuals diagnosed with AAS, categorized by the ADD risk score, exhibited significantly elevated D-dimer levels in both 'low probability' and 'high probability' categories. D-dimer demonstrated favorable diagnostic characteristics for AAS detection, with a sensitivity of 97.6%, specificity of 63.2%, and an impressive AUC of 0.91 on ROC analysis. The

study highlights the utility of the ADD risk score in enhancing the diagnostic accuracy of AAS, particularly when coupled with D-dimer levels.

Deng et al.'s [33] study concludes that the combination of D-dimer and ADD risk score, particularly with thresholds of ADD risk score > 1 and D-dimer > 2000 ng/mL, demonstrated excellent diagnostic performance for identifying AAS. The research involved 462 patients, with 262 in the exploration cohort and the remaining 200 constituting the validation cohort, all presenting symptoms suspected to be associated with AAS. Deng employed a two-part approach, combining ADD risk score and D-dimer levels in a retrospective cohort and validating in a prospective cohort. The combined approach achieved an AUC of 0.929 in the exploration cohort and 0.911 in the validation cohort. Deng et al. consistently applied the ADD risk score in their studies on AAS.

In another study by Morello et al.'s [34], the research delves into the dynamics of copeptin in patients with Acute Aortic Syndrome, revealing an increase within the initial hours for the majority of subjects. Despite this, the accuracy of copeptin in diagnosing AAS is considered suboptimal. The study takes a proactive step by exploring the application of machine learning techniques for predicting AAS, with a specific focus on individuals suspected of having the condition. Conducted between January 2014 and July 2017, the investigation included 313 emergency department outpatients aged over 18. The primary emphasis was on diagnosing and predicting AAS, considering 19 different factors, including age and other fundamental information. Continuous variables were presented as the median with interquartile range, and categorical variables as proportions with a 95% confidence interval.

Morello et al.'s [34] study comprised three main components: Patient recruitment and data collection, Biomarker measurement, and Statistical analysis. The evaluation of copeptin's diagnostic accuracy for AAS involved ROC curve analysis and logistic regression models. A comprehensive approach integrated clinical, laboratory, and statistical methods to assess copeptin's diagnostic and prognostic potential in suspected AAS. When combined with D-dimer, this approach yielded an impressive AUC of 0.92, providing a sensitivity of 95.2% and specificity of 46.6% for excluding AAS when both markers fell below their designated thresholds.

In Song et al.'s [35] study, a novel approach combining ADD risk score, age-adjusted D-dimer, and chest radiography demonstrated significant potential in reducing the failure rate of the AAS exclusion strategy. With a participant cohort of 558 individuals (93 with diagnosed AAS and 465 with chest pain but without an AAS diagnosis), this innovative strategy achieved a specificity of 67.8%, and an impressive AUC of 0.928. Utilizing multivariable logistic regression, the study showcased the effectiveness of this comprehensive strategy in lowering the risk of failure associated with excluding AAS. The statistical method employed in the research serves various purposes, including assessing AAS pretest probability, validating diagnostic accuracy, and evaluating the risk of aortic dissection based on individual signs or symptoms.

Fu et al. [36] and colleagues incorporated a total of 30 variables, with 23 aligning with the Acute Aortic Syndrome Risk Satisfaction Score (AAS-RSS), selected based on their significance in logistic regression analysis (B value), combined with the ADD risk score. Fu et al. directed their attention towards innovating a novel risk assessment tool, engaging a cohort of 630 patients. Employing binary logistic regression, they pinpointed independent risk factors for AAS, culminating in the creation of the AAS-RSS. This newly devised score underwent evaluation using a ROC curve. Notably, AAS-RSS exhibited a significantly higher AUC of 0.919 compared to

ADD risk score's 0.734. Impressively, AAS-RSS demonstrated a superior sensitivity of 89.3% in diagnosing AAS, underscoring its potential for early and effective identification, particularly in populations with limited clinical history awareness.

Tong et al. [37] involved 638 individuals with meticulous patient size matching. This cohort comprised 178 cases of AAS and 460 cases of Non-ST-Elevation Myocardial Infarction. The study employed multivariable logistic regression for various purposes, such as assessing AAS pretest probability, validating diagnostic accuracy, and evaluating the risk of aortic dissection based on individual signs or symptoms. Notably, the research emphasized the exceptional performance of D-dimer, identifying an optimal cutoff of 385 ng/mL for AAS onset of 2 hours or more. This yielded a sensitivity of 91.8% and specificity of 91.3%, with the corresponding nomogram showcasing an outstanding AUC of 0.974. The study underscores the importance of considering the time-dependent quality of D-dimer for discriminating AASs, and both nomogram models hold potential clinical utility for evaluating the probability of AAS.

In a separate investigation conducted in 2021, Morello et al. [38] conducted an extensive study involving 3,330 patients, among whom 13% of the initial 1,848 patients were diagnosed with AAS. In the validation cohorts, comprising 1,035 patients in the high-prevalence group and 447 patients in the low-prevalence group, 23% and 11%, respectively, were diagnosed with AAS. The study utilized multivariate logistic regression to assess AAS pretest probability and validate a simplified AORTAs score for enhanced diagnostic accuracy in acute aortic syndromes. The evaluation of diagnostic performance, measured by AUC, revealed significant progress, with the AORTAs score, incorporating six independent predictors, demonstrating an improved AUC of 0.729 compared to the Aortic Dissection Detection Score AUC of 0.697. AORTAs, a simplified score, showed increased sensitivity, improved AAS classification, and a minor trade-off in specificity,

making it amenable to integration with age-adjusted for diagnostic rule-out. Notably, in this study, one of the highest sensitivities, reaching 100% based on Table 2.1, was achieved.

The study conducted by Xing et al. [39] sheds light on crucial factors impacting the in-hospital outcomes of AAD. Analyzing 569 patients from January 2020 to December 2021, the research incorporated a derivation cohort of 449 individuals and a validation cohort of 120. Xing et al. developed a final prognostic model, integrating five predictors and demonstrating robust calibration and discrimination in both cohorts. Their methodology, detailed in the reference, utilized the Shapiro–Wilk test, chi-squared tests, and logistic regression in R and SPSS to identify survival group differences in AAD. Additionally, the research team employed 10-fold cross-validation during hyperparameter tuning in the training data, emphasizing the exploration of advanced machine learning models. The outcome of their efforts is highlighted in the identification of critical predictors of in-hospital mortality through multivariate logistic regression. This has significantly contributed to the development of a prognostic model characterized by strong calibration and discrimination, as evidenced by a noteworthy AUC of 0.833 in validation cohorts.

## 2.4 Prediction of AAS by Machine Learning Techniques

In prior investigations focusing on AAS, researchers adopted varied methodologies to collect data. Some studies relied on immediate medical experiment results and demographic data from patients arriving at emergency departments, while others concentrated on patients admitted to hospitals, incorporating more detailed parameters and epidemiological experiment results. Certain studies utilized standard data entry forms, capturing extensive information on patient demographics, history, clinical presentations, physical findings, imaging study results, and patient outcomes, including mortality.

In October 2020, Duceau et al. [16] applied machine learning methods to prehospital research, involving 976 patients. Their study focused on prehospital triage and utilized 27 variables and patient characteristics. Two prediction models for AAS were formulated, employing logistic regression and an ensemble machine learning approach named Super Learner (SL). Under-triage, representing the percentage of AAS patients not transported to the specialized aortic center, and over-triage, indicating the percentage of patients with alternative diagnoses transported to the specialized aortic center, were key parameters studied. For internal validation, Duceau et al. employed 5-fold cross-validation in their respective studies. The SL algorithm demonstrated superior performance in predicting AAS, attributed to its ability to comprehend intricate relationships and patterns within the data. The enhanced predictive accuracy of the SL algorithm was evident in its higher AUC value of 0.73 compared to the logistic regression model's 0.67 in the validation cohort, indicating superior predictive capabilities.

In Huo et al.'s [40] study, initially considering 526 patients, 34 cases were excluded from the analysis, resulting in a final population size of 492. The study utilized Correlation-based Feature Selection (CFS) to choose a subset of attributes most relevant for classification, considering each feature's utility in predicting the class and addressing intercorrelations among them. Huo et al. applied 10-fold cross-validation during hyperparameter tuning for advanced machine learning models in the training data. In their study, the Bayesian Network outperformed five other methods, achieving a superior AUC value of 0.857. This underscores the effectiveness of the Bayesian Network in their research.

Wu et al. [41] achieved the highest AUC for predicting in-hospital rupture of type A aortic dissection using the random forest technique. The research retrospectively evaluated 1,133 consecutive patients diagnosed with Thoracic Aortic Aneurysm and Dissection (TAAD),

employing the random forest method. A notable finding from the research was the exceptional performance of the Random Forest classification model, achieving an outstanding AUC of 0.994 for training dataset and AUC of 0.752 for testing. This remarkable result indicates an exceptional discrimination ability in the training dataset, accompanied by sensitivity of 51.4% and a specificity of 94.5%. These metrics collectively emphasize the robust predictive capabilities of the model, highlighting its effectiveness in accurately distinguishing and classifying cases within the context of TAAD.

Guo et al. [42] conducted a thorough research study involving 1,344 patients diagnosed with acute aortic dissection, comprising 1,071 survivors and 273 non-survivors. The study utilized five machine learning models, including Logistic Regression, Decision Tree, K Nearest Neighbor, Gaussian Naïve Bayes, and Extreme Gradient Boosting (XGBoost), to predict in-hospital mortality risk. Among these models, the XGBoost model demonstrated remarkable performance with the highest mean AUC of 0.927 across 10 iterations. The study revealed that the XGBoost model was an effective approach for generating accurate and early predictions of in-hospital mortality in patients with AAD. Evaluation metrics further reinforced XGBoost's superiority, showing high accuracy at 0.918 and robust values for sensitivity, specificity, positive predictive value, and negative predictive value.

In Wu et al.'s [43] study, continuous variables were presented as mean ± standard deviation, and categorical variables as frequencies and percentages. A decision tree, utilizing demographic and clinical data from the training set, revealing distinct risk profiles for various aortic dissection types. A nomogram based on logistic regression analysis established a scoring system using 23 selected variables. The data were randomly divided into training (80%) and testing (20%) sets, with missing values imputed using the median of recorded measurements. This pragmatic strategy ensured

effective model training and testing, offering insights into model performance on unseen data while addressing missing values. The study explored various machine learning techniques, with the XGBoost model demonstrating the highest accuracy and robustness. Shapley Additive explanations analysis highlighted factors such as Stanford type A, maximum aortic diameter >5.5 cm, high variability in HR, high variability in diastolic BP, and involvement of the aortic arch impacting in-hospital deaths before surgery.

Lin et al.'s study [44] showcased the impressive performance of a Convolutional Neural Network (CNN) model in predicting acute type A aortic dissection rupture. The CNN achieved a remarkable AUC of 0.99, along with notable sensitivity (93%) and specificity (90%). The investigation identified age, gender, specific biomarkers, and aortic morphological parameters as independent predictors for acute type A aortic dissection rupture, emphasizing the CNN's efficacy in accurately predicting and identifying potential risk factors. Notably, both random forests and CNN outperformed Logistic Regression (LR) in predicting acute type A aortic dissection risk, while the Support Vector Machine (SVM) demonstrated inferior performance compared to LR.

**Table 2.1 Related Work Summary**

| Ref | Vari-able | Sample Size | Preprocessing | Methods | AUC | Sensi-tivity | Speci-ficity |
|---|---|---|---|---|---|---|---|
| B. Duceau [16] | 32 | 976 | 5-fold Cross Validation | logistic regression Super Learner (SL) | 0.73 | 99% | 18% |
| F.Morello [28] | 35 | 128 | Pretest Probability Assessment Median D-dimer levels compared between patients with non-AAS and alternative diagnoses | ADD-RS Plus D-Dimer | 0.74 | 65% | 83% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| R.McLatchie [29] | 44 | 5,548 | Constructing exact binomial 80% confidence intervals around key proportions. Handling missing data in the calculation of performance indices for clinical decision tools (CDTs). | Multicenter observational EDs assessing with potential AAS, using both prospective and retrospective recruitment, with anonymized patient data collection | 0.958 | 95% | 97% |
| D.Wang [30] | 17 | 1298 | Internal validation: 5-fold -Cross Validation | logistic regression | 0.838 | 84% | 91% |
| R.Ohle [31] | 17 | 1,340 | Continuous variables presented as means ± standard deviation. Categorical variables presented as frequencies and percentages. Two-sided chi-square or Fisher's exact test. | ADDRS, RIPP Von Kodolitsch AORTAs LOVY | 0.828 | 99.7% | 53% |
| R.Gorla [32] | 17 | 376 | Continuous variables presented as means ± standard deviation; categorical variables presented as frequencies and percentages | ROC analysis used to determine the sensitivity and specificity of D-d at the cutoff value of 0.5 mg/l | 0.91 | 97.6% | 63.2% |
| L.Deng [33] | 12 | 462 | Categorical data described as numbers and percentages. Continuous data displayed as mean with standard deviation (SD) or median with interquartile range | ADD-RS Plus D-Dimer | 0.911 | 93.1% | 70.2% |
| F.Morello [34] | 19 | 313 | Categorical variables reported as proportions plus 95% confidence interval (CI). Mann-Whitney U-test for non-parametric continuous variables. Kruskal-Wallis with post-hoc Dunn's multiple comparison test for independent samples. Pearson's $\chi 2$ test for categorical variables. | logistic regression | 0.92 | 95.20 % | 46.60 % |
| D.Song[35] | 7 | 558 | Chi-square test used for categorical variables. Normality tests performed for continuous variables. | Multivariate Logistic regression | 0.928 | 100% | 67.8% |
| F.Fu [36] | 23 | 630 | Two-tailed Student's t-test $\chi 2$ test employed for categorical variables. Independent sample Mann–Whitney U test used for non-normally distributed data. | logistic regression | 0.919 | 89.3% | 80.6% |
| F.Tong [37] | 22 | 638 | Patients categorized into subgroups based on the onset time (<2 h and ≥2 h). t-test, Mann–Whitney U test, or Chi-square test used for normally distributed continuous variables, non-normally distributed continuous variables, and categorical variables, respectively. | Multivariate logistic regression | 0.974 | 91.8% | 91.3% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| F.Morello [38] | 12 | 3330 | Feature Coefficient and heat maps | Multivariable logistic regression | 0.729 | 100% | 43.1% |
| L.Xing [39] | 5 | 569 | Internal validation :20 repetitions of 10-fold Cross validation<br>External validation | Multivariate Logistic regression<br>SPSS set P<0.05 | 0.833 | 91% | 87% |
| D.Huo [40] | 15 | 492 | CFS Feature selection<br><br>10-fold Cross validation | Bayesian Network, Naïve Bayes, SMO<br>J48 (Decision tree) | 0.857 | 93.1% | 70.2% |
| J.Wu [41] | 16 | 1133 | Divided the sample into training (70%) and testing (30%) data | Random Forest | 0.752 | 99% | 51.4% |
| T.Guo [42] | 41 | 1,344 | Treatment Encoding<br>Type of AAD and Gender Encoding<br>Binary Features Encoding<br>10-Fold Cross-Validation | Logistic regression<br>Decision tree<br>Extreme Gradient Boosting<br>nearest neighbor<br>Gaussian naive bayes | 0.927 | 72.9% | 96.6% |
| Z.Wu [43] | 23 | 380 | Divided the sample into training (80%) and testing (20%) data | Logistic regression<br>Decision tree<br>Extreme Gradient Boosting<br>Random Forest (RF)<br>Support Vector Machine (SVM) | 0.926 | 85% | 93% |
| Y.Lin [44] | 10 | 200 | Shapiro-Wilk test used to determine normal distribution.<br>Normal distribution data presented as (mean SD), compared with t-tests.<br>Non-normal distribution data presented as median (M) and interquartile range , compared by two independent sample rank sum tests. | Logistic Regression<br>Random Forest<br>SVM<br>CNN | 0.99 | 93% | 90% |

## 2.5 Machine learning Methods on Clinical Data in Other Diseases

The majority of studies focus on assessing the robustness of the mentioned machine learning models using techniques such as K-fold cross-validation, typically with K = 10 or 5, aimed at addressing overfitting and simplifying model complexity.

In the study by Rahim et al. [45], they introduce the Machine Learning-based Cardiovascular Disease Diagnosis (MaLCaDD) framework designed to predict cardiovascular diseases with high precision. The framework addresses missing values using mean replacement and tackles data imbalance with the SMOTE. Feature Importance technique is then applied for selection, and an ensemble of Logistic Regression and K-Nearest Neighbor (KNN) classifiers is proposed for enhanced prediction accuracy. Comparative analysis shows that MaLCaDD outperforms existing state-of-the-art approaches, even with a reduced set of features. Major contributors to cardiovascular diseases include high blood pressure, family history, stress, age, gender, cholesterol, Body Mass Index (BMI), and an unhealthy lifestyle. While various diagnostic approaches consider these factors, improving accuracy is crucial due to the critical nature and life-threatening risks associated with cardiovascular diseases.

In their extensive study, Li et al. [46] investigated data from 1,637 patients with AAS to discern postoperative Acute Renal Failure (ARF) prevalence and contributing factors. Utilizing the Shapley additive explanations (SHAP) method, they selected 15 crucial variables from an initial set of 134 features and employed the k-nearest neighbors' approach for missing data imputation. Employing SHAP analysis again for feature importance, the study assessed four machine-learning models (XGBoost, Logistic Regression, SVM) across six centers, validated in three external medical centers, with an evaluation based on AUC of the ROC curve. Notably, the XGBoost algorithm outperformed other models with an AUC of 0.82, leading to the development of an online application. Addressing missing data through imputation methods, the study's predictive model aids in the early detection of high ARF risk, allowing clinicians to undertake timely measures for the prevention and treatment of ARF, a common major complication following cardiac surgery.

At the opposite end of the scale, Klaudel et al.'s [47] research stands out with the largest patient cohort in this study, scrutinizing 84,223 records of diagnostic and interventional coronary procedures spanning from 2000 to 2022. This extensive analysis meticulously identified 124 cases of Coronary Inflammatory Disease (CID) within electronic databases. During the assessment, Klaudel et al. implemented a three-fold cross-validation due to the limited occurrences of dissections in the validation dataset. The dataset was then randomly divided into a training set comprising 70% and a testing set comprising 30%. To address the class imbalance, the training dataset underwent oversampling using the SMOTE. Utilizing multivariate logistic regression, this study achieved high accuracy (0.997), precision (0.748), and an AUC of 0.833 in validation cohorts with XGBoost classification. The incorporation of a 3-fold cross-validation approach contributed to effective model optimization, mitigating concerns related to overfitting.

In a noteworthy study conducted in 2020, Lei et al. [48] employed machine learning techniques for predicting acute kidney injury. The research focused on 897 consecutive patients who underwent aortic arch surgery between January 2013 and May 2017. Lei et al. utilized advanced machine learning models and incorporated 10-fold cross-validation during hyperparameter tuning in the training data to assess their classification performance. Among the various machine learning techniques investigated, the gradient boosting machine emerged as the most effective, achieving the highest test AUC of 0.8 for binary classification and 0.71 for multiclass classification, surpassing other methods. Notably, contributors to this superior performance included variables such as body mass index, cardiopulmonary bypass time, and surgery time. This research represents a significant contribution to the use of machine learning in predicting acute kidney injury in the context of aortic arch surgery.

In the construction of prediction models for novel subtypes in patients with abdominal aortic aneurysm, Zhang et al. [49] and their research team compiled 24 preoperative variables and 23 operative and postoperative variables, covering baseline details, comorbidities, and evaluations of pathological vascular anatomy. Employing the K-means clustering algorithm after normalizing the data using the scale function, they determined the optimal clustering (K-value) by calculating the Silhouette Coefficient (SC). This approach facilitated effective and rational classification of patients based on disease characteristics. The study utilized 24 preoperative variables for Unsupervised Machine Learning Algorithm Surgeries (UMLAs), classifying patients into 2 clusters. Subsequently, the analysis of differences in postoperative variables validated the accuracy of clustering by UMLAs. The researchers conducted a thorough analysis of preoperative variables, revealing significant differences between clusters 1 and 2. This analysis led to the identification of six independent clustering prediction factors, encompassing BMI, neck angulation, left common iliac artery diameter and angulation, and right common iliac artery diameter and angulation. These factors collectively contributed to the development of a nomogram with an AUC of 0.933. This comprehensive approach highlights the use of advanced clustering techniques and predictive modeling to refine patient classification, providing valuable insights for surgical planning in the context of UMLAs.

In Ostberg et al.'s [50] study, a comprehensive preparation phase included imputing missing values through multiple imputations, normalizing continuous variables, and employing a 70-30 training-testing split. Diverse machine-learning models, such as the Generalized Additive Model, Random Forest, K-nearest neighbor, Support Vector Machine, Neural Network, and Bayesian Additive Regression Trees, were utilized to address specific challenges. Imputation methods and 10-fold cross-validation enhanced model reliability and generalizability. To tackle time dependence and

competing risks, they incorporated inverse probability of censoring weighting with bootstrap aggregation. Hyperparameter optimization through grid search and AUROC evaluation on the training set optimized model performance. The study focused on six diverse models, achieving peak performance in predicting outcomes related to type B dissection, type B dissection or rupture, and type B dissection, rupture, or all-cause mortality, with AUC ranging from 0.820 to 0.872.

Wang et al. [51] applied LASSO regression with an optimal alpha coefficient of 0.0019, refining feature selection to 30 features and obtaining feature coefficients and heatmaps. Using Pyradiomics, they extracted Radiomics features of Abdominal Aortic Aneurysm (AAA). Through Pearson correlation analysis, and LASSO regression, predictors for Endovascular Aneurysm Repair related severe adverse events were identified. Predictive models including logistic regression, Naive Bayes, Support Vector Machine (SVM), Random Forest (RF) and XGBoost were applied. Wang et al. employed a random division of data into training 80% and testing 20% sets, addressing missing values through median imputation. This pragmatic strategy facilitated effective model training and testing, providing insights into performance on unseen data and ensuring appropriate handling of missing values. Additionally, 5-fold cross-validation contributed to effective model optimization and mitigated overfitting concerns. The Logistic regression model presented superior predictive value with an AUC of 0.93 and an F1 score of 0.91.

Ke et al. [52] conducted a comprehensive study on predictive models for patients with Acute Coronary Syndrome (ACS), involving 6,482 patients grouped into ST-Elevation Myocardial Infarction (STEMI), Non-ST-Elevation Myocardial Infarction (NSTEMI), and Unstable Angina (UA). The study utilized meticulous statistical methods, summarizing continuous and categorical variables, and employed appropriate statistical tests for comparing training and testing sets. Logistic regression, random forest, SVM, and Gradient-Boosting Decision Tree (GBDT) models

were constructed, with GBDT showing the highest predictive accuracy (AUC of 0.918) and top sensitivity (93.8%). Key predictors such as D-dimer, and Killip suggested potential improvements in clinical outcomes for ACS patients. This methodology aligns with broader studies employing multivariable logistic regression for assessing pretest probability, validating diagnostic accuracy, and evaluating the risk of aortic dissection. Data collection involved diverse patient information from electronic medical records, hospital systems, and clinical repositories, with ACS diagnosed.

The study conducted by Emakhu et al. [53] for acute coronary syndrome prediction in emergency care employed a multifaceted approach to address missing data during preprocessing, incorporating one-hot encoding, removal of redundant data, and Multiple Imputations by Chained Equations (MICE) across 58 variables. Three pivotal components, Cost-Sensitive Classification (CSC), Feature Selection (FS), and Resampling (RS), was integrated into a proposed framework (FS + CSC + RS). SMOTE addressed missing values and balanced data, using Canopy and K-means clustering for enhanced minority class samples. Tomek Links identified pairs with different classes, and the resampling process balanced the dataset with SMOTE while removing Tomek links. This comprehensive approach bolstered data integrity and imbalances handling, contributing to robust analyses. Emakhu et al. also applied Adaptive Boost (AdaBoost), Gradient Boosting, and Extreme Gradient Boosting methods, alongside BorutaShap for feature selection. Ensemble learning integrated outcomes from diverse algorithms, achieving superior performance, with 11 significant risk factors for ACS identified among 31,228 patients. The proposed framework demonstrated a high sensitivity of 86.3% and an impressive AUROC of 0.9333, showcasing the efficacy of their comprehensive approach in predictive modeling.

In 2020, Sax et al. [54] employed 58 variables, incorporating an extended logistic regression model with 71 predictors, including the original 13 variables form acute heart failure risk-

stratification rule (STRATIFY) and 58 additional variables from electronic health records. The study extensively applied logistic regression, decision trees, random forests, and boosting methods, with XGBoost achieving the highest AUC of 0.85 during hyperparameter tuning with 10-fold cross-validation. Logistic regression using the original STRATIFY variables showed good accuracy with an AUC of 0.76. Notably, XGBoost demonstrated improvements in sensitivity, specificity, and positive predictive value, particularly at higher-risk thresholds. The research aimed to optimize risk prediction using additional patient clinical data and machine-learning models. The derived model, incorporating machine learning and additional variables, outperformed conventional approaches, highlighting enhanced 30-day risk prediction for acute heart failure patients in the emergency department.

## 2.6 Summary

In summary, the reviewed studies collectively advance the refinement of diagnostic and predictive models for AAS, demonstrating progress across diverse methodologies. The emphasis on the AUC in comparing diagnostic performance underscores the effectiveness of notable models like XGBoost and Random Forest, indicating their accuracy in predicting and classifying cases. These findings contribute to a comprehensive understanding of factors influencing outcomes and risk estimation in Acute Aortic Syndrome and Its subsets.

The survey's notable discovery was the impressive AUC of 0.99 achieved using the CNN in predicting the risk of Acute Aortic Dissection rupture [42]. The dataset was strategically divided, with 70% designated for training and 30% for testing, enhancing the robustness of the model evaluation. Specifically for predicting AAS, the highest AUC recorded was 0.974, accompanied by a notable 91.8% sensitivity, utilizing the Multivariate Logistic Regression [37].

In the domain of AAS research, the study with the smallest patient cohort, as indicated in Table 2.1, is associated with Morello et al. [28], conducted in 2023. This investigation focused on the identification and analysis of 128 cases of AAS evaluated in two Emergency Departments. On the contrary, the study with the largest patient cohort encountered in this research on AAS is McLatchie et al.'s [29] conducted in 2023. Initially centered on the identification and analysis of cases in three UK Emergency Departments, this study expanded to include 5,548 patients and considered 44 variables, providing a comprehensive exploration of the subject.

One limitation observed in the mentioned studies is the absence of a sample size calculation. The study recognizes the importance of having at least 10 events for each predictor parameter when developing prediction models, as small sample sizes can compromise statistical power and the reliability of predictive models. However, this comprehensive thesis addresses this limitation by incorporating a substantial total sample size of 148,707 cases with 42 clinical variables data that will be explain more in next section about the data. This marks a significant transition into the realm of big data compared to earlier, smaller studies in the field of Acute Aortic Syndrome (AAS). Significantly, this extensive dataset also holds potential applicability for predicting various heart diseases beyond AAS.

# Chapter 3
# Data and Preprocessing

## 3.1 Clinical Dataset

Clinical datasets in AAS play a pivotal role in advancing our understanding of this complex cardiovascular condition. These datasets consist of a comprehensive collection of patient information, encompassing clinical histories, diagnostic imaging results, laboratory tests, and treatment outcomes [55]. The richness of these datasets allows researchers and healthcare professionals to discern patterns, identify risk factors, and refine diagnostic and treatment strategies. By delving into the nuances of patient data, clinicians can gain valuable insights into the various presentations of AAS, aiding in early detection, risk prediction, and personalized patient care [1], [56]. This wealth of clinical information not only contributes to the scientific knowledge base surrounding aortic disorders but also enhances the precision and effectiveness of medical interventions.

The utilization of clinical datasets in AAS research is a dynamic process that involves data integration, cleaning, and analysis [31], [57]. Researchers take great care to sort and arrange data, fixing issues such as missing information to ensure its accuracy. Afterward, they employ advanced computer techniques, like machine learning, to uncover hidden patterns in the data. The results of these analyses assist doctors in making evidence-based decisions, forming treatment plans, and

ultimately improving patient health. Constantly refining and adding more information about AAS allows for ongoing progress in heart medicine. This helps us better understand the condition and develop specific ways to treat it, including finding effective ways to predict the disease in hospital emergency departments.

## 3.2 Structure of Dataset

This dataset is gathered from 150 emergency departments located in both Canada and the USA. It includes the medical histories of close to 150,000 patients, covering the timeframe from 2021 to 2022. It includes information about patients' health, covering both numbers and categories, to give a detailed picture of their health experiences.

The dataset is structured into four main parts—demographics, procedures, lab results, and vitals—enabling a holistic view of personal details, medical treatments, laboratory findings, and vital signs. This thorough collection of information helps healthcare professionals better understand patients' health conditions, enabling them to provide precise and effective medical care. Additionally, the dataset contributes to the objective of predicting AAS early on in Emergency Departments within hospitals.

### 3.2.1 Demographic Dataset

In this particular section of the complete dataset, outlined in Table 3.1, there are 11 distinct files, each containing 6 columns, amounting to a total of 335,201 rows. The Visit-UID acts as a unique identifier to trace individual patient records. This Visit-UID is consistently employed across diverse datasets, such as procedures, lab results, and vitals, ensuring smooth integration of information across various aspects of each patient's medical history.

### 3.2.2 Procedure Dataset

In this segment of the dataset, presented in Table 3.1, there are 15 individual files, each with 3 columns—Visit-UID, procedure code, and Procedure description—collecting information about various medical procedures carried out on patients. In preparation for analysis, rows with similar medical details were grouped together by assigning them new titles or brief abbreviations. This arrangement enhances the structure of the dataset, making the analysis of medical procedures more efficient.

This organized structure aids in obtaining a clearer understanding of patient categories with AAS and assists in identifying which procedures were conducted or overlooked during their hospital stay.

### 3.2.3 Lab Results Dataset

In the lab results dataset, consisting of 15 separate files with 5 columns and a total of 4,202,716 rows (as indicated in Table 3.1), stands as the largest segment of our dataset. This extensive dataset corresponds to 92,892 unique Visit-UIDs. Notably, a significant number of patients lack lab result values, treated as missing data. In the upcoming segment focusing on handling missing data, detailed explanations will be provided regarding the methods used to effectively address and manage the missing data.

### 3.2.4 Vital Dataset

The Vital dataset, consisting of 15 separate files with 4 columns and a total of 1,048,575 rows (as indicated in Table 3.1), is the second biggest dataset in this research. The vital dataset in this study includes important health measurements that give us a good view of how patients are doing.

These diverse parameters offer a comprehensive overview of patients' physiological conditions, enabling a thorough analysis of vital signs and contributing valuable insights into cardiovascular, respiratory, neurological, and overall health. This information is pivotal for evidence-based clinical decisions and the development of targeted treatment approaches in the context of AAS.

**Table 3.1 Dataset Overview**

| | Categories | Files | Column | Number of Row | Unique Visit-UID |
|---|---|---|---|---|---|
| 1 | Demographic | 11 | 6 | 335,201 | 146,114 |
| 2 | Procedure | 15 | 3 | 308,828 | 109,736 |
| 3 | Lab Result | 15 | 5 | 4,202,716 | 92,892 |
| 4 | Vital | 15 | 4 | 1,048,575 | 57,465 |
| **Total** | | **56** | **18** | **5,895,320** | **148,707** |

## 3.3 Preprocessing

### 3.3.1 Standardization and Normalization

One significant challenge posed by this extensive raw dataset lies in the fact that it was recorded across 150 different emergency departments with varied formats. The initial step in the integration process was to create consistent labels and titles within each of the four distinct dataset categories.

Utilizing the Visit-UID, the integration of distinct pieces of information about each patient was accomplished, consolidating them into a single table. Each row in the final table corresponds to a unique Visit-UID, representing an individual patient. The columns of the table contain values of relevant variables for each patient, effectively organizing and linking diverse information related to the same individual in a cohesive manner.

### 3.3.2 Standardizing Labels for Consistency in Data Analysis

The process of making labels consistent and categorizing them in a common format is known as "standardization" or "label normalization" [58]. This involves ensuring that similar terms or labels across different datasets are unified and follow a standardized format, making it easier to analyze and interpret the data consistently [59]. Standardization helps in creating a common language or taxonomy, reducing ambiguity, and facilitating seamless integration and comparison of information from diverse sources.

### 3.3.3 Standardizing Measurements for Consistency in Data Analysis

The process of making common measurements and uniting values is known as "standardizing" or "normalizing" measurements [58]. This involves converting different units of measurement for the same quantity into a consistent and standardized format [59]. Standardization ensures that all data points are expressed in a uniform unit, making it easier to compare and analyze the information. This process is essential when dealing with datasets that may use various units for the same measurement, such as converting height from feet to meters or weight from pounds to kilograms. Standardizing measurements promotes consistency and accuracy in data analysis.

## 3.4 Preprocessing Implementation Before Integrating Separate Datasets

The implementation of preprocessing before integrating separate datasets involves a series of steps to clean, organize, and standardize the individual datasets [60]. This preprocessing stage is crucial to ensure that the data is in a suitable and uniform format for seamless integration. Common preprocessing tasks include handling missing values, addressing outliers, standardizing units or scales, and resolving inconsistencies in data formats. By performing these preprocessing steps, the aim is to enhance the quality, reliability, and compatibility of the individual datasets, laying the foundation for a more accurate and coherent analysis upon integration [60].

### 3.4.1 Preprocessing Demographic Dataset

The demographic dataset in this section includes crucial details like age, gender, chief complaints, diagnosis disposition, and associated codes. To ensure consistency in labels, a categorization process was undertaken. The final labels for the two mentioned columns are presented in Table 3.2 and Table 3.3.

The "Chief Complaint" column in Table 3.2 records the primary symptoms or reasons prompting patients to seek medical attention, such as chest pain, shortness of breath, and discomfort. It provides essential insights into patients' initial health concerns, guiding healthcare professionals in diagnostic and treatment decisions.

**Table 3.2 Chief Complaint list**

| Chief Complaint list |
|---|
| ATYPICAL-CHEST-PAIN |
| CHEST-CONGESTION |
| CHEST-DISCOMFORT |
| CHEST-HEAVINESS |
| CHEST-INJURY |
| CHEST-PAIN |
| CHEST-PAIN-TIGHTNESS |
| CHEST-PAIN-UNSPECIFIED |
| CHEST-PRESSURE |
| CHEST-WALL-PAIN |
| LEFT-SIDE-CHEST-PAIN |
| PATIENT-CLASS |
| SOB-CHEST-PAIN |

The "DX-Disposition" column in Table 3.3 records the disposition status of patients following medical evaluation, indicating outcomes such as admission, discharge, transfer, leaving against medical advice (AMA), or observation. This information provides a concise summary of the patient's post-assessment status, aiding in understanding the outcomes and decisions made during their healthcare journey.

**Table 3.3 Dx-Disposition list**

| DX-Disposition list |
|---|
| Acute-Care-Facility |
| ADMITTED |
| AMA |
| ANWA |
| Died |
| DISCHARGE |
| ED -OBSERVATION-STATUS |
| GROUP-HOME |
| LEFT-AGAINST-MEDICAL-ADVICE |
| LTWC |
| LWBS |
| LWBS - BT |
| LWOB |
| LWTC |
| Nur-Home |
| OBTAIN-BED-ASSIGNMENT |
| OTH |
| PAT |
| Request -for-Bed |
| RMETA |
| RSTMA |
| SNF |
| TRANSFER |

A notable addition to this segment of dataset is the "AAS" column, based on medical results, indicating that among the 148,707 patients, 129 individuals have been diagnosed with acute aortic syndrome, while the remaining do not exhibit this condition. This supplementary information enriches the dataset by identifying patients with this particular medical issue. This facilitates focused analyses, allowing for targeted medical interventions and predictive modeling based on the presence or absence of acute aortic syndrome.

### 3.4.2 Preprocessing Procedure Dataset

The Procedure Dataset went through a similar fixing and organizing process. In preparation, various medical procedure titles and explanations related to things like AAA repair, ablation, angiogram, antibiotics, anticoagulation, aorta-related treatments, heart surgeries such as AVR, and imaging like CT and ECHO were organized and given shorter names. This helps make the information clearer and more straightforward, making it easier to create a detailed report and understand different medical procedures. The details obtained from this process show the specific treatments patients received during their healthcare experiences (refer to Table 3.4).

**Table 3.4 Procedure dataset list**

| Procedure Dataset list |
|---|
| AAA repair |
| Ablation |
| AD repair |
| Angiogram |
| Antibiotics |
| Anticoagulated |
| AORTA |
| AVR |
| CAD |
| Cardioversion |
| CT abdomen |
| CT aorta |
| CT Chest |

| |
|---|
| CT head |
| Dissection repair |
| ECHO |
| EMS |
| HTN |
| ICU |
| IVF |
| Pacemaker |
| Pain-meds |
| SOB |
| Troponin |

### 3.4.3 Preprocessing Lab Results Dataset

The lab results dataset in this research encompasses a diverse array of biomarkers and measurements related to patients' physiological conditions (as shown in Table 3.5).

The most challenging aspect of this section was dealing with different units and formats for the same items standardizing the units and labels for consistency across all records and creating metadata for clear documentation.

It is divided into six main components, namely D-Dimer (Dim), hemoglobin (HGB), lymphocytes (LYM), neutrophils (NEU), platelets (PLT), and troponins (TRP). Each of these categories encompasses multiple sub-measurements, providing detailed insights into various aspects of patients' blood and biochemical profiles.

For instance, under hemoglobin (HGB), parameters like HBA1C, mean corpuscular hemoglobin concentration, and reticulocyte hemoglobin are included. The lymphocytes (LYM) category covers absolute lymphocyte count, atypical lymphocytes, and reactive lymphocytes, among others. Neutrophils (NEU) include measurements such as neutrophil absolute count, segmented count, and immature neutrophil percentage. Platelets (PLT) encompass details like platelet count, mean platelet volume, and platelet morphology. Lastly, troponins (TRP) involve multiple measurements

related to troponin I and troponin T, including high-sensitivity troponin measurements at different time intervals.

This comprehensive dataset of laboratory results serves as a valuable resource for comprehending patients' blood and biochemical parameters. It enables the exploration of correlations, identification of patterns, and potentially the discovery of diagnostic markers for aortic dissection in various ways across different emergency department (ED) hospitals.

One of the significant challenges in preprocessing this segment is the inconsistency in labeling, especially when different labels are employed for the same parameter across the 150 emergency department (ED) hospitals, along with variations in the units used to record them. For example, D-DIMER values were documented using diverse units such as mg/L, ng/mL, ug/mL (FEU), and mcg/mL, and they were labeled differently despite conveying the same meaning.

When the number of items in the lab result dataset becomes large, as indicated by more than 4 million rows in Table 3.1, the standardization process and unifying units become crucial for enhancing the efficiency of the subsequent modeling steps. Standardizing and unifying units ensure consistency in the representation of data, making it more manageable and comparable.

**Table 3.5 Lab results list**

| Lab Results list |
|---|
| **(Dim)** |
| (Dim)-D-DIMER |
| (Dim)-D-DIMER-ALERT-COMMENT-INTERPRETATION |
| (Dim)-D-DIMER-QUANTITATIVE |
| **(HGB)** |
| (HGB)-HBA1C |
| (HGB)-HEMOGLOBIN |
| (HGB)-MEAN-CORPUSCULAR-HEMOGLOBIN |
| (HGB)-RETICULOCYTE-HEMOGLOBIN |
| (HGB)-VENOUS-BLOOD-OXY-HEMOGLOBIN |

| |
|---|
| **(LYM)** |
| (LYM)-ABSOLUTE-LYMPHOCYTE-COUNT |
| (LYM)-ATYPICAL-ABSOLUTE-LYMPHOCYTE-COUNT |
| (LYM)-ATYPICAL-LYMPHOCYTE |
| (LYM)-ATYPICAL-LYMPHOCYTE-PERCENTAGE |
| (LYM)-ATYPICAL-LYMPHOCYTE-REL |
| (LYM)-LYMPHOCYTE-PERCENTAGE |
| (LYM)-REACTIVE-LYMPHOCYTE |
| (LYM)-REACTIVE-LYMPHOCYTE-ABSOLUTE |
| (LYM)-REACTIVE-LYMPHOCYTE-PERCENTAGE |
| (LYM)-VARIANT-LYMPHOCYTE |
| **(NEU)** |
| (NEU)-BAND-NEUTROPHIL-ABSOLUTE |
| (NEU)-BAND-NEUTROPHIL-PERCENTAGE |
| (NEU)-HYPERSEGMENTED-NEUTROPHIL |
| (NEU)-IMMATURE-NEUTOPHIL-PERCENTAGE |
| (NEU)-IMMATURE-NEUTROPHIL-ABSOULTE |
| (NEU)-IMMATURE-NEUTROPHIL-TOTAL |
| (NEU)-NEUTROPHIL-ABSOULTE |
| (NEU)-NEUTROPHIL-ABSOULTE-DIFF |
| (NEU)-NEUTROPHIL-ABSOULTE-SEGMENTED-COUNT |
| (NEU)-NEUTROPHIL-PERCENTAGE |
| (NEU)-NEUTROPHIL-PERCENTAGE-BUDY-FLUID |
| (NEU)-NEUTROPHIL-SEGMENTED-PERCENTAGE |
| (NEU)-NEUTROPHILS-VACUOLATED |
| **(PLT)** |
| (PLT)-ENLARGED-PLATELET |
| (PLT)-FIBRINOGEN-IN-PLATELET-POOR-PLASMA |
| (PLT)-GIANT-PLATELET |
| (PLT)-HYPOGRANULAR-PLATELET |
| (PLT)-IMMATURE-PLATELET |
| (PLT)-INR-IN-PLATELET-POOR-PLASMA |
| (PLT)-LARGE-PLATELET |
| (PLT)-MEAN-PLATELET-VOLUME |
| (PLT)-PLATELET-CLUMP-IN-BLOOD |
| (PLT)-PLATELET-COUNT |
| (PLT)-PLATELET-COUNT-LESS-THAN-100K |

| |
|---|
| (PLT)-PLATELET-ESTIMATE |
| (PLT)-PLATELET-MORPHOLOGY |
| (PLT)-PLATELET-PRODUCT-READY |
| (PLT)-PLATELET-SLIDE-REVIEW |
| (PLT)-PLATELET-STATELLITISM |
| **(TRP)** |
| (TRP)-BASELINE-TROPONIN T |
| (TRP)-GONADOTROPIN-CHORIONIC-(HCG) |
| (TRP)-GONADOTROPIN-CHORIONIC-(HCG)-QUAL |
| (TRP)-HIGH-SENSITIVITY-TROPONIN I |
| (TRP)-HIGH-SENSITIVITY-TROPONIN T |
| (TRP)-HIGH-SENSITIVITY-TROPONIN T-2HOUR |
| (TRP)-HIGH-SENSITIVITY-TROPONIN T-6HOUR |
| (TRP)-HIGH-SENSITIVITY-TROPONIN T-DELTA |
| (TRP)-TROPONIN I |
| (TRP)-TROPONIN I-2HOUR |
| (TRP)-TROPONIN I-2HOUR-DELTA |
| (TRP)-TROPONIN I-4HOUR |
| (TRP)-TROPONIN I-4HOUR-DELTA |
| (TRP)-TROPONIN I-4TH-GENERATION |
| (TRP)-TROPONIN I-5TH-GENERATION |
| (TRP)-TROPONIN I-5TH-GENERATION-BASELINE |
| (TRP)-TROPONIN I-5TH-GENERATION-DELTA-ABSOLUTE |
| (TRP)-TROPONIN I-5TH-GENERATION-PERCENTILE |
| (TRP)-TROPONIN I-6HOUR |
| (TRP)-TROPONIN I-INTERPRETATION |
| (TRP)-TROPONIN I-MMC |
| (TRP)-TROPONIN T |
| (TRP)-TROPONIN T-DELTA |
| (TRP)-TROPONIN T-INTERPRETATION |

### 3.4.4 Preprocessing Vital Dataset

The vital sign list encompasses crucial indicators for monitoring and evaluating a patient's overall health. For this dataset, both the steps of ensuring uniformity in labels and standardizing measurements and units have been applied. This dataset includes various items with corresponding values, such as blood pressure (Diastolic-BP and Systolic-BP), breathing rate, body temperature

(Temperature-C), pain scores, pulse rate, and details about height, weight, and body mass index (BODY-MASS-INDEX), as outlined in Table 3.6. The incorporation of these measurements provides valuable insights into patients' overall health, particularly in terms of their heart and respiratory conditions.

Besides the usual health signs, the dataset also has special measures like the Glasgow Coma Score for checking neurological health and information about heart monitoring, including heart rate (HEART-RATE-MONITORED), ectopy type, and rhythm. Including these different signs gives us a complete picture of how patients' hearts, lungs, and overall health are working during their medical check-ups.

Additionally, the dataset tells us about oxygen levels (OXYGEN) and records the time each measurement is taken (RECORDED-TIME). This time aspect helps us see how these health signs change over time, allowing us to spot any patterns and understand if there's a connection between these vital signs and the occurrence or development of aortic dissection in this study.

**Table 3.6 Vital Sign list**

| Vital-Sign list |
|---|
| Diastolic-BP |
| Systolic-BP |
| Respiratory-Rate |
| Temperature-C |
| Pain-Score |
| Pulse-Rate |
| Pulse-Ox |
| Height-CM |
| WEIGHT-KILOGRAM |
| HEART-RATE-MONITORED |
| BODY-MASS-INDEX |
| OXYGEN |
| GLASGOW-COMA-SCORE |
| Ectopy-Type |
| OB-GYN-STATUS |
| BP-ICU |
| Rhythm |
| Mean |
| MAP |
| Intensity |
| RECORDED-TIME |

## 3.5 The Role of Metadata in a Standardized Dataset

The metadata, or data dictionary, serves as a comprehensive guide defining the labels, measurements, and units used in a dataset. After completing the steps of ensuring uniformity in labels and measurements across varied datasets, the metadata becomes a crucial resource for understanding and interpreting the dataset. It provides a detailed description of each variable, including its name (label), the type of measurement it represents, and the unit in which the measurement is expressed.

This robust metadata or data dictionary, can effectively communicate and share information about the dataset, ensuring consistency and clarity in the interpretation of variables. This documentation becomes especially valuable when using the dataset for other research purposes, allowing us to comprehend the nuances of each variable, its significance, and the standardized units used. In summary, the metadata acts as a key reference tool, facilitating transparency, reproducibility, and the meaningful utilization of the dataset in various research contexts [61].

## 3.6 Data Loading and Exploration

To effectively load, analyze, and visualize clinical datasets in Python, several key libraries come into play [62]. Pandas serves as the cornerstone for data manipulation and analysis, offering robust data structures like Data Frames ideal for tabular data. NumPy provides support for numerical operations, essential for tasks involving arrays and mathematical functions [63]. When it comes to plotting and visualization, Matplotlib is a widely used library, while Seaborn builds on Matplotlib, offering a high-level interface for creating attractive statistical graphics.

For machine learning tasks and predictive modeling, Scikit-learn offers a comprehensive suite of tools, including modules for data splitting, model evaluation, and classification reports [63]. Lastly, Stats models is valuable for incorporating statistical models and tests into analysis pipelines. Installing these libraries via `pip` equips users with a powerful ecosystem for handling, exploring, and gaining insights from clinical datasets in Python.

By leveraging the demographic information and data, a comprehensive understanding of the population's distribution across different age groups and gender categories can be obtained. This analysis forms the foundation for exploring patterns, trends, and variations within the dataset, providing valuable insights into the diversity and composition of the studied population.

### 3.6.1 Exploring Demographic Dataset

Table 3.7 details demographic characteristics for two distinct groups: non-AAS (individuals without acute aortic syndrome) and AAS (patients diagnosed with acute aortic syndrome). Firstly, regarding age distribution, the AAS group shows a notably higher average age of 64, compared to the non-AAS group with an average age of 48. The age breakdown highlights differences in population distribution, with the AAS group having a higher percentage of individuals in the 60-80 age brackets, as also illustrated in Figure 3.2. This discrepancy in average age and age distribution suggests potential variations in health demographics and healthcare needs between the two groups, pointing to an increased risk for predicting AAS in the age range observed in the Emergency Department.

The distribution of genders across the two groups reveals significant differences. In the non-AAS category, 45.1% of the population comprises males, and 53.2% constitute females. Conversely, the AAS group exhibits a shift in gender distribution, with males making up 65.1%, and females

accounting for 34.9%, as also indicated in Figure 3.1. Examining the potential health implications, it is important to explore the risk and likelihood of AAS-related diseases in the context of gender within the AAS group. The data indicates that men in the AAS group may encounter a heightened risk, given their predominant representation, with the risk for men being 1.82 times greater than that for women in the same group.

Finally, it is crucial to acknowledge the presence of missing data, evident in both age and gender categories. In the non-AAS group, 1.7% of the data is absent, emphasizing the importance of addressing this issue in the subsequent section. Effectively managing missing data is pivotal for ensuring the reliability of machine learning analyses. Approaches to development of models equipped to handle missing values adeptly are discussed in the following section. This meticulous attention to missing data is essential for maintaining the integrity and validity of conclusions drawn from machine learning methodologies.

**Table 3.7 Demographics dataset details**

| Demographics | Non-AAS (n=148,578) | | | AAS (n=129) | | |
|---|---|---|---|---|---|---|
| | Average | Count | Percentage | Average | Count | Percentage |
| **Age** | 48 | | | 64 | | |
| <18 | | 4,848 | 3.3% | | 0 | 0% |
| 18-40 | | 50,328 | 33.9% | | 11 | 8.5% |
| 40-60 | | 50,796 | 34.2% | | 40 | 31% |
| 60-80 | | 32,691 | 22% | | 60 | 46.5% |
| 80+ | | 7,322 | 4.9% | | 18 | 14% |
| Missing Data | | 2,593 | 1.7% | | 0 | 0% |
| **Gender** | Non | | | Non | | |
| Male | | 66,960 | 45.1% | | 84 | 65.1% |
| Female | | 79,025 | 53.2% | | 45 | 34.9% |
| Missing Data | | 2,593 | 1.7% | | 0 | 0% |

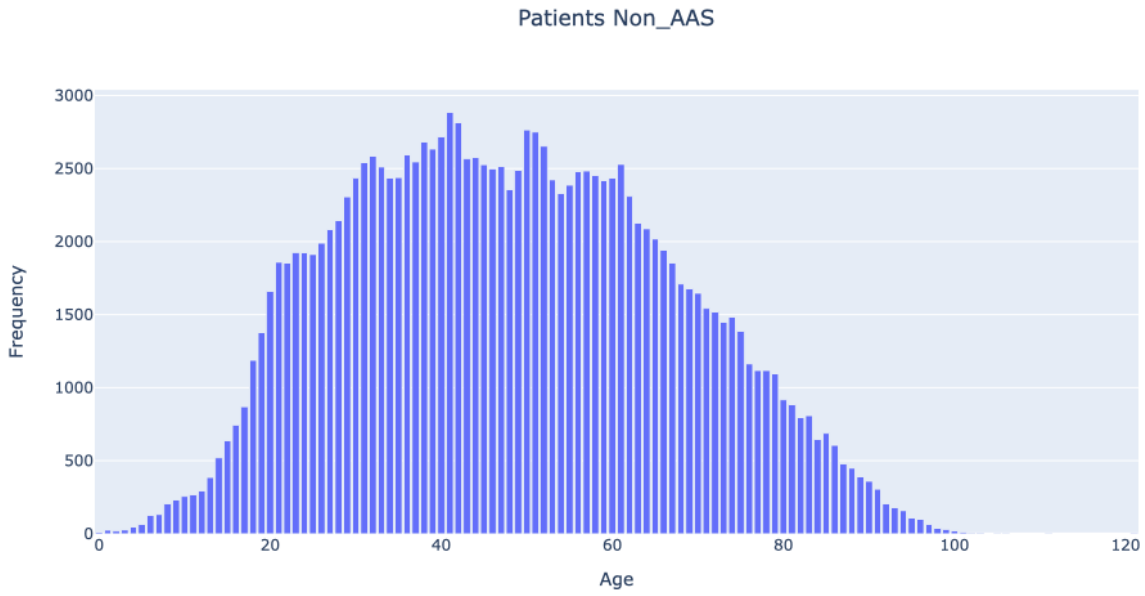


**Figure 3.1 Age Frequency Distribution of Patients non-AAS**

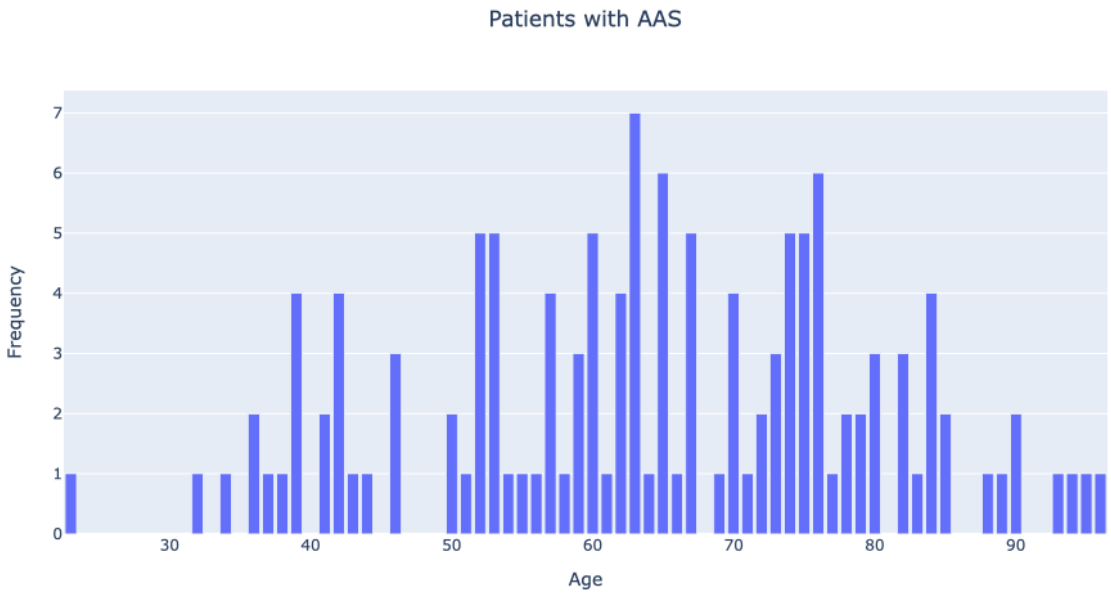


**Figure 3.2 Age Frequency Distribution of Patients with AAS**

The comparison of gender distributions between patients without and with AAS reveals noteworthy findings. In the absence of acute aortic syndrome (AAS) (Figure 3.3), the gender distribution is relatively balanced, with women at 54.13% and men at 45.87%. However, in the

presence of AAS (Figure 3.4), there is a significant shift, indicating a higher proportion of male patients at 65.12%, with females comprising 34.88%. This gender disparity suggests that AAS may exhibit gender-specific patterns or predispositions. Further investigation into the underlying factors contributing to this gender-related difference could provide valuable insights into the characteristics and risk factors associated with acute aortic syndrome in different demographic groups.

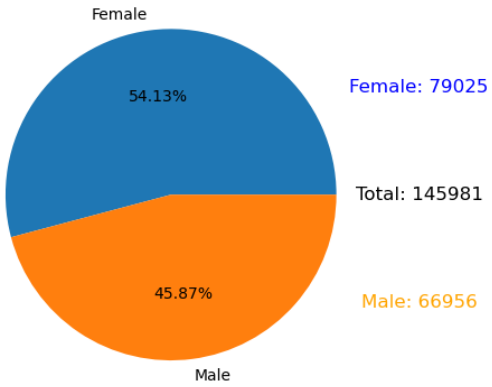


**Figure 3.3 Age Distribution of Patients non-AAS**

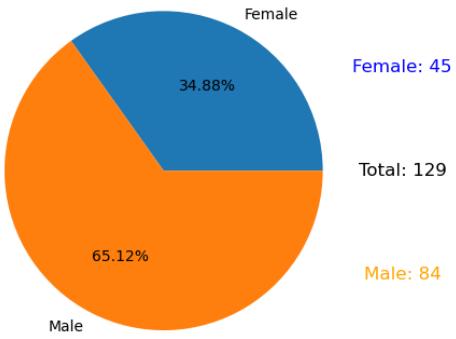


**Figure 3.4 Age Distribution of Patients with AAS**

Table 3.8 delves into the top repeated items of disposition outcomes for individuals categorized into two distinct groups: Non-AAS and AAS patients. Disposition outcomes play a critical role as indicators of the final status of patients within the healthcare system.

In the non-AAS group, a substantial majority, accounting for 79.65%, experienced the outcome of being discharged, suggesting that the majority of individuals sought medical attention but were ultimately deemed fit for release. Moreover, 16.25% were admitted to the hospital, underscoring the need for inpatient care in a significant portion of cases. Other dispositions, such as transfers and cases against medical advice, reflect diverse healthcare scenarios within this group.

In contrast, the AAS group shows a unique trend in disposition outcomes. Significantly, there were no instances recorded as discharged in this group, marking a clear distinction from the non-AAS group. Instead, a majority of cases (52.19%) resulted in hospital admissions, emphasizing a substantial need for inpatient care among individuals in this category. Additionally, a notable proportion (44.02%) underwent transfers to other medical facilities, underscoring the intricate nature of healthcare management for this demographic.

The outcomes in the Emergency Department are particularly noteworthy. In the non-AAS group, a minimal percentage (0.01%) resulted in ED deaths, whereas the AAS group witnessed a higher proportion (3.21%) facing this unfortunate outcome. This difference prompts further examination into potential contributing factors, such as access to healthcare, severity of conditions, and cultural considerations. This investigation aims to provide a better understanding of the varying patterns of ED outcomes between the two groups.

Lastly, it's crucial to acknowledge the cases recorded as 'Died.' While no cases were recorded in the non-AAS group, the AAS group had 2 cases (0.58%) resulting in death. This stark difference underscores the importance of analyzing mortality rates within ethnic populations to identify potential disparities and inform targeted healthcare interventions. Overall, this nuanced

exploration of disposition outcomes provides valuable insights into the distinct healthcare needs and outcomes within the Non-AAS and AAS groups.

**Table 3.8 Analysis of Top Repeated Disposition Outcomes**

| Disposition | Non-AAS | AAS |
|---|---|---|
| DISCHARGE | 245968 (79.65%) | 0 |
| ADMITTED | 50171 (16.25%) | 179 (52.19%) |
| TRANSFER | 9435 (3%) | 151 (44.02%) |
| AMA | 2094 (0.68 %) | 0 |
| LWTC | 405 (0.13%) | 0 |
| LWBS | 254 (0.08%) | 0 |
| Obtain-Bed-Assignment | 140 (0.05%) | 0 |
| Transfer | 135 (0.04%) | 0 |
| ED DEATH | 42 (0.01%) | 11 (3.21%) |
| ANWA | 40 (0.01%) | 0 |
| Died | 0 | 2 (0.58%) |

Table 3.9 provides a detailed analysis of the commonly encountered diagnosis codes for two separate groups: Non-AAS and AAS patients. Diagnoses are represented by specific medical codes, and the percentages offer insights into the prevalence of each diagnosis within its corresponding group.

In the non-AAS group, the most prevalent diagnosis is represented by the code R07.89, constituting 26.21% of cases. This code corresponds to "Other Chest Pain" within the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) coding system [64]. The prevalence of this code suggests a substantial occurrence of chest pain or related symptoms within this group, indicating a focus on respiratory issues. The second most common diagnosis, Z20.822 (6.88%), suggests a focus on preventing health issues or getting vaccines. This code is related to "Contact with and (suspected) exposure to infections with a predominantly sexual

mode of transmission", indicating that some people in the non-AAS group sought healthcare for preventive measures or possible exposure to infections.

Other notable diagnoses include R07.9 (5.96%), I10 (3.28%), and R07.2 (3.19%). R07.9 represents "Chest Pain, Unspecified," and could encompass a range of chest discomforts. I10 corresponds to "Essential (primary) hypertension," indicating cases related to high blood pressure. R07.2 relates to "Precordial pain," which involves pain in the area of the chest over the heart. These diagnoses collectively cover various respiratory and cardiovascular conditions within the non-AAS group.

In the AAS group, the most prominent diagnosis is represented by the code I71.01, accounting for a significant 15.8% of cases. This code is associated with "Dissection of aorta, thoracic," and it aligns with the categorization of the AAS group, indicating that a substantial portion of individuals in this group experienced aortic dissection—a critical cardiovascular condition. The second most common diagnosis in the AAS group is Z20.822 (8.99%), which suggests a focus on preventive health measures or immunizations. While this code might seem unusual in the context of Acute Aortic Syndrome, it could be related to healthcare measures taken during the treatment or diagnosis of AAS cases.

Other notable diagnoses in the AAS group include I71.03 (8.45%), I71.00 (6.81%), and I71.02 (3%). These codes are associated with different types of aortic aneurysms or dissections. I71.03 corresponds to "Dissection of aorta, unspecified," I71.00 to "Dissection of aorta, ruptured," and I71.02 to "Dissection of aorta, abdominal". These diagnoses highlight the prevalence of severe cardiovascular conditions within the AAS group. The absence of occurrences for certain diagnoses (e.g., R07.89, R06.02, R00.2, U07.1, R10.13, R03.0) in the AAS group emphasizes the specificity

of these conditions to the non-AAS group, reinforcing the distinct healthcare challenges faced by individuals diagnosed with Acute Aortic Syndrome.

The breakdown provides insights into common health conditions in non-AAS and AAS groups. Non-AAS reveals chest pain, respiratory, and cardiovascular issues. AAS emphasizes critical cardiovascular conditions, particularly aortic dissections, indicating unique healthcare needs for this severe condition.

**Table 3.9 Analysis of Top Repeated Diagnosis code**

| Diagnosis Code | Non-AAS | AAS |
|---|---|---|
| R07.89 | 83499 (26.21%) | 32 (8.72%) |
| Z20.822 | 21913 (6.88%) | 33 (8.99%) |
| R07.9 | 18981 (5.96%) | 14 (3.81%) |
| R06.02 | 11671 (3.66%) | 0 |
| I10 | 10458 (3.28%) | 17. (4.63%) |
| R07.2 | 10160 (3.19%) | 11 (3%) |
| R00.2 | 5330 (1.67%) | 0 |
| U07.1 | 5125 (1.61%) | 0 |
| R10.13 | 4553 (1.43%) | 0 |
| R03.0 | 4061 (1.27%) | 0 |
| I71.01 | 0 | 58 (15.8%) |
| I71.02 | 0 | 11 (3%) |
| I71.03 | 0 | 31 (8.45%) |
| I71.00 | 0 | 25 (6.81%) |
| R00.1 | 0 | 6 (1.63%) |

This table presents a comprehensive breakdown of the most frequently reported chief complaints by individuals in two distinct groups: non-AAS (those without acute aortic syndrome) and AAS (patients with acute aortic syndrome) group patients. Chief complaints serve as key indicators of the primary reasons individuals sought medical attention, offering a detailed analysis of the prevalent symptoms and concerns within each group.

In the non-AAS group, the overwhelming majority (95.84%) reported "CHEST-PAIN" as their chief complaint, indicating a significant prevalence of individuals experiencing chest pain symptoms. This suggests a diverse range of potential causes for chest pain within this group, ranging from respiratory issues to cardiovascular conditions. Other reported chief complaints in this group, such as "CHEST-WALL-PAIN", "CHEST-PRESSURE", and "CHEST-DISCOMFORT", while less prevalent, still contribute to the overall understanding of diverse symptoms encountered.

Conversely, in the AAS group, an even higher percentage (98.45%) reported "CHEST-PAIN" as the chief complaint, aligning with the critical nature of AAS. This high prevalence underscores the severity and specificity of chest pain as a key symptom in individuals with AAS. Other reported chief complaints, such as "CHEST-WALL-PAIN" and "CHEST-PRESSURE", though less common, further contribute to the nuanced understanding of symptoms in this group.

Notably, "CHEST-DISCOMFORT", "CHEST-INJURY", "CHEST-CONGESTION", and "PATIENT-CLASS" are chief complaints present in the non-AAS group but not in the AAS group. These variations in reported chief complaints highlight the diversity of health concerns encountered in the non-AAS population, underscoring the importance of considering a broad range of symptoms when addressing their healthcare needs.

In summary, this table illuminates the chief complaints reported by individuals in both the Non-AAS and AAS groups, emphasizing the prevalence of chest pain symptoms in both cohorts and shedding light on the unique symptom profiles and healthcare needs within each population.

**Table 3.10 Analysis of Top Repeated Chief Complain**

| Chief Complaint | Non-AAS | | AAS | |
|---|---|---|---|---|
| CHEST-PAIN | 139913 | (95.84%) | 127 | (98.45%) |
| CHEST-WALL-PAIN | 2350 | (1.61%) | 1 | (0.78%) |
| CHEST-PRESSURE | 1196 | (0.82%) | 1 | (0.78%) |
| CHEST-DISCOMFORT | 782 | (0.54%) | | |
| CHEST-INJURY | 436 | (0.3%) | | |
| CHEST-CONGESTION | 421 | (0.29%) | | |
| PATIENT-CLASS | 418 | (0.29%) | | |
| SOB-CHEST-PAIN | 165 | (0.11%) | | |
| CHEST-HEAVINESS | 136 | (0.09%) | | |
| CHEST-PAIN-UNSPECIFIED | 49 | (0.03%) | | |

### 3.6.2 Exploring Lab Results Dataset

Table 3.11 presents a comparison of D-Dimer levels in the Lab Results dataset for two groups: Non-AAS (comprising 148,578 individuals) and AAS (consisting of 129 individuals). D-dimer serves as a blood marker commonly employed to identify the formation or breakdown of blood clots. The unit of measurement for D-Dimer is nanograms per milliliter (ng/mL).

In the non-AAS group, the average D-Dimer level is 768 ng/mL, while in the AAS group, it is higher at 1757 ng/mL. The distribution of D-Dimer levels across different ranges shows variations between the two groups. Notably, the percentage of individuals with D-Dimer levels in the 100-2000 ng/mL range is higher in the AAS group (7.0%) compared to the non-AAS group (9.4%).

Furthermore, examining the Missing Data category reveals that a substantial percentage of individuals in both groups (90.10% in non-AAS and 90.7% in AAS) have missing or unrecorded D-Dimer values. Addressing missing data is crucial for a comprehensive analysis of D-Dimer patterns in relation to aortic dissection.

**Table 3.11 D-Dimer Analysis**

| Lab Results | non-AAS (n=148,578) | | | AAS (n=129) | | |
|---|---|---|---|---|---|---|
| | Average | Count | Percentage | Average | Count | Percentage |
| **D-DIMER (ng/mL)** | 768 | | | 1757 | | |
| 100-2000 | | 13,914 | 9.4% | | 9 | 7.0% |
| 2001-4000 | | 171 | 0.12% | | 2 | 1.55% |
| 4001-6000 | | 77 | 0.05% | | 0 | 0% |
| 6001-8000 | | 16 | 0.01% | | 0 | 0% |
| 8001-10000 | | 454 | 0.31% | | 1 | 0.78% |
| 10001-111082 | | 16 | 0.01% | | 0 | 0% |
| Missing Data | | 133,930 | 90.10% | | 117 | 90.7% |

Table 3.12 presents data on hemoglobin (HGB) levels in the Lab Results dataset for both non-AAS (148,578 individuals) and AAS (129 individuals) groups. Hemoglobin is a vital component of red blood cells responsible for carrying oxygen throughout the body. The unit for hemoglobin measurement is grams per deciliter (g/dL).

In the non-AAS group, the average hemoglobin level is 14 g/dl, and the same value is observed in the AAS group. The distribution of hemoglobin levels across different ranges reveals differences between the two groups. For instance, in the 5-10 g/dl range, the percentage of individuals is higher in the AAS group (2.3%) compared to the non-AAS group (1.2%). Additionally, the proportion of individuals with hemoglobin levels in the 10-15 g/dl range is notably higher in the AAS group (34%) compared to the non-AAS group (25%).

The table also indicates a substantial percentage of missing data, with 63.8% of individuals in the non-AAS group and 45% in the AAS group lacking recorded hemoglobin values. Addressing missing data is essential for a comprehensive analysis of hemoglobin patterns in the context of aortic dissection.

**Table 3.12 Hemoglobin Analysis**

| Lab Results | non-AAS (n=148,578) | | | AAS (n=129) | | |
|---|---|---|---|---|---|---|
| | Average | Count | Percentage | Average | Count | Percentage |
| **Hemoglobin (g/dl)** | 14 | | | 14 | | |
| 0-5 | | 26 | 0.017% | | 0 | 0% |
| 5-10 | | 1,757 | 1.2% | | 3 | 2.3% |
| 10-15 | | 37,150 | 25% | | 44 | 34% |
| 15-20 | | 14,788 | 10% | | 24 | 19% |
| 20-25 | | 21 | 0.014% | | 0 | 0% |
| 25-28 | | 1 | 0.001% | | 0 | 0% |
| Missing Data | | 94,834 | 63.8% | | 58 | 45% |

Table 3.13 provides data on Troponin levels in the Lab Results dataset for both the non-AAS (148,578 individuals) and AAS (129 individuals) groups. Troponin is a protein associated with heart muscle contraction, and its levels in the blood can indicate cardiac injury. The unit for Troponin measurement is nanograms per liter (ng/L).

In the non-AAS group, the average Troponin level is 2 ng/L, while in the AAS group, it slightly rises to 4 ng/L. The examination of Troponin level distribution across various ranges reveals differences between the two groups. For example, in the 0-1 ng/L range, a higher percentage of individuals is observed in the AAS group (41.86%) compared to the non-AAS group (37.13%). Additionally, the proportion of individuals with Troponin levels in the 15-25 ng/L range is higher in the AAS group (1.55%) than in the non-AAS group (0.42%).

The table also underscores the presence of missing data, with 55% of individuals in the non-AAS group and 45% in the AAS group lacking recorded Troponin values.

**Table 3.13 Troponin Analysis**

| Lab Results | non-AAS (n=148,578) | | | AAS (n=129) | | |
|---|---|---|---|---|---|---|
| | Average | Count | Percentage | Average | Count | Percentage |
| **Troponin (ng/mL)** | 2 | | | 4 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 0-1 | | 55,172 | 37.13% | | 54 | 41.86% |
| 1-3 | | 2,888 | 1.94% | | 1 | 0.78% |
| 3-8 | | 5,879 | 3.96% | | 7 | 5.43% |
| 8-12 | | 1,192 | 0.80% | | 4 | 3.10% |
| 12-15 | | 419 | 0.28% | | 1 | 0.78% |
| 15-25 | | 626 | 0.42% | | 2 | 1.55% |
| 25-50 | | 464 | 0.31% | | 1 | 0.78% |
| 50-80 | | 124 | 0.08% | | 0 | 0% |
| 80-130 | | 47 | 0.03% | | 1 | 0.78% |
| Missing Data | | 81,755 | 55% | | 58 | 45% |

### 3.6.3 Exploring Vital Dataset

Table 3.14 offers a summary of diastolic blood pressure (Diastolic-BP) measurements in both the non-AAS (148,578 individuals) and AAS (129 individuals) groups within the Vitals dataset. The unit for diastolic blood pressure is millimeters of mercury (mmHg). Diastolic blood pressure signifies the pressure in the arteries when the heart is at rest between beats.

In the non-AAS group, the average diastolic blood pressure is 85 mmHg, while in the AAS group, it is slightly lower at 83 mmHg. The table illustrates the distribution of diastolic blood pressure across different ranges for both groups. For instance, in the 60-90 mmHg range, a higher percentage of individuals is observed in the non-AAS group (20.13%) compared to the AAS group (10.85%). Conversely, in the 90-120 mmHg range, the AAS group has a higher percentage (19.38%) than the non-AAS group (9.40%).

Moreover, the table highlights the presence of missing data, with 62% of individuals in the non-AAS group and 56.6% in the AAS group lacking recorded diastolic blood pressure values. Addressing missing data is crucial for a comprehensive analysis of blood pressure patterns, especially in the context of aortic dissection.

**Table 3.14 Diastolic-BP Analysis**

| Vitals | non-AAS (n=148,578) | | | AAS (n=129) | | |
|---|---|---|---|---|---|---|
| | **Average** | **Count** | **Percentage** | **Average** | **Count** | **Percentage** |
| **Diastolic-BP (mmHg)** | 85 | | | 83 | | |
| 0-30 | | 64 | 0.04% | | 2 | 1.55% |
| 30-60 | | 7,237 | 4.87% | | 13 | 10.08% |
| 60-90 | | 29,903 | 20.13% | | 14 | 10.85% |
| 90-120 | | 13,963 | 9.40% | | 25 | 19.38% |
| 120-150 | | 3,954 | 2.66% | | 1 | 0.78% |
| 150-180 | | 839 | 0.56% | | 0 | 0% |
| 180-240 | | 70 | 0.05% | | 1 | 0.78% |
| >240 | | 1 | 0% | | 0 | 0% |
| Missing Data | | 92,547 | 62% | | 73 | 56.6% |

Table 3.15 presents a summary of systolic blood pressure (Systolic-BP) measurements in both the non-AAS (148,578 individuals) and AAS (129 individuals) groups within the Vitals dataset. The unit for systolic blood pressure is millimeters of mercury (mmHg). Systolic blood pressure denotes the maximum pressure in the arteries during a heartbeat.

In the non-AAS group, the average systolic blood pressure is 121 mmHg, and in the AAS group, it is slightly lower at 118 mmHg. The table illustrates the distribution of systolic blood pressure across different ranges for both groups. For instance, in the 90-120 mmHg range, the AAS group has a higher percentage (24.81%) compared to the non-AAS group (14.93%). On the other hand, in the 120-150 mmHg range, the non-AAS group has a higher percentage (15.22%) than the AAS group (7.75%).

Additionally, the table indicates the presence of missing data, with 63% of individuals in the non-AAS group and 56.6% in the AAS group lacking recorded systolic blood pressure values. Addressing missing data is crucial for a comprehensive analysis of blood pressure patterns, particularly in the context of aortic dissection.

**Table 3.15 Systolic-BP Analysis**

| Vitals | non-AAS (n=148,578) | | | AAS (n=129) | | |
|---|---|---|---|---|---|---|
| | **Average** | **Count** | **Percentage** | **Average** | **Count** | **Percentage** |
| **Systolic-BP (mmHg)** | 121 | | | 118 | | |
| 0-30 | | 283 | 0.19% | | 0 | 0% |
| 30-60 | | 1,471 | 0.99% | | 3 | 2.33% |
| 60-90 | | 3,884 | 2.61% | | 2 | 1.55% |
| 90-120 | | 22,176 | 14.93% | | 32 | 24.81% |
| 120-150 | | 22,618 | 15.22% | | 10 | 7.75% |
| 150-180 | | 4,941 | 3.33% | | 5 | 3.88% |
| 180-240 | | 571 | 0.38% | | 4 | 3.10% |
| >240 | | 4 | 0.003% | | 0 | 0% |
| Missing Data | | 92,881 | 63% | | 73 | 56.6% |

## 3.7 Missing Data

Missing data presents a pervasive challenge in medical scientific research, particularly in cardiovascular health, where participant withdrawal can markedly reduce the sample size of initially large cohorts. The term "missing data" or "missingness" refers to situations where specific information is absent, potentially introducing bias and consistently diminishing efficiency in analyses [65]. To address missing data, understanding the underlying reasons, known as the missingness mechanism, is crucial for conducting accurate analyses that consider the specific circumstances surrounding the absent information [65].

In the extensive datasets provided, as evident in previous tables, a substantial number of values were observed to be missing. To address this issue and boost method efficiency, the multiple imputation method was employed. Various methods exist for handling missing data, each with distinct advantages and limitations.

### 3.7.1 Missing Data Mechanisms

Various factors can result in missing data, and it's crucial to acknowledge them. The choice of methods to address missing data in statistical analysis relies on assumptions about the underlying mechanisms, emphasizing the importance of understanding the reasons behind missing data [66], [67]. Table 3.16 categorizes missing data mechanisms for predictors into three types: MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Missing Not at Random). It outlines the scenarios associated with each mechanism, providing a foundational understanding for selecting suitable methods to address missing data in statistical analyses.

**Table 3.16 Types of missing data mechanisms for predictors**

| Label | Missing mechanism | Description |
|---|---|---|
| **MCAR** | Missing completely at random | Administrative errors, accidents |
| **MAR** | Missing at random | Missingness related to known patient characteristics, time, or place ("MAR on x"), or to the outcome ("MAR on y") |
| **MNAR** | Missing not at random | Missingness related to the value of the predictor, or to variables not available in the analysis |

#### 3.7.1.1 Missing Completely at Random (MCAR)

Involves missing data occurring entirely randomly, with the probability of data being missing unrelated to the values of any variables, whether observed or unobserved. MCAR implies that the causes of missing data are unrelated to the data itself. An example is a weighing scale running out of batteries, causing some data to be missing due to sheer bad luck. Another instance is a random sample from a population, where each member has an equal chance of being included in the sample [68], [69].

#### 3.7.1.2 Missing at Random (MAR)

Occurs when the probability of data being missing depends on the values of other observed variables but is not related to unobserved variables. If the probability of being missing is the same only within groups defined by the observed data, it is considered MAR. Unlike MCAR, MAR is a broader class, allowing for dependencies on observed variables. For instance, a weighing scale may produce more missing values when placed on a soft surface, making the data MAR if the surface type is known. Modern missing data methods often assume MAR [68], [69].

#### 3.7.1.3 Missing Not at Random (MNAR)

If neither MCAR nor MAR holds, we have Missing Not at Random (MNAR), also known as NMAR (not missing at random). MNAR implies that the probability of being missing varies for reasons unknown to us. An example is a weighing scale mechanism wearing out over time, leading to more missing data as time progresses. Strategies to handle MNAR involve seeking more data about the causes of missingness or conducting sensitivity analyses under various scenarios. MNAR is the most complex case and may involve scenarios challenging to recognize and address [68], [69].

### 3.7.2 Methods for Handling Missing Data in Statistical Analysis

Dealing with missing values is a crucial step in the data analysis process because incomplete datasets can significantly impact the reliability and validity of results. Missing values can occur for various reasons, such as data collection errors, participant non-response, or system failures. To handle this challenge, researchers and analysts use a variety of approaches, each with its own

advantages and limitations. Here's a more detailed explanation of the commonly used methods [68] :

1. **Complete Case Analysis (CCA):** This approach involves analyzing only the cases with complete information and excluding those with missing data. While simple, it may lead to biased results if missingness is not completely random.

2. **Mean or Median Imputation**: Missing values are replaced with the mean or median of the observed values for that variable. It's a straightforward method but may distort the distribution and relationships in the data.

3. **Regression Imputation**: Missing values are predicted using a regression model based on other observed variables. This method is more sophisticated but assumes a linear relationship.

4. **Multiple Imputation (MI):** MI involves creating multiple datasets, each with different imputed values for missing data, and analyzing each dataset separately. It accounts for the uncertainty associated with missing values.

5. **K-Nearest Neighbors (KNN) Imputation:** Missing values are imputed based on the values of their k-nearest neighbors in the feature space. It's a non-parametric method and can capture complex relationships.

6. **Expectation-Maximization (EM) Algorithm:** EM is an iterative method that estimates missing values based on the observed data's likelihood. It's commonly used for data with missing values following a specific distribution.

7. **Last Observation Carried Forward (LOCF):** This method assumes that the most recent observed value for a variable can be carried forward to fill in missing values. It's often used in longitudinal studies.

8. **Data Augmentation:** This Bayesian approach models missing data as additional parameters and estimates them along with the model parameters.

9. **Inverse Probability Weighting (IPW):** This method assigns weights to observed cases to account for missing data patterns, helping to reduce bias.

10. **Bootstrap Methods:** Techniques like bootstrapping can be used to estimate standard errors and confidence intervals while accounting for missing data.

The selection of a method for handling missing values is indeed influenced by several factors, including the nature of the data, the missing data mechanism, and the assumptions that can be reasonably made about the missing values [68]. In the context of this research, multiple imputation was chosen as the preferred method.

### 3.7.3 Multiple Imputation

Multiple imputation is a preferred method for handling missing data due to its ability to account for uncertainty by creating multiple datasets with varied imputed values [70]. This approach preserves the variability in the original data, yielding more accurate standard errors and confidence intervals. It is flexible, allowing for sophisticated modeling techniques and accommodating different types of missingness mechanisms. Multiple imputation reduces bias compared to simpler methods and is compatible with various statistical procedures, making it suitable for diverse research scenarios [70], [71]. The choice of multiple imputation over other methods is driven by

its robustness, flexibility, and ability to provide more reliable and nuanced results in the presence of missing data.

Furthermore, in this method the approach to handling missing values extends to both predictor variables (X) and the outcome variable (y). This method assumes that genuine predictor values are hidden behind the missing values, and it takes into account missingness in both the input and output variables [66]. Traditional statistical approaches often exclude subjects with any missing value from analysis, such as incomplete case analysis, leading to the exclusion of subjects with missing values for any potential predictor or outcome [66], [72].

#### 3.7.3.1 Imputation Principle

The imputation principle involves estimating or replacing missing values in a dataset with plausible substitutes based on statistical techniques [73]. This process aims to address the impact of missing data on the analysis, allowing for a more complete and reliable interpretation of the dataset. Imputation methods substitute the missing values with plausible values so that the completed data can then be analyzed with standard statistical techniques. In some datasets, we may find a characteristic or combination of characteristics that closely defines the predictor with missing values, for example, when variables are strongly related to the same underlying phenomenon. The goal is to produce a dataset that retains as much information as possible while minimizing bias and preserving the statistical properties of the original dataset [74].

Table 3.17 classifies approaches for handling missing values based on the use of predictor variables (X) and outcome variables (Y). The methods include Conditional Mean (CM) for imputation using the conditional mean of X, Single Imputation (SI) involving a random draw from

the predictive distribution of X and Y, and Multiple Imputation (MI) utilizing random draws from the predictive distribution of both X and Y, repeated for increased reliability.

**Table 3.17 Approaches to dealing with missing values, including imputation methods**

| Label | X/Y used? | Approach |
|---|---|---|
| **CM** | X | Performs single imputation using the conditional mean, estimated through a regression model. |
| **SI** | X+Y | Involves single imputation with a random draw from the predictive distribution of an imputation model (stochastic regression imputation). |
| **MI** | X+Y | Utilizes multiple imputation with random draws from the predictive distribution of an imputation model, repeated several times (e.g., 20 times) |

### 3.7.3.2 Imputation Methods

Multiple imputation involves generating several completed datasets instead of a single one. Each dataset is filled multiple times for missing values using independent draws from a model. The goal of multiple imputation is to estimate the actual relationship between missing data and the available information. In essence, it predicts plausible values for the missing data based on other known information for each variable with missing values [65], [73], [71].

Imputing missing values for a predictor when other predictors also have missing values poses a challenge [70]. Addressing the predicament of predicting missing values for one predictor using other predictors with missing values may benefit from employing data augmentation methods. These methods follow an iterative process of imputation and posterior steps until convergence, generating multiple imputed datasets to account for uncertainties associated with predicting missing values from observed data [75], [76].

The choice of the imputation method depends on various factors, including the distribution and characteristics of data, the types of variables, and the underlying patterns of missingness. Table 3.18 provides a comparison of imputation methods for both continuous and categorical variables, highlighting their suitability and considerations.

**Table 3.18 Comparison of Imputation Methods**

| Imputation Method | Continuous Variables | Categorical Variables | consideration |
|---|---|---|---|
| **Mean or Median Imputation** | Suitable for continuous variables | Not applicable | May not be suitable for skewed data or when extreme values are present. |
| **Mode Imputation** | Not applicable | Ideal for categorical variables | May oversimplify the variability in the data. |
| **Predictive Mean Matching (PMM)** | Suitable for continuous variables | Applicable for both continuous and categorical variables | May not work well if the data has complex relationships or if there are outliers. |
| **k-Nearest Neighbors (KNN)** | Effective for capturing local patterns | Applicable | Sensitive to the choice of the number of neighbors (k). |
| **Multiple Imputation by Chained Equations (MICE)** | Versatile for mixed variable types | Well-suited for categorical variables | Assumes the missing data mechanism is Missing at Random (MAR) or Missing Completely at Random (MCAR). |
| **Iterative Imputer (scikit-learn)** | Adaptable for continuous variables | Can handle categorical variables | Assumes a certain level of linearity in relationships. |
| **Regression Imputation** | Appropriate for clear linear relationship | Can be used but assumes linearity | Assumes a linear relationship between variables. |
| **Decision Tree Imputation** | Effective for capturing non-linear and complex relationships | Suitable for both continuous and categorical variables | Prone to overfitting. |
| **Matrix Factorization Methods** | Suitable for high-dimensional datasets | Can be applied but more common for continuous variables | Interpretability may be challenging. |

Based on the characteristics and relationships among our dataset values, the choice between imputation methods is determined. In the case of a dataset where variables are neither continuous nor correlated, the selection of an appropriate imputation technique becomes crucial. Specifically, when dealing with non-continuous variables that exhibit no discernible relationships, Multiple

Imputation by Chained Equations (MICE) emerges as a suitable option. It is particularly well-suited for handling missing values in categorical variables, allowing for the estimation of probabilities associated with certain events.

#### 3.7.3.3 Multiple Imputation by Chained Equations (MICE)

This method involves imputing missing values in a dataset by modeling each incomplete variable conditional on the observed values of the other variables. The imputation equations are created for each variable with missing data. The general idea is to perform imputation in a series of steps, with each step updating one variable at a time. This process is iterated until convergence. The imputation models can vary depending on the nature of the variables (continuous, categorical, etc.) and the assumptions you make about the data [77].

Multiple Imputation by Chained Equations (MICE) is an iterative imputation method that imputes missing values in a dataset by modeling each variable with missing data as a function of other variables. The method is based on the idea of imputing missing values multiple times, creating multiple datasets with plausible imputed values. These datasets are then analyzed separately, and the results are combined to account for the uncertainty introduced by the imputation process [65].

The general equation for imputing a missing value in MICE can be described as follows:

$$Y_i^m = \mu_i + \epsilon_i^m$$

Here, $Y_i^m$ represents the imputed value for the missing data in variable $i$ in the $m$-th imputed dataset. $\mu_i$ is the predicted value for the missing data, and $\epsilon_i^m$ is a random error term sampled from the distribution of the residuals of the predictive model [69].

The MICE algorithm proceeds through multiple cycles, where each cycle involves updating the imputed values for each variable based on the imputed values of other variables. This process is repeated until convergence, and the final imputed values are obtained by combining the results from all imputed datasets [76].

## 3.8 Balancing data with SMOTE

To address the imbalance challenge in the dataset, with only 129 positive cases among 148,500 negative cases, the Synthetic Minority Over-sampling Technique (SMOTE) was employed.

### 3.8.1 What is SMOTE

SMOTE stands for Synthetic Minority Over-sampling Technique. It is a popular method used to address class imbalance in machine learning datasets, particularly in classification tasks. Class imbalance occurs when one class (the minority class) is significantly underrepresented compared to the others (the majority classes).

The main goal of SMOTE is to balance the class distribution by generating synthetic samples for the minority class. In this study, SMOTE was applied to address the imbalance between positive and negative cases of Acute Aortic Syndrome (AAS). With only 129 positive cases among 148,500 negative cases, the dataset faced an imbalance challenge. By generating synthetic samples of the minority class, SMOTE ensured that the predictive model is not biased towards the majority class. This improved the model's ability to accurately identify both positive and negative cases of AAS, thereby enhancing its predictive performance.

### 3.8.2 Application of SMOTE

In this study, as evident in Figure 3.5, before applying SMOTE, the imbalance between the minority and majority classes was evident. In the graph, the minority class appears as a smaller cluster of orange data points compared to the majority class, represented by blue data points. This significant class imbalance can lead to biased model predictions and poor performance.

However, as shown in Figure 3.6 after applying SMOTE, the class distribution becomes more balanced. In the graph, the minority class(es) are better represented, with synthetic data points generated by SMOTE. This helps to alleviate the class imbalance issue and ensures that the predictive model is not biased towards the majority class, resulting in improved performance and more accurate predictions.

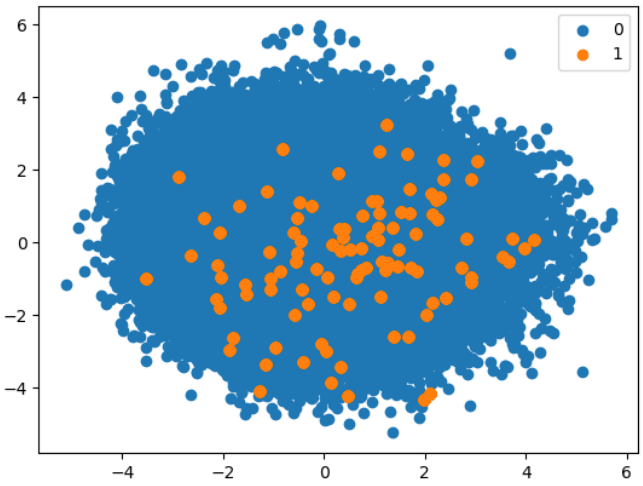


**Figure 3.5 Data set Before Applying SMOTE**

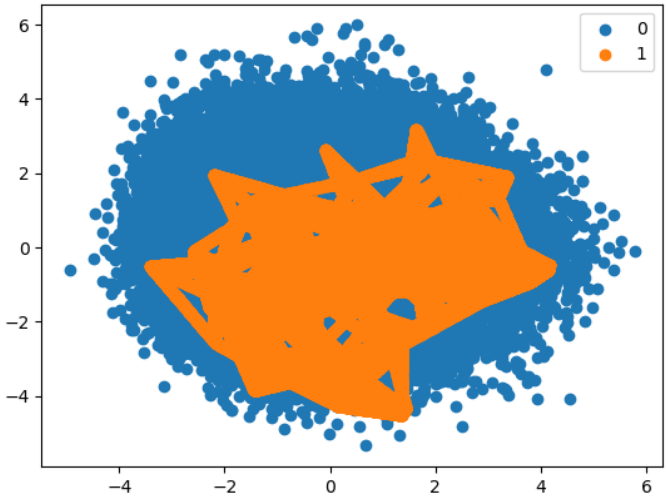


**Figure 3.6 Data Set After Applying SMOTE**

# Chapter 4
# Methods

## 4.1 Methodology

After addressing missing data in the previous step, the focus shifts to preparing the data for the next analytical phase. In this context, the interquartile range (IQR) method is employed. The IQR helps gauge the spread of the dataset by identifying the range within which the central 50% of the data lies. By calculating the first quartile (Q1) and the third quartile (Q3), the IQR highlights the middle portion of the data, excluding extreme values. This information is crucial for understanding data variability and revealing potential outliers.

Once the IQR is determined, its application extends to outlier detection. Values falling outside a specified range, often defined as 1.5 times the IQR, are considered potential outliers. This insight aids in identifying and managing extreme values that might impact the reliability of subsequent analyses. Consequently, the interquartile range method plays a pivotal role in refining the dataset, making it more robust and suitable for the analytical steps that follow.

Following the handling of missing data and the utilization of the interquartile range method, the methodology progresses to feature extraction from the preprocessed data, as indicated in the subsequent step of the flowchart illustrated in Figure 4.1. This step holds paramount importance for various reasons. The original dataset inherently possesses significant complexity and dimensionality, posing considerable computational and analytical challenges. Feature extraction becomes crucial in simplifying and transforming the data, reducing its complexity by identifying and selecting key variables or features that are most relevant to the analysis at hand.

The techniques utilized for feature extraction play a pivotal role in managing the dimensionality of numerical clinical datasets, focusing on distilling the essential elements of the data signal. By reducing dimensionality, these methods expedite subsequent analyses, as highlighted in [78]. This process aims to streamline the information contained in the dataset, emphasizing the most pertinent, aspects and facilitating more efficient and insightful analysis.

In the context of machine learning applications, as illustrated in the flowchart, the significance of feature extraction is crucial. Its primary role is to counter overfitting by eliminating noise and irregularities inherent in the raw data. This contributes to refining the accuracy and resilience of the resulting models [78].

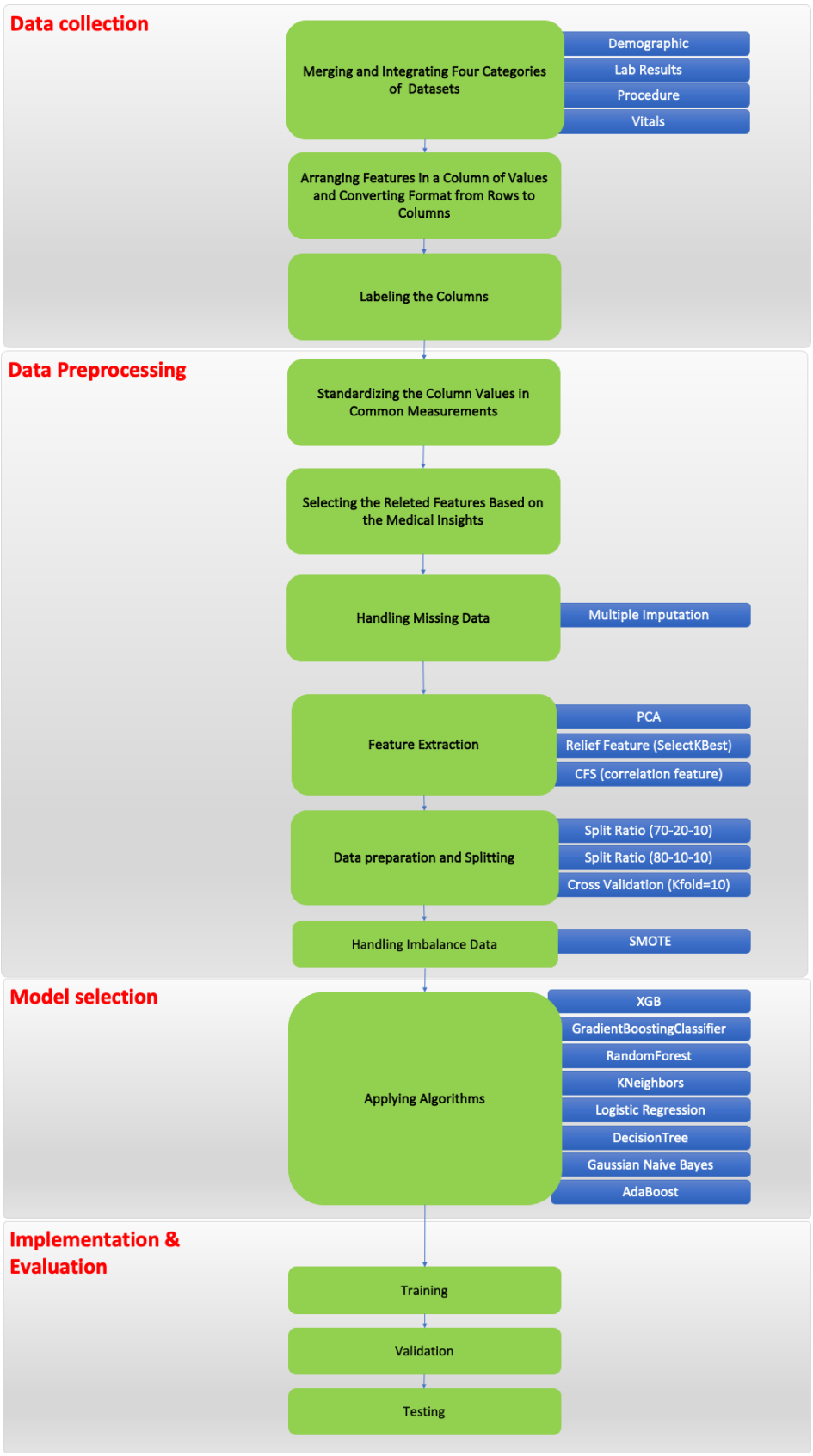


**Figure 4.1 Machine Learning Workflow for Predicting Acute Aortic Syndrome**

As indicated in the flowchart, this chapter will specifically focus on the extraction of features from the clinical dataset. The emphasis is on isolating and highlighting relevant characteristics or attributes from the clinical data, a crucial step in the overall process outlined in the flowchart. This feature extraction process is essential for gaining insights and understanding patterns within the clinical dataset, potentially facilitating further analysis and interpretation. The specificity of the chapter's concentration on this aspect suggests its significance in the broader context of the research or study being conducted.

## 4.2 Feature Extraction

In the pursuit of enhancing the interpretability and efficiency of data analysis, this study strategically utilized three diverse feature extraction methods: Principal Component Analysis (PCA), the Relief feature employing SelectKBest, and the Correlation Feature Subset (CFS) method. PCA, a dimensionality reduction technique, aims to transform the original feature space into a more compact yet representative set of variables, capturing the essential information embedded in the data. Concurrently, the Relief algorithm, coupled with SelectKBest, accentuates the importance of features that effectively discriminate between instances of the target class, ensuring the retention of the most discriminative attributes. Additionally, the application of the CFS method facilitates the identification of feature subsets that exhibit high correlation with the target class while minimizing intercorrelation among the selected features.

Through the integration of these feature extraction methods, the study strives to unravel the underlying patterns and intrinsic structure within the dataset. Each method brings its distinctive approach and criteria to the table, offering a nuanced perspective on feature relevance. This

comprehensive feature extraction process is instrumental in streamlining the dataset for subsequent modeling or interpretation, paving the way for more focused and accurate analyses. The amalgamation of PCA, Relief feature selection, and CFS not only refines the dataset but also enriches the understanding of the key factors influencing the target variable, ultimately fostering a more informed and insightful exploration of the study's underlying data.

### 4.2.1 Principal Component Analysis

PCA, or Principal Component Analysis, is a statistical technique used for dimensionality reduction and feature extraction in data analysis. Its primary objective is to transform a dataset with potentially correlated variables into a new set of variables, known as a linearly uncorrelated principal component [79].

The transformation is designed in a way that the first principal component captures the maximum possible variance in the original data. Each subsequent principal component captures the highest remaining variance while being orthogonal (uncorrelated) to the preceding ones. By doing this, PCA simplifies the dataset by highlighting the most essential information and reducing the dimensionality, making it easier to analyze and interpret.

In essence, PCA helps identify the key patterns or features in a dataset, allowing for a more concise representation of the data while preserving as much variability as possible [79]. This technique finds applications in various fields, including signal processing, image analysis, and bioinformatics, among others.

In the realm of AAS, where datasets may be high-dimensional and include diverse clinical parameters, PCA plays a crucial role in feature extraction. By reducing the dimensionality of the

dataset, PCA enables a more focused analysis while retaining the essential information embedded in the data. This proves particularly beneficial for identifying key factors contributing to the diagnosis, prognosis, or understanding of AAS.

### 4.2.2 Correlation-based Feature Selection

Correlation-based Feature Selection (CFS) is a method used in data analysis and machine learning to identify and retain relevant features from a dataset based on their correlation with the target variable. The primary goal of CFS is to select a subset of features that not only individually correlate well with the target but also have low intercorrelation among themselves [80].

The CFS algorithm evaluates the relationship between each feature and the target variable, emphasizing the importance of features that provide unique information. It considers both the correlation of individual features with the target and the redundancy among features, aiming to select a subset that collectively contributes valuable insights without unnecessary overlap.

Utilizing CFS can simplify datasets and enhance the effectiveness of subsequent analyses, particularly in situations with a high number of features. This approach contributes to improved model interpretability, decreased computational complexity, and frequently leads to the development of more resilient and precise models by prioritizing the most informative features. The applicability of CFS extends to diverse domains such as biology, finance, and healthcare, where the selection of relevant features is vital for comprehending relationships between variables and ensuring accurate predictions [80].

For AAS research, where understanding the relationships between various clinical parameters is vital, CFS aids in selecting a subset of features that collectively contribute to a more comprehensive understanding of the syndrome. By prioritizing features that are both individually

informative and exhibit low redundancy with one another, CFS can enhance the interpretability of models, streamline datasets, and potentially improve the accuracy of predictions related to AAS diagnosis or patient outcomes. The application of CFS in AAS studies aligns with the broader goal of optimizing feature selection methods for more effective and meaningful analyses in the medical domain.

### 4.2.3 Relief Feature (Select-KBest)

The Relief algorithm is a feature selection technique employed in machine learning to sift through a dataset and identify the most crucial features for classification tasks [81]. It operates by comparing each instance in the dataset with its nearest neighbors, assessing how well individual features distinguish between instances of the same class (hit neighbors) and instances of different classes (miss neighbors). The algorithm assigns importance scores to features based on the differences in their values for a given instance and its neighbors. Features that consistently contribute to distinguishing between instances of different classes receive higher scores.

In more technical terms, Relief computes the feature importance scores by iteratively examining a subset of instances and updating the scores based on the differences in feature values [81]. This iterative process provides a nuanced understanding of which features are most relevant for the classification task at hand. After the algorithm has gone through the dataset, it selects the top k features with the highest importance scores, where k is a predetermined or user-defined value.

The practical application of the Relief algorithm is often seen in conjunction with other feature selection methods, such as Select-KBest, where the aim is to streamline the dataset and retain a subset of features that significantly contribute to the predictive power of machine learning models. In essence, Relief aids in distilling complex datasets down to their essential components, facilitating more effective and interpretable machine learning analyses.

In the context of Acute Aortic Syndrome (AAS) or any medical application, the Relief algorithm can be utilized to identify and prioritize relevant features from clinical datasets. AAS is a serious and potentially life-threatening condition related to the aorta, and understanding the key factors contributing to its diagnosis or prognosis is crucial for effective medical decision-making.

Applying the Relief algorithm to a clinical dataset associated with AAS involves examining patient records and relevant clinical parameters. The algorithm assesses the importance of each feature in distinguishing between patients with different outcomes or conditions related to AAS. For example, it might evaluate the significance of blood pressure readings, heart rate, biochemical markers, or other clinical variables in predicting the presence or severity of AAS.

By employing three different feature extraction methods with eight different machine learning techniques, the best integrated method can be found. Principal Component Analysis (PCA), Correlation-based feature selection, and Relief feature selection methods enhance the robustness and interpretability of the predictive model. PCA reduces dataset dimensionality by transforming original variables into a smaller set of uncorrelated variables, preserving important information while reducing computational complexity. Correlation-based feature selection identifies and selects the most relevant features by measuring the correlation between each feature and the target variable, thereby improving predictive accuracy. Relief feature selection is particularly effective in identifying relevant features in noisy and redundant datasets, further refining the feature set and enhancing the model's predictive performance. By integrating these three methods, a comprehensive and informative feature set is obtained, capturing the most relevant information from the dataset while minimizing the impact of noise and irrelevant features. This improves the overall effectiveness and interpretability of the predictive model.

## 4.3 Classification Methods

Classification in machine learning is a supervised learning task that involves categorizing input data into predefined classes or labels based on its features. In this type of problem, the algorithm is provided with a labeled dataset during the training phase, where each data point has a set of features and an associated class label. The primary objective is to learn a mapping from these features to the correct class labels, enabling the algorithm to generalize its learning and accurately classify new, unseen instances.

The classification process relies on identifying patterns and relationships within the training data to create a model that can make predictions on unseen data. Various algorithms are employed for classification tasks, each with its approach and strengths.

The success of classification methods lies in their ability to efficiently learn from labeled data, generalize patterns, and make accurate predictions on new instances, contributing to the advancement of automated decision-making in various domains.

Classification involves a diverse range of algorithms, each possessing distinctive traits and applications. Prominent instances in this study encompass XGBoost (Extreme Gradient Boosting), Random Forest, and Logistic Regression, alongside other algorithms employed for classification tasks.

### 4.3.1 Random Forest (RF)

Random Forest is a machine learning algorithm renowned for its versatility and effectiveness in both classification and regression tasks. At its core, it operates as an ensemble learning technique, harnessing the strength of multiple decision trees to build a more robust and accurate model [82].

Unlike a single decision tree prone to overfitting, Random Forest introduces diversity by constructing a multitude of trees, each trained on a random subset of features and data points. This diversity, achieved through the combination of bagging and random feature selection, mitigates the risk of overfitting and enhances the model's generalization ability [82].

The algorithm employs a voting mechanism to make predictions in classification tasks and average predictions in regression tasks [47]. This ensemble approach ensures that the final prediction is a collective decision from all the individual trees, contributing to the model's stability and reliability. Random Forest's robustness to overfitting and ability to handle complex datasets make it a popular choice in various domains where accurate predictions are essential.

Furthermore, Random Forest provides insights into feature importance, highlighting which features contribute significantly to the model's prediction [43]. This feature importance analysis can aid in understanding the underlying patterns in the data, making Random Forest not only a powerful predictive tool but also a valuable tool for feature selection. The algorithm's application spans diverse fields such as finance, healthcare, and remote sensing, illustrating its adaptability and utility across different domains. Its widespread use is a testament to its effectiveness in tackling complex problems in machine learning.

In the context of AAS, the algorithm's ability to handle numerical features such as blood pressure readings, biochemical markers, and imaging metrics is crucial. By constructing numerous decision trees on different subsets of the dataset, Random Forest captures the complex relationships between these numerical variables and the presence or severity of AAS. This diversity in model construction enables a more nuanced understanding of the numerical clinical parameters contributing to accurate predictions.

### 4.3.2 Logistic Regression (LR)

Logistic Regression is a statistical method commonly used for binary classification tasks, where the outcome variable has two possible categories [16]. Despite its name, it is employed for classification rather than regression. This algorithm is particularly effective when predicting the probability of an event happening or not happening based on one or more predictor variables [36]. In logistic regression, the logistic function (sigmoid function) is utilized to transform a linear combination of input features into a value between 0 and 1. This transformed value represents the probability of the event occurring. If the probability is above a certain threshold (often 0.5), the model predicts the positive class; otherwise, it predicts the negative class.

Unlike linear regression, logistic regression is well-suited for problems where the dependent variable is categorical, making it widely used in fields such as medicine (disease diagnosis), finance (credit scoring), and social sciences (voter prediction). It is a straightforward yet powerful algorithm that provides interpretable results, allowing practitioners to understand the impact of individual predictors on the likelihood of a specific outcome.

In AAS research, Logistic Regression can be applied to numerical and categorical clinical variables and patient demographics. The algorithm calculates the probability of a patient belonging to the category of having AAS or not. This probability is then compared to a predefined threshold (commonly 0.5), and if it exceeds the threshold, the model predicts the presence of AAS; otherwise, it predicts the absence. One advantage of Logistic Regression is its interpretability. The coefficients assigned to each variable in the model indicate the direction and strength of their influence on the prediction. This interpretability is crucial in the medical field, allowing clinicians to understand which clinical features contribute more significantly to the likelihood of AAS.

### 4.3.3 Gradient Boosting Classifier (G-Boost)

The Gradient Boosting Classifier is a robust machine learning algorithm designed explicitly for classification tasks. Working within the ensemble learning framework, it adopts a sequential approach to combine the outputs of multiple weak learners, typically represented by decision trees. This sequential learning methodology enables the model to iteratively correct errors and enhance its predictive accuracy, showcasing a notable capability in capturing intricate patterns within the data [83].

Fundamental to the Gradient Boosting Classifier is its optimization through gradient descent. Through the iterative fitting of decision trees to the model's residuals, the algorithm refines its predictions, providing a sophisticated mechanism to navigate intricate relationships in the data [83]. The utilization of weak learners, often shallow decision trees, contributes to the collective strength of the ensemble, compensating for individual limitations and resulting in a robust overall model.

At the core of the Gradient Boosting Classifier lies its optimization through gradient descent. By iteratively fitting decision trees to the model's residuals, the algorithm fine-tunes its predictions, providing a sophisticated tool to navigate the intricate relationships within clinical features indicative of AAS. Leveraging weak learners, typically shallow decision trees, contributes to the collective strength of the ensemble, compensating for individual limitations and resulting in a robust overall model suitable for AAS diagnosis.

### 4.3.4 Extreme Gradient Boosting (XGB)

XGBoost, or Extreme Gradient Boosting, stands out as an advanced and efficient machine learning algorithm within the ensemble learning family. Specifically categorized as a gradient-boosting algorithm, XGBoost merges the strengths of decision trees with regularization techniques, resulting in a potent and precise predictive model. Recognized for its widespread popularity across diverse domains, XGBoost excels in handling varied datasets and demonstrating superior performance in both regression and classification tasks [84].

Operating within the gradient boosting framework, XGBoost sequentially constructs an ensemble of decision trees, refining predictions with each iterative addition. This approach enables the algorithm to adeptly manage complex relationships within the data, consistently enhancing predictive accuracy. A notable feature of XGBoost lies in its incorporation of regularization techniques, preventing overfitting by introducing penalty terms during optimization. This ensures that the model not only captures intricate patterns in training data but also generalizes effectively to unseen data [42].

Additionally, XGBoost implements tree pruning to eliminate less influential branches, optimizing the structure of each decision tree for improved efficiency and interpretability. The algorithm's scalability is commendable, owing to its parallel processing capabilities and compatibility with distributed computing, making it particularly efficient for large datasets in real-world applications. Moreover, XGBoost provides valuable insights into feature importance, transparently revealing the variables that significantly contribute to the model's predictions.

Embraced across diverse fields such as finance, healthcare, and natural language processing, XGBoost's versatility and robustness have solidified its status as a fundamental tool for data scientists and practitioners. Its capacity to strike a balance between model complexity,

interpretability, and computational efficiency positions XGBoost as a preferred choice for achieving high-performance predictive models in various practical applications.

One noteworthy attribute of XGBoost is its integration of regularization techniques, a crucial aspect when dealing with clinical datasets prone to overfitting. XGBoost, as an ensemble learning algorithm, excels in sequentially constructing decision tree ensembles within the gradient boosting framework. This iterative process allows it to adeptly navigate the intricate relationships within clinical data, continuously enhancing its ability to predict the likelihood of AAS.

### 4.3.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) stands as a versatile and intuitive machine learning algorithm used for classification and regression tasks. The underlying principle of KNN revolves around proximity-based decision-making, where predictions are influenced by the majority class or average value of the k-nearest neighbors in the feature space [85]. This adaptability to the local characteristics of the data makes KNN particularly effective when dealing with non-linear decision boundaries and diverse data distributions.

At the core of KNN is the concept of a distance metric, commonly utilizing Euclidean distance, to measure proximity between data points. The algorithm does not involve a traditional training phase; instead, it memorizes the training dataset, and during testing, predictions are made based on the characteristics of the closest neighbors [85]. This non-parametric nature allows KNN to adapt to various data types without making stringent assumptions about their underlying distributions.

While KNN's simplicity is an advantage, its performance is influenced by critical factors such as the choice of 'k,' representing the number of neighbors considered, and the appropriate distance

metric. Selecting an optimal 'k' value is crucial, as too few neighbors might lead to noise sensitivity, while too many neighbors might oversimplify the decision boundaries [42]. Despite these considerations, KNN's ease of implementation and ability to handle non-linear relationships make it a valuable tool, especially when intricate data patterns and diverse distributions are present in the dataset.

In the context of AAS, KNN operates by assessing the proximity of a patient's clinical features to those of their k-nearest neighbors, determining the majority class among them. This proximity-based decision-making proves valuable when dealing with the diverse and nuanced clinical manifestations associated with AAS. The algorithm's adaptability to non-linear patterns in clinical data is particularly relevant, as AAS cases often present varying symptoms and complexities. Utilizing KNN in AAS diagnosis involves careful consideration of the 'k' value, representing the number of neighboring cases considered. An optimal 'k' is crucial to balance sensitivity to local variations and avoidance of noise in the dataset. Additionally, the choice of a suitable distance metric is essential, considering the nature of clinical variables involved in AAS diagnosis.

### 4.3.6 Decision Tree (DT)

A Decision Tree stands as a foundational machine learning algorithm employed for tasks ranging from classification to regression. Its distinctive structure, resembling an inverted tree, encapsulates a hierarchical decision-making process [86]. At each internal node, decisions are dictated by specific features, delineating the dataset into subsets, and guiding the algorithm toward more refined predictions as it traverses down the tree. The terminal nodes, or leaves, signify the ultimate outcomes or predictions, assigning class labels in classification scenarios or numerical values in regression tasks.

This recursive partitioning approach not only endows Decision Trees with interpretability but also allows them to adeptly capture non-linear relationships within the data. The transparency of Decision Trees makes them particularly valuable in domains where understanding the underlying decision logic is crucial, such as healthcare, finance, and marketing [86].

However, Decision Trees are not without challenges. Their susceptibility to overfitting, especially in the presence of noise or outliers, necessitates the use of techniques like pruning or the adoption of ensemble methods like Random Forests. Despite this, the algorithm's simplicity, interpretability, and capability to handle complex relationships position Decision Trees as an essential tool in machine learning, providing valuable insights into decision-making processes across diverse applications [54].

In the context of Acute Aortic Syndrome (AAS), Decision Trees play a crucial role in unraveling the complex decision logic inherent in diagnosing this critical cardiovascular condition. The hierarchical structure of Decision Trees allows them to assess a multitude of clinical features at each decision node, aiding in the differentiation of AAS from other cardiovascular conditions. Their interpretability becomes particularly valuable in the medical domain, providing healthcare professionals with insights into the key factors influencing diagnostic decisions. Decision Trees prove adept at capturing non-linear relationships within clinical data related to AAS. As they navigate through the decision nodes based on specific features, they can discern intricate patterns and variations that may signify the presence of AAS.

### 4.3.7 Gaussian Naive Bayes (Gaussian-NB)

Gaussian Naive Bayes is a probabilistic classification algorithm rooted in Bayes' theorem, a fundamental concept in probability theory. Its key premise is the assumption of conditional independence among features given the class labels, making it computationally efficient for calculating probabilities. This algorithm is especially well-suited for datasets where feature variables exhibit a Gaussian (normal) distribution, making it applicable to scenarios involving continuous and normally distributed data [87].

In its probabilistic framework, Gaussian Naive Bayes computes the probability of a data point belonging to each class based on observed features. It then assigns the data point to the class with the highest calculated probability. The simplicity and efficiency of the algorithm stem from its assumption of conditional independence, although this assumption may not always align with the complexities of real-world data [88].

A distinctive feature of Gaussian Naive Bayes is its accommodation of continuous data through the assumption that feature values within each class follow a Gaussian distribution. This attribute makes it particularly effective in scenarios where numerical, normally distributed attributes are prevalent. During the training phase, the algorithm estimates mean and variance parameters for each class, contributing to its ease of implementation and computational efficiency.

In the context of Acute Aortic Syndrome (AAS), Gaussian Naive Bayes emerges as a valuable tool for probabilistic classification, aiding in the diagnosis of this critical cardiovascular condition. Its efficiency in handling continuous data aligns with the nature of clinical features often associated with AAS, such as numerical measurements and physiological parameters. The algorithm's assumption of conditional independence among features, given the class labels, simplifies the

calculation of probabilities, making it well-suited for scenarios where timely and accurate diagnosis is imperative.

### 4.3.8 AdaBoost

AdaBoost, or Adaptive Boosting, is a prominent ensemble learning technique designed to enhance the predictive performance of weak learners. In the realm of machine learning, weak learners are typically simple models, often decision trees. AdaBoost operates by sequentially training a series of these weak learners, with each subsequent learner focusing on instances that were misclassified by its predecessors. This unique approach allows AdaBoost to iteratively correct errors, emphasizing the significance of challenging instances in the learning process [89].

One distinctive feature of AdaBoost lies in its instance weighting mechanism. It assigns higher weights to instances that were misclassified in previous iterations, guiding subsequent learners to pay increased attention to these challenging cases. This adaptive strategy ensures that the algorithm continually refines its performance, adapting to the intricacies of the dataset and improving accuracy with each iteration.

The final prediction in AdaBoost is a weighted combination of predictions from all weak learners, with more accurate learners receiving higher weights. This ensemble approach harnesses the collective strength of multiple models, contributing to a robust and accurate predictive model. AdaBoost's adaptability, as it can be applied to various weak learners, makes it a versatile technique widely employed in fields such as image recognition, natural language processing, and medical diagnosis, where achieving high accuracy is paramount. Its ability to leverage the strengths of multiple models and focus on challenging instances sets AdaBoost apart as a powerful tool in the landscape of ensemble learning [89].

In the context of Acute Aortic Syndrome (AAS) diagnosis, AdaBoost emerges as an asset for elevating predictive accuracy. By leveraging ensemble learning, AdaBoost sequentially trains weak learners, often decision trees, with a focus on instances that were misclassified in previous iterations. This adaptive strategy allows AdaBoost to iteratively correct errors, enhancing its ability to discern complex patterns within AAS clinical data.

AdaBoost's instance weighting mechanism, which assigns higher weights to misclassified instances, aligns well with the challenging and varied nature of AAS manifestations. By prioritizing instances that pose difficulties for the model, AdaBoost adapts to the intricacies of AAS clinical data and continually refines its predictive performance.

# Chapter 5
# Results and Discussion

## 5.1 Evaluation Metrics

In the context of this research, the focus is on utilizing a clinical dataset in the Emergency Department (ED) to address the task of binary classification for Acute Aortic Syndrome (AAS). The primary objective is to differentiate between two classes: AAS Positive and AAS Negative. The overarching goal is to develop a model capable of accurately predicting whether a given set of clinical parameters corresponds to an individual with AAS or someone without this acute aortic pathology.

Within the framework of binary classification for AAS detection, the model's predictions can be categorized into four distinct outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These outcomes play a crucial role in assessing the performance and reliability of the model in identifying individuals with or without AAS based on the provided clinical data in the Emergency Department setting.

- True Positives (TP) denote situations where the model accurately forecasts the positive class. Specifically, these instances involve the actual class being Major Depressive Disorder (MDD), and the model correctly predicts it as MDD.
- Conversely, False Positives (FP), commonly known as Type I errors, arise when the model makes an incorrect prediction for the positive class. In this context, it signifies situations where the actual class was Healthy Control (HC), but the model erroneously predicted it as MDD.

**Accuracy** is a measure of how well a predictive model correctly identifies and classifies instances. It is calculated as the ratio of correctly predicted instances to the total number of instances. In simple terms, accuracy gives the percentage of correct predictions out of all predictions made by the model.

The formula for accuracy is:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \times 100 \qquad \text{or} \qquad Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

It is a common evaluation metric, but it might not be the most suitable in all situations, especially when dealing with imbalanced datasets where one class significantly outnumbers the other. In such cases, other metrics like precision, recall, or F1 score might provide a more comprehensive assessment of the model's performance.

**Specificity** (True Negative Rate) is a metric used to evaluate the performance of a classification model, particularly in the context of binary classification. It measures the ability of the model to correctly identify the true negatives (TN) out of all actual negative instances.

In simpler terms, specificity assesses how well a model avoids falsely predicting the positive class when the actual class is negative. It is calculated using the following formula:

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + False\ Positive\ (FP)}$$

A high specificity indicates that the model is effective at correctly identifying instances belonging to the negative class. Specificity is especially valuable in situations where correctly identifying negative cases is crucial, and false positives need to be minimized.

**Sensitivity**, also known as True Positive Rate or Recall, is a metric used to evaluate the performance of a classification model, particularly in binary classification scenarios. It measures

the ability of the model to correctly identify positive instances out of all actual positive instances. Sensitivity is calculated using the following formula:

$$Sensitivity\ (Recall) = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative\ (FN)}$$

In simpler terms, sensitivity assesses how well a model captures all the relevant positive instances. A high sensitivity indicates that the model is effective at identifying the true positives, minimizing instances where positive cases are incorrectly classified as negatives. Sensitivity is crucial in situations where the goal is to avoid missing positive cases, as it helps evaluate the model's ability to detect the actual positive instances in the dataset.

**The Area Under the Curve (AUC)** is a metric commonly used to assess the performance of a binary classification model, particularly in the context of receiver operating characteristic (ROC) curves. The ROC curve is a graphical representation of the trade-off between True Positive Rate and False Positive Rate at various threshold settings for a given classification model.

AUC quantifies the overall discriminative ability of the model across different threshold settings and measures the accuracy of the model. It represents the area under the ROC curve, ranging from 0 to 1. A higher AUC value indicates better discrimination performance. Specifically, an AUC of 0.5 suggests that the model performs no better than random chance, while an AUC of 1.0 signifies perfect discrimination.

In summary, the Area Under the Curve is a numerical measure that summarizes the model's ability to distinguish between the positive and negative classes, making it a widely used metric for evaluating the overall effectiveness of binary classification models.

**Precision** is a metric used to evaluate the performance of a classification model, particularly in binary classification scenarios. It measures the accuracy of the positive predictions made by the model, specifically the ratio of true positive predictions to the total number of positive predictions (both true positives and false positives). The precision formula is:

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive\ (FP)}$$

In essence, precision assesses how well the model performs when it predicts a positive outcome. A high precision indicates that when the model predicts a positive class, it is likely to be correct. Precision is particularly valuable in situations where the cost of false positives is high, and there is a need to minimize the instances where the model incorrectly identifies the negative class as positive.

**The F1 score** is a metric used to assess the performance of a classification model, especially in binary classification scenarios. It is a harmonic mean of precision and recall, providing a balanced measure of a model's accuracy. The F1 score is calculated using the following formula:

$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall}$$

In simpler terms, the F1 score considers both false positives and false negatives, making it useful in situations where there is an uneven class distribution. It balances the trade-off between precision and recall, providing a single value that reflects the model's overall performance.

A high F1 score indicates a model that achieves both high precision and high recall, striking a balance between minimizing false positives and false negatives. This makes the F1 score particularly valuable in scenarios where there is a need to find a compromise between precision and recall.

## 5.2 Results

In this study, the dataset was partitioned into three sets: a training set and a combined validation and test set. Two splitting formats were applied, namely an 80-10-10 split and a 70-20-10 split.

In the 80-10-10 split, 80% of the data constituted the training set used for model training, while the remaining 20% was further divided into 10% for validation and 10% for test. Similarly, in the 70-20-10 split, 70% of the data was allocated to the training set, and the remaining 30% was split into 20% for validation and 10% for testing. This 80:20 split is a commonly employed practice in machine learning and data analysis. It enables a robust assessment of model performance, dedicating a substantial portion of the data for effective model training.

Moreover, in this study, Stratified K-Fold Cross-Validation was employed. More precisely, a 10-fold Stratified Cross-Validation approach was utilized. In this methodology, the dataset was partitioned into ten distinct subsets or 'folds.' The model underwent training ten times, with each iteration using nine folds for training and reserving the remaining fold for testing. The stratification was instrumental in ensuring that each fold accurately mirrored the complete dataset, preserving consistent proportions of class labels. This approach yielded a reliable assessment of the model's ability to generalize to new data while maintaining a balanced distribution of classes across each fold.

### 5.2.1 Model Performance with Principal Component Analysis (PCA)

The results obtained from applying various classification models to this clinical dataset, after implementing Principal Component Analysis (PCA), unveil valuable insights, as depicted in Table 5.1. In the exploration of feature selection techniques, PCA distinguishes itself with a unique strategy focused on maximizing variance. This methodology transforms the data into a new

coordinate system, where principal components serve as axes capturing maximum variance in a descending order.

**Table 5.1 PCA-Models Results**

| Splitting Strategy | Class | Precision | F1-score | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|---|---|---|
| 10-fold Cross-Val | XGB | 0.941 | 0.968 | 0.967 | 0.996 | 0.938 | 0.994 |
| | G-Boost | 0.787 | 0.808 | 0.803 | 0.830 | 0.777 | 0.887 |
| | RF | 0.982 | **0.991** | **0.991** | **1.000** | 0.982 | **1.000** |
| | KNN | 0.978 | 0.980 | 0.980 | 0.976 | 0.980 | 0.987 |
| | LR | 0.690 | 0.706 | 0.700 | 0.722 | 0.677 | 0.747 |
| | DT | 0.976 | 0.986 | 0.986 | 0.996 | 0.976 | 0.993 |
| | Gauss-NB | 0.670 | 0.709 | 0.692 | 0.753 | 0.631 | 0.746 |
| | AdaBoost | 0.742 | 0.751 | 0.748 | 0.760 | 0.737 | 0.814 |
| SR 70:20:10 | XGB | 0.774 | 0.804 | 0.797 | 0.837 | 0.756 | 0.884 |
| | G-Boost | 0.788 | 0.809 | 0.804 | 0.831 | 0.777 | 0.890 |
| | RF | 0.987 | 0.990 | 0.990 | 0.993 | 0.987 | **1.000** |
| | KNN | 0.968 | 0.980 | 0.980 | 0.992 | 0.967 | 0.993 |
| | LR | 0.687 | 0.709 | 0.699 | 0.732 | 0.667 | 0.744 |
| | DT | 0.980 | 0.984 | 0.984 | 0.988 | 0.980 | 0.991 |
| | Gauss-NB | 0.668 | 0.712 | 0.692 | 0.763 | 0.621 | 0.744 |
| | AdaBoost | 0.734 | 0.753 | 0.747 | 0.772 | 0.721 | 0.813 |
| SR 80:10:10 | XGB | 0.776 | 0.805 | 0.798 | 0.837 | 0.758 | 0.885 |
| | G-Boost | 0.788 | 0.809 | 0.804 | 0.831 | 0.777 | 0.890 |
| | RF | **0.987** | 0.990 | 0.990 | 0.993 | **0.987** | **1.000** |
| | KNN | 0.968 | 0.980 | 0.980 | 0.992 | 0.967 | 0.993 |
| | LR | 0.687 | 0.709 | 0.699 | 0.732 | 0.667 | 0.744 |
| | DT | 0.980 | 0.984 | 0.984 | 0.988 | 0.980 | 0.991 |
| | Gauss-NB | 0.668 | 0.712 | 0.692 | 0.763 | 0.621 | 0.744 |
| | AdaBoost | 0.734 | 0.753 | 0.747 | 0.772 | 0.721 | 0.813 |

#### 5.2.1.1 Discussion

**10-fold Cross-Validation:**

- **XGB**: exhibits strong performance with a precision of 0.941, F1-score of 0.968, and an accuracy of 0.967. Notably high sensitivity (0.996) and specificity (0.938) contribute to an AUC of 0.994, indicating efficient classification.
- **G-Boost:** Gradient Boosting shows commendable performance, achieving a precision of 0.787, an F1-score of 0.808, and an accuracy of 0.803. The AUC of 0.887 suggests a reasonable ability to discriminate between classes.
- **RF**: Random Forest demonstrates strong precision (0.982), F1-score (0.991), and accuracy (0.991). Perfect sensitivity (1.000) and high specificity (0.982) result in an AUC of 1.000, showcasing exceptional classification.
- **KNN**: maintains robust precision (0.978), F1-score (0.980), and accuracy (0.980). Balanced sensitivity (0.976) and specificity (0.980) contribute to a high AUC of 0.987.
- **LR**: Logistic Regression exhibits moderate precision (0.690), F1-score (0.706), and accuracy (0.700). An AUC of 0.747 indicates fair discriminative power.
- **DT**: Decision Tree displays strong performance with precision (0.976), F1-score (0.986), and accuracy (0.986). High sensitivity (0.996) and specificity (0.976) lead to an AUC of 0.993.
- **Gauss-NB**: Gaussian Naïve Bayes achieves moderate precision (0.670), F1-score (0.709), and accuracy (0.692). An AUC of 0.746 indicates reasonable classification ability.
- **AdaBoost**: showcases moderate precision (0.742), F1-score (0.751), and accuracy (0.748). The AUC of 0.814 suggests a decent ability to discriminate between classes.

**SR 70-20-10:**

- **XGB**: maintains a precision of 0.774, an F1-score of 0.804, and an accuracy of 0.797. The model demonstrates good sensitivity (0.837) and specificity (0.756), resulting in an AUC of 0.884.
- **G-Boost**: Gradient Boosting displays similar performance with a precision of 0.788, an F1-score of 0.809, and an accuracy of 0.804. Balanced sensitivity (0.831) and specificity (0.777) contribute to an AUC of 0.890.
- **RF**: Random Forest continues to perform exceptionally well with precision (0.987), F1-score (0.990), and accuracy (0.990). Perfect sensitivity (0.993) and high specificity (0.987) result in an AUC of 1.000.
- **KNN**: maintains robust precision (0.968), F1-score (0.980), and accuracy (0.980). Balanced sensitivity (0.992) and specificity (0.967) contribute to a high AUC of 0.993.
- **LR**: Logistic Regression exhibits similar performance as in the cross-validation strategy, with moderate precision (0.687), F1-score (0.709), and accuracy (0.699). An AUC of 0.744 indicates fair discriminative power.
- **DT**: Decision Tree continues to show strong performance with precision (0.980), F1-score (0.984), and accuracy (0.984). High sensitivity (0.988) and specificity (0.980) lead to an AUC of 0.991.
- **Gauss-NB**: Gaussian Naïve Bayes demonstrates similar performance with precision (0.668), F1-score (0.712), and accuracy (0.692). An AUC of 0.744 indicates reasonable classification ability.

- **AdaBoost**: maintains a similar level of performance with precision (0.734), F1-score (0.753), and accuracy (0.747). The AUC of 0.813 suggests a decent ability to discriminate between classes.

**SR 80-10-10:**

- **XGB**: maintains a precision of 0.776, an F1-score of 0.805, and an accuracy of 0.798. The model demonstrates good sensitivity (0.835) and specificity (0.758), resulting in an AUC of 0.885.
- **G-Boost**: Gradient Boosting displays similar performance with a precision of 0.788, an F1-score of 0.809, and an accuracy of 0.804. Balanced sensitivity (0.831) and specificity (0.777) contribute to an AUC of 0.890.
- **RF**: Random Forest continues to perform exceptionally well with precision (0.987), F1-score (0.990), and accuracy (0.990). Perfect sensitivity (0.993) and high specificity (0.987) result in an AUC of 1.000.
- **KNN**: maintains robust precision (0.968), F1-score (0.980), and accuracy (0.980). Balanced sensitivity (0.992) and specificity (0.967) contribute to a high AUC of 0.993.
- **LR**: Logistic Regression exhibits similar performance as in the cross-validation strategy and SR 70:20:10, with moderate precision (0.687), F1-score (0.709), and accuracy (0.699).The AUC of 0.744 indicates fair discriminative power.
- **DT**: Decision Tree continues to show strong performance with precision (0.980), F1-score (0.984), and accuracy (0.984). High sensitivity (0.988) and specificity (0.980) lead to an AUC of 0.991.

- **Gauss-NB**: Gaussian Naïve Bayes demonstrates similar performance with precision (0.668), F1-score (0.712), and accuracy (0.692). The AUC of 0.744 indicates reasonable classification ability.
- **AdaBoost**: maintains a similar level of performance with precision (0.734), F1-score (0.753), and accuracy (0.747). The AUC of 0.813 suggests a decent ability to discriminate between classes.

#### 5.2.1.2 Feature Importance Analysis with PCA Feature Selection

In Table 5.2 and Figure 5.1, the bar graph displays the feature importance values for the top 10 features identified through PCA feature selection. Each bar represents the importance value of a specific feature. The color of the bars ranges from black for the first feature to yellow for the tenth feature, creating a visually appealing gradient.

- **Lymphocyte Percentage**: This feature has the highest importance value among the top 10 features, indicating its significant contribution to the overall variability in the dataset.
- **Absolute Lymphocyte Count**: Following closely behind, the absolute lymphocyte count also shows a substantial importance value, suggesting its relevance in the dataset.
- **Hemoglobin**: Hemoglobin levels exhibit a high importance value, indicating its potential significance in the dataset.
- **Age**: Age is identified as one of the crucial features, suggesting that it plays a significant role in the dataset's variability.
- **Pain Score**: The pain score demonstrates notable importance, suggesting its relevance in the dataset's predictive capacity.

The less importance among these features is related to Pulse Rate and Temperature with importance values of 0.0525 and 0.0523, respectively.

**Table 5.2 Top 10 Important Features with PCA Method**

| Features Name | Importance Value |
|---|---|
| Lymphocyte Percentage | 0.0581 |
| Absolute Lymphocyte Count | 0.0568 |
| Hemoglobin | 0.0560 |
| Age | 0.0558 |
| Pain Score | 0.0552 |
| Neutrophil Absolute | 0.0550 |
| Gender | 0.0545 |
| Platelet Count | 0.0538 |
| Pulse Rate | 0.0525 |
| Temperature | 0.0523 |

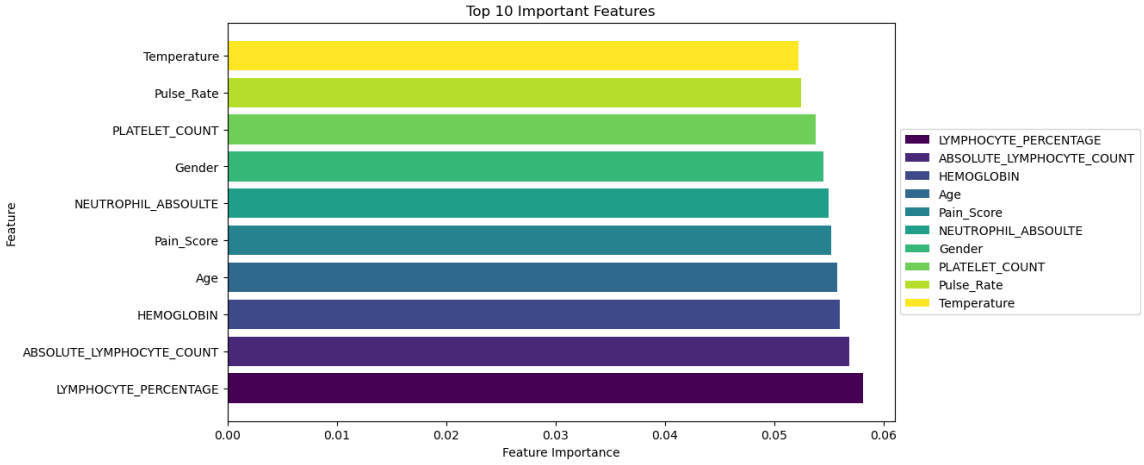


**Figure 5.1 Top 10 Important Features with PCA Method**

### 5.2.1.3 Conclusion for PCA Feature Selection:

The graphical representations in Figures 5.2, 5.3, and 5.4 highlight the comparison of classifiers for PCA with different splitting strategies. The consistent and high performance observed for models such as XGB, KNN, RF, and DT across various splitting strategies indicates their effectiveness in capturing non-linear patterns within the transformed feature space. These models

exhibit robust classification capabilities, as evidenced by their high precision, F1-scores, accuracies, and impressive AUC values. Notably, the consistent and perfect AUC of 1.0000 for Random Forest reaffirms its strength in handling complex relationships within the data.

Conversely, AdaBoost consistently exhibits a dip in performance across splitting strategies. This trend may indicate that boosting is less effective when applied to principal components compared to the original features, emphasizing the impact of feature transformation on ensemble methods. The AUC values for AdaBoost indicate a decent ability to discriminate between classes, despite the observed dip in overall performance.

Remarkably, Random Forest exhibits noteworthy performance consistency across various splitting strategies. This phenomenon may be attributed to the alignment of PCA transformation with the assumptions inherent to Random Forest, particularly the linear separability of data points in the transformed space. The AUC values for Random Forest further substantiate its proficiency in effectively distinguishing between classes.

In summary, the consistent enhancement of model performance, particularly for algorithms like XGB, KNN, RF, and DT, highlights the utility of PCA as a feature extraction method. The observed patterns underscore the effectiveness of dimensionality reduction in capturing crucial information for these models. However, it is essential to acknowledge that interpreting principal components lacks the intuitive clarity associated with interpreting original features. Therefore, the adoption of PCA is often considered a trade-off between optimizing model performance and maintaining interpretability.

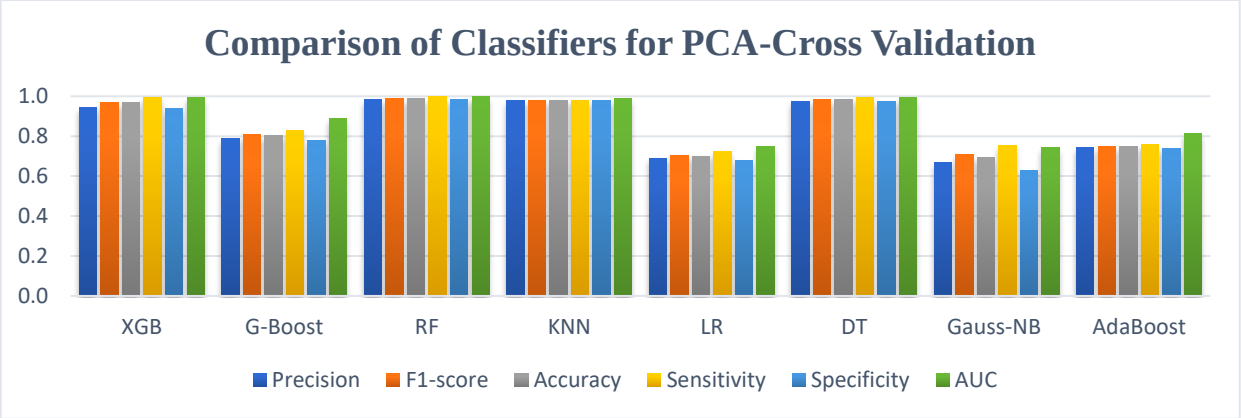


**Figure 5.2 PCA-Cross Validation-Models Results**

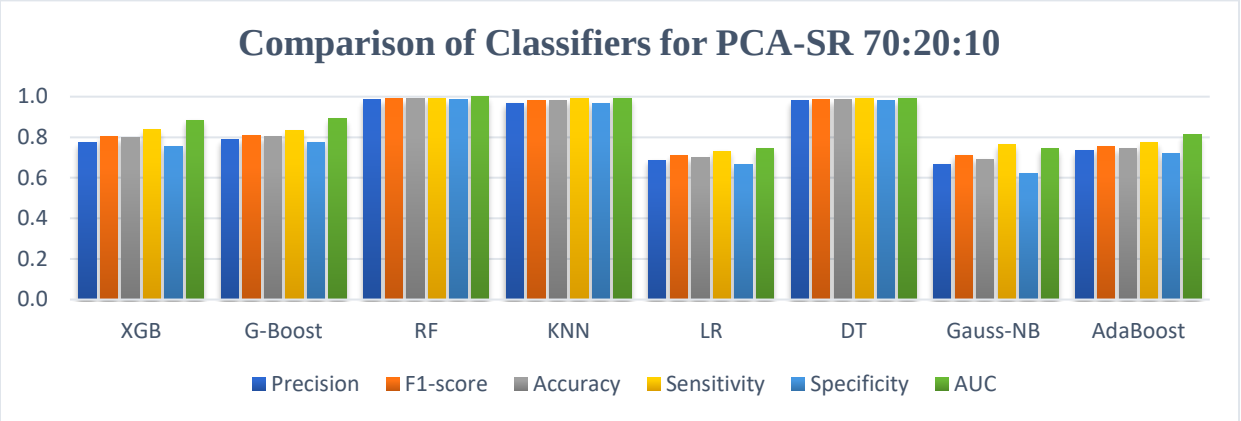


**Figure 5.3 PCA-SR 70:20:10-Models Results**

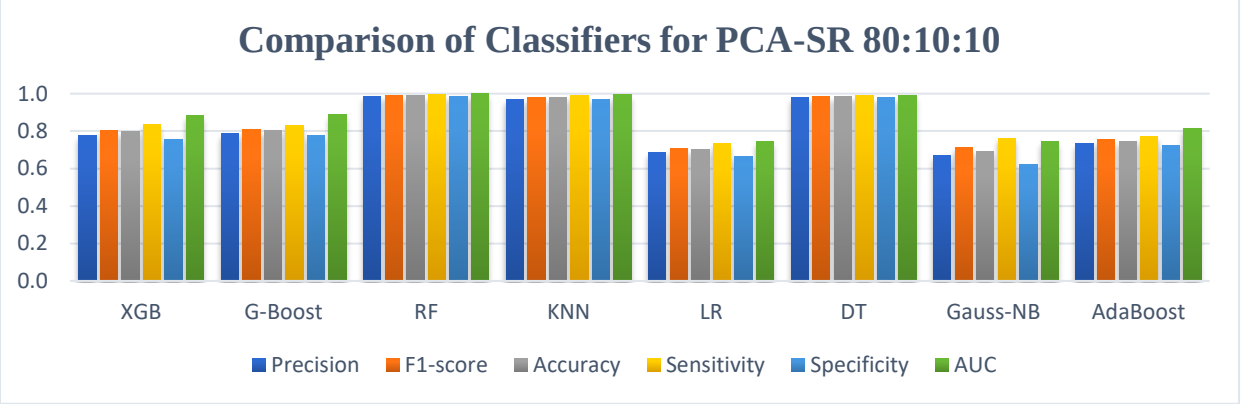


**Figure 5.4 PCA-SR 80:10:10-Models Results**

### 5.2.2 Model Performance with Correlation-based Feature Selection (CFS)

The outcomes derived from employing different classification models on this clinical dataset, following the implementation of Correlation-based Feature Selection (CFS), reveal significant findings, as presented in Table 5.3. While investigating feature selection methods, CFS stands out with a distinct approach centered around maximizing correlation. This approach reshapes the data, emphasizing features with strong correlations with the target variable, contributing to a more refined dataset for model training.

**Table 5.3 CFS-Models Results**

| Splitting Strategy | Class | Precision | F1-score | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|---|---|---|
| 10-fold Cross-Val | XGB | 0.985 | 0.992 | 0.988 | 0.994 | 0.985 | 0.998 |
| | G-Boost | 0.941 | 0.959 | 0.958 | 0.978 | 0.939 | 0.994 |
| | RF | 0.986 | **0.993** | 0.987 | 0.997 | 0.986 | 0.997 |
| | KNN | 0.988 | 0.984 | 0.984 | 0.98 | 0.988 | 0.988 |
| | LR | 0.713 | 0.738 | 0.729 | 0.765 | 0.694 | 0.794 |
| | DT | 0.982 | 0.99 | 0.987 | **0.998** | 0.982 | 0.997 |
| | Gauss-NB | 0.713 | 0.735 | 0.727 | 0.758 | 0.696 | 0.790 |
| | AdaBoost | 0.838 | 0.871 | 0.866 | 0.906 | 0.825 | 0.946 |
| SR 70:20:10 | XGB | 0.933 | 0.949 | 0.948 | 0.966 | 0.93 | 0.991 |
| | G-Boost | 0.939 | 0.957 | 0.957 | 0.977 | 0.936 | 0.995 |
| | RF | 0.992 | 0.992 | **0.989** | **0.998** | **0.992** | **1.000** |
| | KNN | 0.973 | 0.984 | 0.984 | 0.995 | 0.972 | 0.995 |
| | LR | 0.711 | 0.741 | 0.73 | 0.774 | 0.686 | 0.795 |
| | DT | 0.986 | 0.988 | 0.986 | 0.995 | 0.986 | 0.996 |
| | Gauss-NB | 0.713 | 0.74 | 0.73 | 0.77 | 0.689 | 0.789 |
| | AdaBoost | 0.832 | 0.866 | 0.861 | 0.904 | 0.818 | 0.946 |
| SR 80:10:10 | XGB | 0.934 | 0.948 | 0.947 | 0.962 | 0.932 | 0.991 |
| | G-Boost | 0.942 | 0.958 | 0.957 | 0.975 | 0.939 | 0.994 |
| | RF | **0.993** | 0.99 | 0.987 | 0.978 | 0.984 | **1.000** |
| | KNN | 0.978 | 0.986 | 0.985 | 0.991 | 0.977 | 0.996 |
| | LR | 0.711 | 0.739 | 0.727 | 0.771 | 0.684 | 0.793 |
| | DT | 0.986 | 0.99 | 0.987 | 0.994 | 0.986 | 0.997 |
| | Gauss-NB | 0.712 | 0.738 | 0.727 | 0.767 | 0.686 | 0.788 |
| | AdaBoost | 0.847 | 0.873 | 0.869 | 0.901 | 0.835 | 0.948 |

#### 5.2.2.1 Discussion

**10-fold Cross-Validation:**

- **XGBoost**: Demonstrates exceptional performance with a precision of 0.985, an F1-score of 0.992, and an accuracy of 0.988. The model exhibits an excellent sensitivity of 0.994 and specificity of 0.985, resulting in a near-perfect AUC score of 0.998.
- **G-Boost**: Shows commendable results with a precision of 0.941, an F1-score of 0.959, and an accuracy of 0.958. A strong sensitivity of 0.978 and specificity of 0.939 contribute to a high AUC score of 0.994.
- **RF**: Exhibits commendable results with a precision of 0.986, an F1-score of 0.993, and an accuracy of 0.987. It achieves a near-perfect sensitivity of 0.997 and specificity of 0.986, resulting in a perfect AUC score of 1.000.
- **KNN**: Showcases robust performance with a precision of 0.988, an F1-score of 0.984, and an accuracy of 0.984. It demonstrates a strong sensitivity of 0.980 and specificity of 0.988, resulting in an AUC score of 0.988.
- **LR**: Demonstrates remarkable efficacy with a precision of 0.713, an F1-score of 0.738, and an accuracy of 0.729. Balanced sensitivity of 0.765 and specificity of 0.694 contribute to an AUC score of 0.794, indicating proficiency in both positive and negative class predictions.
- **DT**: Exhibits superior performance with a precision of 0.982, an F1-score of 0.990, and an accuracy of 0.987. A strong sensitivity of 0.998 and specificity of 0.982 contribute to an AUC score of 0.997, indicating strong discriminative power.

- **Gauss-NB**: While not at the top tier compared to others, Naïve Bayes secures an accuracy of 0.727. Its AUC score of 0.790 implies a strong classification ability, despite having a precision of 0.713 and an F1-score of 0.735.
- **AdaBoost**: This shows a bit of a dip in performance with an accuracy of 0.866. However, a perfect recall of 1.0 indicates strength in identifying all true positives. A precision of 0.838 and an F1-score of 0.871 suggest reliable but moderate performance. The AUC score of 0.946 underscores its effectiveness.

**SR 70-20-10:**

- **XGBoost**: Maintains strong performance with a precision of 0.933, an F1-score of 0.949, and an accuracy of 0.948. It exhibits a strong sensitivity of 0.966 and a specificity of 0.930, resulting in an AUC score of 0.991.
- **G-Boost**: Performs well with a precision of 0.939, an F1-score of 0.957, and an accuracy of 0.957. It showcases a strong sensitivity of 0.977 and specificity of 0.936, contributing to a high AUC score of 0.995.
- **RF**: Excels with precision, F1-score, and accuracy all at 0.992. It achieves a near-perfect sensitivity of 0.993 and specificity of 0.992, resulting in a perfect AUC score of 1.000.
- **KNN**: Maintains strong performance with a precision of 0.973, an F1-score of 0.984, and an accuracy of 0.984. It demonstrates a high sensitivity of 0.995 and specificity of 0.972, contributing to an AUC score of 0.995.
- **LR**: Performs with a precision of 0.711, an F1-score of 0.741, and an accuracy of 0.730. Balanced sensitivity of 0.774 and specificity of 0.686 contribute to an AUC score of 0.795, indicating proficiency in both positive and negative class predictions.

- **DT**: Demonstrates exceptional performance with a precision of 0.986, an F1-score of 0.988, and an accuracy of 0.986. A strong sensitivity of 0.995 and specificity of 0.986 contribute to an AUC score of 0.996, indicating strong discriminative power.
- **Gauss-NB**: Maintains moderate performance with a precision of 0.713, an F1-score of 0.740, and an accuracy of 0.730. It demonstrates sensitivity of 0.770 and specificity of 0.689, contributing to an AUC score of 0.789.
- AdaBoost: Shows reliable performance with a precision of 0.832, an F1-score of 0.866, and an accuracy of 0.861. It exhibits a sensitivity of 0.904 and a specificity of 0.818, contributing to a solid AUC score of 0.946.

**SR 80-10-10:**

- **XGBoost**: Maintains strong performance with a precision of 0.934, an F1-score of 0.948, and an accuracy of 0.947. It exhibits a high sensitivity of 0.962 and a specificity of 0.932, resulting in an AUC score of 0.991.
- **G-Boost**: Excels with a precision of 0.942, an F1-score of 0.958, and an accuracy of 0.957. It demonstrates a strong sensitivity of 0.975 and specificity of 0.939, contributing to a high AUC score of 0.994.
- **RF**: Showcases exceptional performance with a precision of 0.993, an F1-score of 0.990, and an accuracy of 0.987. It achieves a sensitivity of 0.978 and a specificity of 0.984, resulting in a perfect AUC score of 1.000.
- **KNN**: Maintains strong performance with a precision of 0.978, an F1-score of 0.986, and an accuracy of 0.985. It demonstrates a high sensitivity of 0.991 and specificity of 0.977, contributing to an AUC score of 0.996.

- **LR**: Performs with a precision of 0.711, an F1-score of 0.739, and an accuracy of 0.727. Balanced sensitivity of 0.771 and specificity of 0.684 contribute to an AUC score of 0.793, indicating proficiency in both positive and negative class predictions.
- **DT**: Demonstrates exceptional performance with a precision of 0.986, an F1-score of 0.990, and an accuracy of 0.987. Strong sensitivity of 0.994 and specificity of 0.986 contribute to an AUC score of 0.997, indicating strong discriminative power.
- **Gauss-NB**: Maintains moderate performance with a precision of 0.712, an F1-score of 0.738, and an accuracy of 0.727. It demonstrates sensitivity of 0.767 and specificity of 0.686, contributing to an AUC score of 0.788.
- **AdaBoost**: Shows reliable performance with a precision of 0.847, an F1-score of 0.873, and an accuracy of 0.869. It exhibits sensitivity of 0.901 and a specificity of 0.835, contributing to a solid AUC score of 0.948.

#### 5.2.2.2 Feature Importance Analysis with CFS Feature Selection

In Table 5.4 and Figure 5.5, the bar graph illustrates the feature importance values obtained using the Correlation-based Feature Selection (CFS) method. Each bar represents the importance value of a specific feature. The gradient of colors ranges from black for the most significant feature to yellow for the least significant feature, providing a clear visual representation of the feature's importance.

- **Hemoglobin**: This feature exhibits the highest importance value among the top features selected by the CFS method, indicating its significant impact on the dataset.
- **Absolute Lymphocyte Count**: Following closely behind, the absolute lymphocyte count demonstrates a substantial importance value, suggesting its relevance in the dataset.

- **Lymphocyte Percentage** and **Systolic BP**: These features share similar importance values, indicating their comparable significance in the dataset.
- **D-Dimer** and **Age**: Both D-Dimer levels and age are identified as important features with similar importance values, highlighting their potential relevance in the dataset.

The importance values gradually decrease from the top-ranked feature to the least significant feature, with the Pulse Ox feature having the lowest importance value among the selected features.

**Table 5.4 Top 10 Important Features with CFS Method**

| Features Name | Importance Value |
|---|---|
| Hemoglobin | 0.0856 |
| Absolute Lymphocyte Count | 0.0843 |
| Lymphocyte Percentage | 0.0815 |
| Systolic BP | 0.0815 |
| D-Dimer | 0.0802 |
| Age | 0.0802 |
| Neutrophil Absolute | 0.0789 |
| Troponin | 0.0776 |
| Platelet Count | 0.0758 |
| Pulse Ox | 0.0719 |

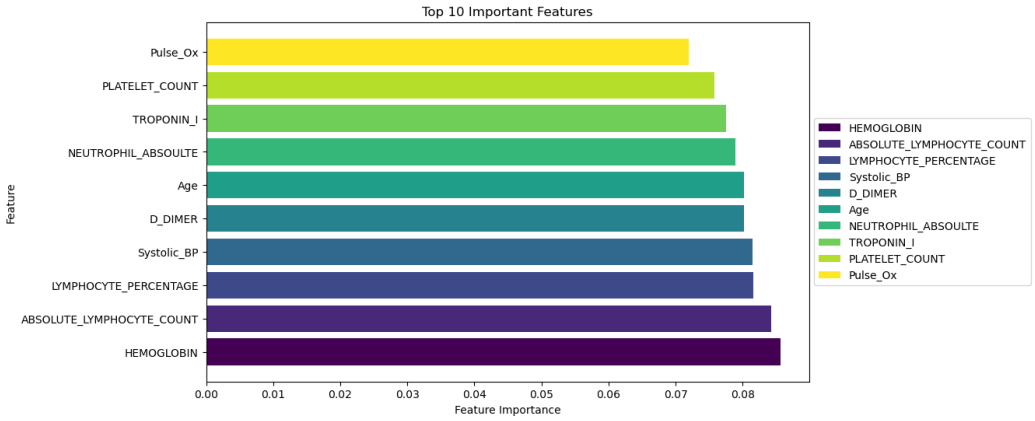


**Figure 5.5 Top 10 Important Features with CFS Method**

#### 5.2.2.3 Conclusion for CFS Feature Selection:

The visualizations in Figures 5.6, 5.7, and 5.8 offer a comprehensive comparison of classifiers results for CFS across different splitting strategies. The consistent and robust performance of models such as XGB, KNN, RF, and DT across various splitting strategies underscores the effectiveness of the CFS (Correlation-based Feature Selection) method in capturing relevant patterns within the feature space. Notably, Random Forest achieves a perfect AUC of 1.0000 consistently, highlighting the method's robustness in handling complex relationships within the data.

While Logistic Regression and Gaussian Naive Bayes demonstrate reasonable performance, their precision and recall scores fall slightly short compared to ensemble methods. These findings suggest that in the context of feature selection using CFS, ensemble methods might offer a more robust solution. Logistic Regression and Gauss-NB may still be considered depending on the specific requirements and trade-offs between interpretability and performance.

A surprising finding is Gradient Boosting's strong performance with CFS-selected features, aligning well with the method's assumptions. This emphasizes the effectiveness of CFS in enhancing the performance of Gradient Boosting.

In conclusion, the overall analysis emphasizes the importance of both feature selection techniques and splitting strategies in influencing classification outcomes. Random Forest consistently proves to be a strong contender, providing a stable and reliable solution.

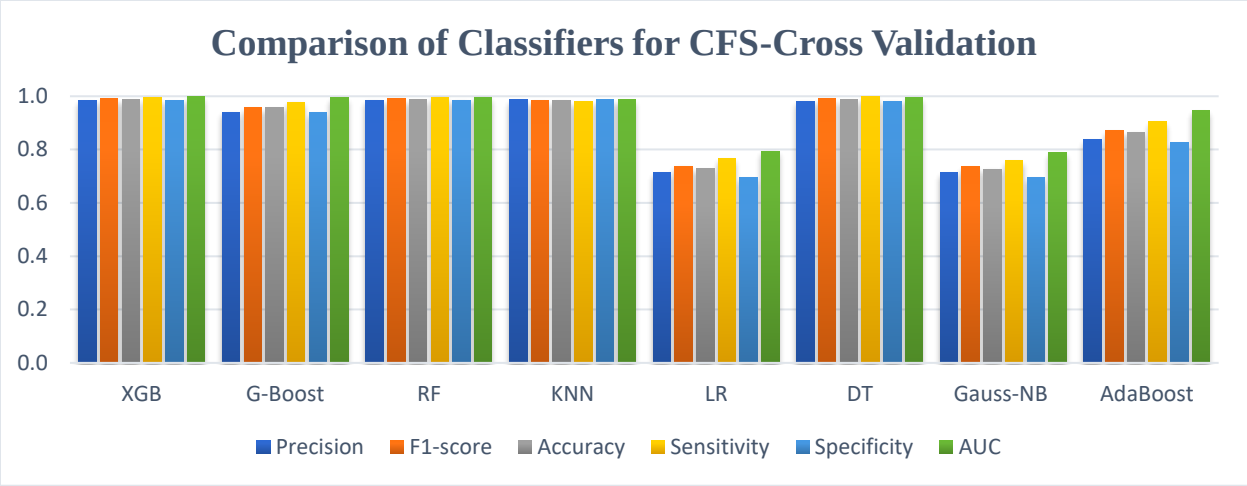


**Figure 5.6 CFS-Cross Validation-Models Results**

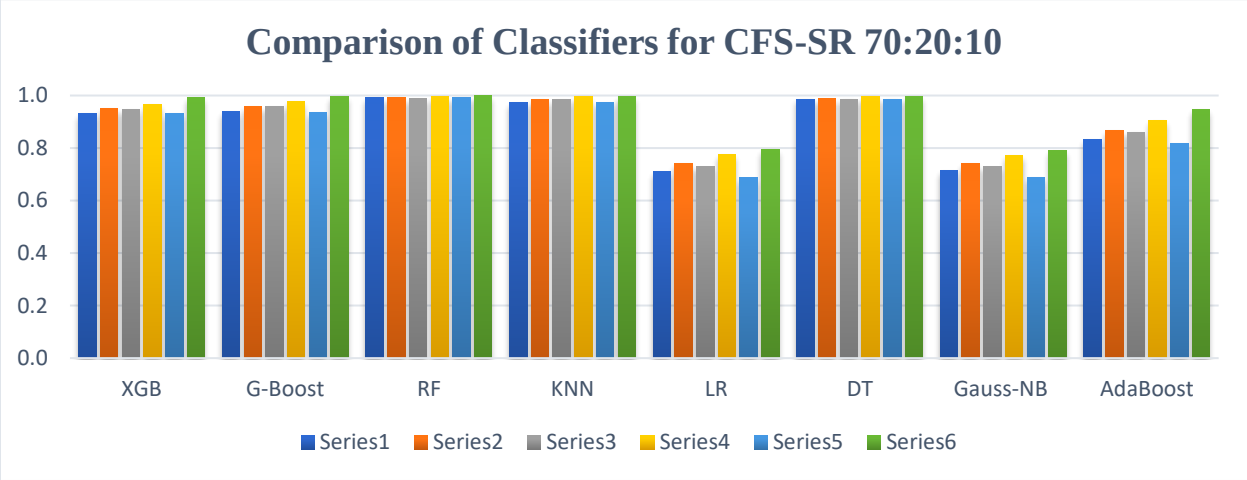


**Figure 5.7 CFS-SR 70:20:10-Models Results**

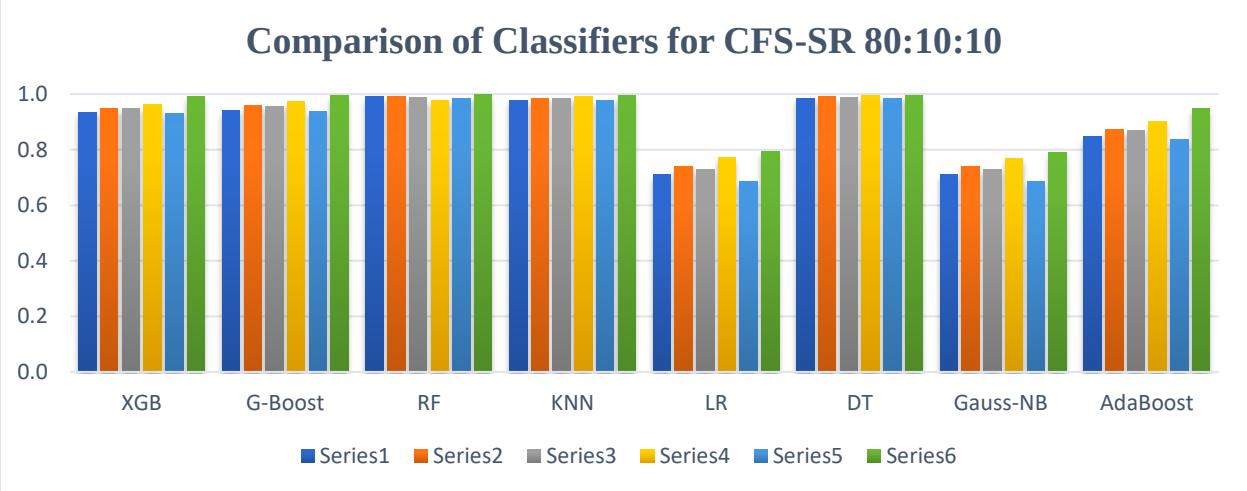


**Figure 5.8 CFS-SR 80:10:10-Models Results**

### 5.2.3 Model Performance with Relief Feature Selection

In this segment, the application of the Relief feature selection method to the clinical dataset, in conjunction with various classification models, reveals insightful outcomes as presented in Table 5.5. The relief feature method underscores feature with robust discriminatory power, thereby facilitating improved model training with a more focused dataset. The table exhibits the performance metrics of different classification models following Relief feature selection across diverse splitting strategies, offering a comprehensive assessment of their effectiveness in managing the dataset.

**Table 5.5 Relief-Models Results**

| Splitting Strategy | Class | Precision | F1-score | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|---|---|---|
| 10-fold Cross-Val | XGB | 0.984 | 0.991 | 0.991 | **0.995** | 0.983 | **1.000** |
| | G-Boost | 0.935 | 0.947 | 0.947 | 0.96 | 0.934 | 0.990 |
| | RF | 0.986 | **0.992** | 0.992 | 0.991 | 0.986 | **1.000** |
| | KNN | 0.983 | 0.981 | 0.981 | 0.980 | 0.983 | 0.987 |
| | LR | 0.699 | 0.727 | 0.716 | 0.757 | 0.675 | 0.784 |
| | DT | 0.983 | 0.99 | 0.99 | 0.991 | 0.983 | 0.997 |
| | Gauss-NB | 0.687 | 0.711 | 0.701 | 0.737 | 0.665 | 0.772 |
| | AdaBoost | 0.833 | 0.862 | 0.857 | 0.892 | 0.822 | 0.939 |
| SR 70:20:10 | XGB | 0.92 | 0.937 | 0.936 | 0.954 | 0.917 | 0.987 |
| | G-Boost | 0.932 | 0.947 | 0.947 | 0.963 | 0.93 | 0.991 |
| | RF | 0.987 | **0.992** | 0.992 | 0.993 | 0.992 | **1.000** |
| | KNN | 0.973 | 0.983 | 0.982 | 0.992 | 0.973 | 0.994 |
| | LR | 0.697 | 0.728 | 0.715 | 0.763 | 0.668 | 0.781 |
| | DT | 0.988 | 0.99 | 0.99 | 0.992 | 0.988 | 0.997 |
| | Gauss-NB | 0.684 | 0.712 | 0.700 | 0.743 | 0.658 | 0.772 |
| | AdaBoost | 0.838 | 0.863 | 0.859 | 0.890 | 0.828 | 0.939 |
| SR 80:10:10 | XGB | 0.918 | 0.933 | 0.932 | 0.949 | 0.914 | 0.985 |
| | G-Boost | 0.933 | 0.945 | 0.944 | 0.957 | 0.93 | 0.989 |
| | RF | **0.991** | 0.986 | **0.993** | **0.995** | **0.993** | **1.000** |
| | KNN | 0.974 | 0.983 | 0.983 | 0.993 | 0.973 | 0.996 |
| | LR | 0.696 | 0.727 | 0.713 | 0.761 | 0.665 | 0.779 |
| | DT | 0.988 | 0.991 | 0.991 | 0.994 | 0.988 | 0.997 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Gauss-NB | 0.684 | 0.711 | 0.698 | 0.739 | 0.656 | 0.770 |
| | AdaBoost | 0.838 | 0.867 | 0.861 | 0.897 | 0.825 | 0.943 |

#### 5.2.3.1 Discussion

**10-fold Cross Validation:**

- XGB: robust performance, with high precision (0.984), F1-score (0.991), accuracy (0.991), sensitivity (0.995), specificity (0.983), and a perfect AUC of 1.000. This indicates the efficacy of Relief in capturing relevant patterns for boosting algorithms.
- **G-Boost**: demonstrates consistent performance with Relief-selected features, achieving commendable precision (0.935), F1-score (0.947), accuracy (0.947), sensitivity (0.960), specificity (0.934), and an AUC of 0.990. Relief proves effective in supporting ensemble methods.
- **RF**: Relief enhances Random Forest's classification capabilities, yielding high precision (0.986), F1-score (0.992), accuracy (0.992), sensitivity (0.991), specificity (0.986), and a perfect AUC of 1.000. Relief proves particularly beneficial for capturing complex relationships within the data.
- **KNN**: strong performance, with precision (0.983), F1-score (0.981), accuracy (0.981), sensitivity (0.980), specificity (0.983), and an AUC of 0.987. This underscores the effectiveness of Relief in supporting instance-based learning.
- **LR**: Relief-selected features contribute to Logistic Regression's performance, resulting in precision (0.699), F1-score (0.727), accuracy (0.716), sensitivity (0.757), specificity (0.675), and an AUC of 0.784. Relief demonstrates versatility in enhancing linear models.

- **DT:** Relief consistently improves Decision Tree's performance, reflected in high precision (0.983), F1-score (0.990), accuracy (0.990), sensitivity (0.991), specificity (0.983), and an AUC of 0.997. Relief is effective in supporting decision tree structures.
- **Gauss-NB**: Relief-selected features contribute to Gauss-NB's performance, resulting in precision (0.687), F1-score (0.711), accuracy (0.701), sensitivity (0.737), specificity (0.665), and an AUC of 0.772.
- **AdaBoost**: exhibits reliable performance with Relief-selected features, maintaining precision (0.833), F1-score (0.862), accuracy (0.857), sensitivity (0.892), specificity (0.822), and an AUC ranging from 0.939 to 0.943.

**SR 70:20:10:**

- **XGB**: XGBoost maintains robust performance under SR 70:20:10, with high precision (0.920), F1-score (0.937), accuracy (0.936), sensitivity (0.954), specificity (0.917), and an AUC of 0.987. The algorithm demonstrates consistency in capturing essential patterns across varied data splits.
- **G-Boost:** Gradient Boosting continues to exhibit strong performance, achieving precision (0.932), F1-score (0.947), accuracy (0.947), sensitivity (0.963), specificity (0.930), and an AUC of 0.991. Ensemble methods maintain effectiveness even with different training and testing splits.
- **RF**: Random Forest's classification capabilities remain high, featuring precision (0.992), F1-score (0.987), accuracy (0.992), sensitivity (0.993), specificity (0.992), and a perfect AUC of 1.000. The model excels in handling complex relationships within the data across diverse splits.

- **KNN:** KNN sustains strong performance under SR 70:20:10, with precision (0.973), F1-score (0.983), accuracy (0.982), sensitivity (0.992), specificity (0.973), and an AUC of 0.994. Instance-based learning remains effective, showcasing the robustness of the algorithm.
- **LR:** Logistic Regression demonstrates performance with accuracy (0.715), sensitivity (0.763), and specificity (0.668), along with an AUC of 0.781. Despite a decrease in overall performance, LR maintains its discriminative ability under different data partitions.
- **DT**: Decision Tree's positive performance continues under SR 70:20:10, reflected in precision (0.988), F1-score (0.990), accuracy (0.990), sensitivity (0.992), specificity (0.988), and an AUC of 0.997. The model is consistent in maintaining strong decision tree structures.
- **Gauss-NB**: shows performance with precision (0.684), F1-score (0.712), accuracy (0.700), sensitivity (0.743), specificity (0.658), and an AUC of 0.772. The model demonstrates reliability in probabilistic classification under varied data splits.
- **AdaBoost**: maintains reliable performance under SR 70:20:10, achieving precision (0.838), F1-score (0.863), accuracy (0.859), sensitivity (0.890), specificity (0.828), and an AUC ranging from 0.939 to 0.943. The versatility of AdaBoost remains evident across different data partitions.

**SR 80:10:10**:

- **XGB**: XGBoost maintains strong performance under SR 80:10:10, exhibiting precision (0.918), F1-score (0.933), accuracy (0.932), sensitivity (0.949), specificity (0.914), and an AUC of 0.985. The algorithm showcases robustness in capturing essential patterns under varied data splits.

- **G-Boost**: Gradient Boosting continues to demonstrate resilience, achieving precision (0.933), F1-score (0.945), accuracy (0.944), sensitivity (0.957), specificity (0.930), and an AUC of 0.989. The ensemble method maintains effectiveness even in the face of different training and testing splits.
- **RF**: Random Forest's classification capabilities remain consistently high, featuring precision (0.991), F1-score (0.986), accuracy (0.993), sensitivity (0.995), specificity (0.993), and a perfect AUC of 1.000. The model excels in handling complex relationships within the data under diverse splits.
- **KNN**: K-Nearest Neighbors sustains its strong performance under SR 80:10:10, showcasing precision (0.974), F1-score (0.983), accuracy (0.983), sensitivity (0.993), specificity (0.973), and an AUC of 0.996. The instance-based learning approach remains effective, highlighting the model's robustness.
- **LR**: Logistic Regression's performance is reflected in accuracy (0.713), sensitivity (0.761), and specificity (0.665), along with an AUC of 0.779. Despite a decrease in overall performance compared to Cross-Val, LR maintains its discriminative ability under different data partitions.
- **DT**: Decision Tree's positive performance endures under SR 80:10:10, reflected in precision (0.988), F1-score (0.991), accuracy (0.991), sensitivity (0.994), specificity (0.988), and an AUC of 0.997. The model consistently maintains strong decision tree structures.
- **Gauss-NB:** Gaussian Naive Bayes exhibits performance with precision (0.684), F1-score (0.711), accuracy (0.698), sensitivity (0.739), specificity (0.655), and an AUC of 0.770.

The model demonstrates reliability in probabilistic classification even under varied data splits.

- **AdaBoost:** maintains reliable performance under SR 80:10:10, achieving precision (0.838), F1-score (0.867), accuracy (0.861), sensitivity (0.897), specificity (0.825), and an AUC ranging from 0.943 to 0.948. AdaBoost's versatility persists across different data partitions.

#### 5.2.3.2 Feature Importance Analysis with Relief Feature Selection

In Table 5.6 and Figure 5.9, the values correspond to the feature importance obtained using the Relief feature selection method. Each feature is represented along with its importance value. The importance values range from highest to lowest, with Age having the highest importance value of 0.1516 and Gender having the lowest importance value of 0.0033. The bar graph provides a visual representation of these importance values, with the color gradient ranging from black for the most significant feature to yellow for the least significant feature, aiding in the interpretation of feature importance.

- **Age**: emerges as the most significant feature, with the highest importance value of 0.1516, indicating its substantial impact on the dataset.
- **Troponin**: Troponin follows closely behind Age, with a high importance value of 0.1471, signifying its significance in the dataset.
- **Temperature**: Temperature exhibits a notable importance value of 0.1285, suggesting its relevance as a key feature.
- **Weight** and **Height**: Weight and Height, with importance values of 0.124 and 0.1115 respectively, represent physical characteristics within the dataset.

- **Respiratory Rate**: Respiratory Rate demonstrates a significant importance value of 0.1092, indicating its impact on the dataset.
- **Body Mass Index**: Body Mass Index (BMI) shows a moderate importance value of 0.085, indicating its relevance in the dataset.
- **Pulse Ox** and **Pain Score**: The combined importance value of Pulse Ox and Pain Score is 0.1398, suggesting their collective impact on the dataset.
- **Gender**: exhibits the lowest importance value among the selected features, with a value of 0.0033, suggesting its minimal impact on the dataset.

**Table 5.6 Top 10 Important Features with Relief Method**

| Features Name | Importance Value |
|---|---|
| Age | 0.1516 |
| Troponin | 0.1471 |
| Temperature | 0.1285 |
| Weight-kilogram | 0.124 |
| Height-cm | 0.1115 |
| Respiratory Rate | 0.1092 |
| Body Mass Index | 0.085 |
| Pulse Ox | 0.0722 |
| Pain Score | 0.0676 |
| Gender | 0.0033 |

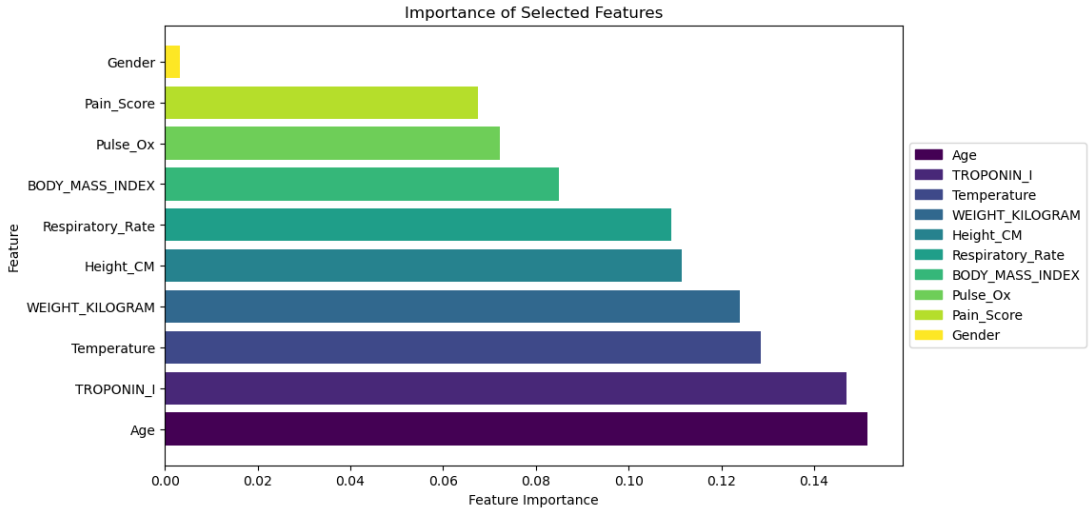


**Figure 5.9 Top 10 Important Features with Relief Method**

#### 5.2.3.3 Conclusion for Relief Feature Selection:

The graphical representations in Figures 5.10, 5.11, and 5.12 highlight the comparison of classifier results for Relief Feature selection with different splitting strategies.

Ensemble methods, particularly Random Forest and X-GBoost, consistently stand out in terms of overall performance. Random Forest achieves outstanding precision, F1-score, and accuracy across various splitting strategies. Similarly, X-GBoost demonstrates robustness with high precision and F1-score, making both these ensemble methods strong contenders for classification tasks.

Gradient Boosting (G-Boost), K-Nearest Neighbors (KNN), Decision Tree (DT), and AdaBoost also exhibit competitive performance. These methods consistently maintain good precision, F1-score, and accuracy, making them reliable choices across different scenarios. Their ability to balance sensitivity and specificity contributes to their overall effectiveness in classification.

While Logistic Regression and Gaussian Naive Bayes (Gauss-NB) show respectable performance, they tend to have slightly lower precision and F1-score compared to ensemble methods. Logistic Regression, in particular, exhibits lower accuracy in certain scenarios. These methods may be considered in contexts where ensemble approaches are not suitable or where a trade-off between interpretability and performance is acceptable.

In summary, Random Forest consistently stands out as a top-performing method across all scenarios, closely followed by X-GBoost. The choice of the best method may depend on specific requirements, such as precision, sensitivity, or overall balanced performance.

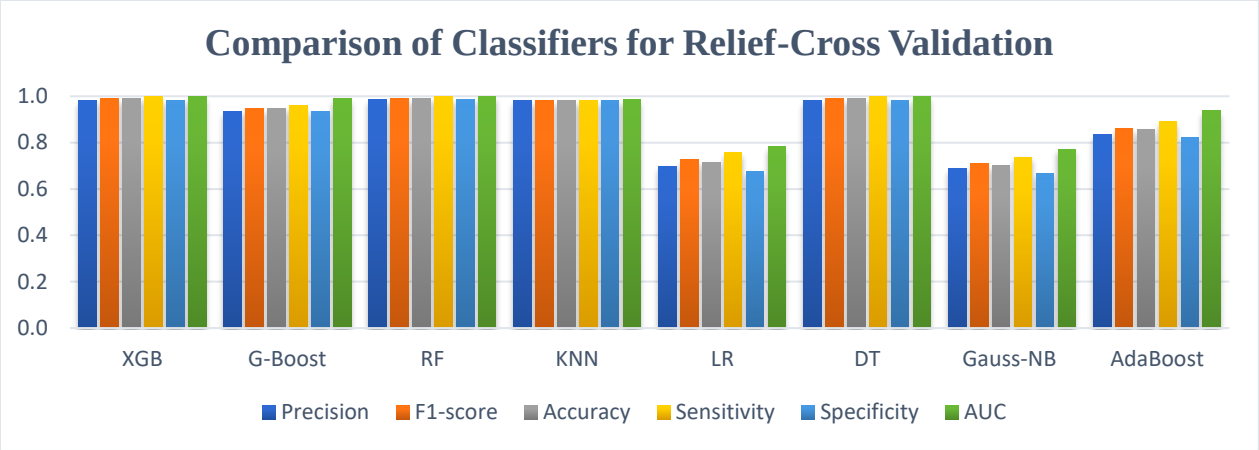


**Figure 5.10 Relief-Cross Validation-Models Results**

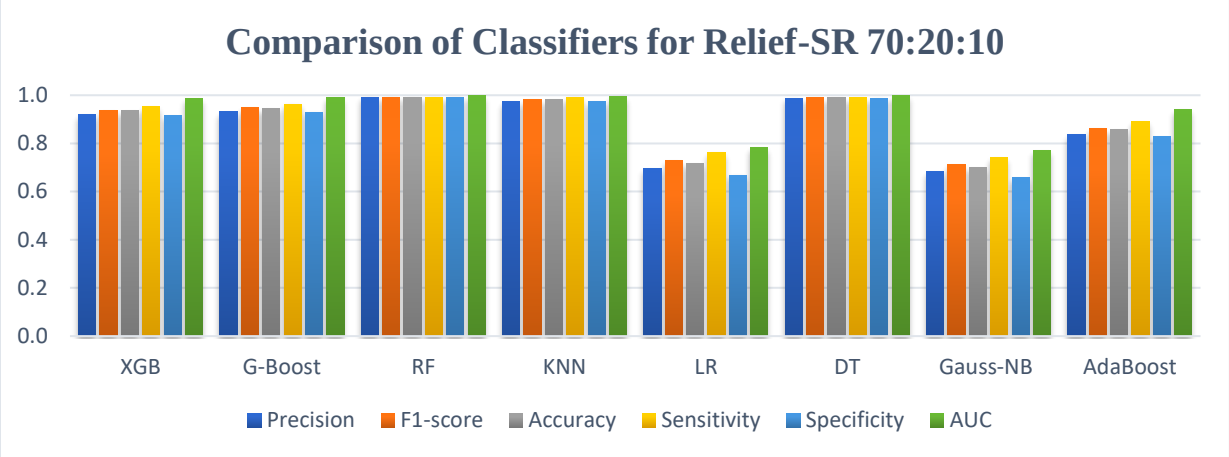


**Figure 5.11 Relief-SR 70:20:10-Models Results**

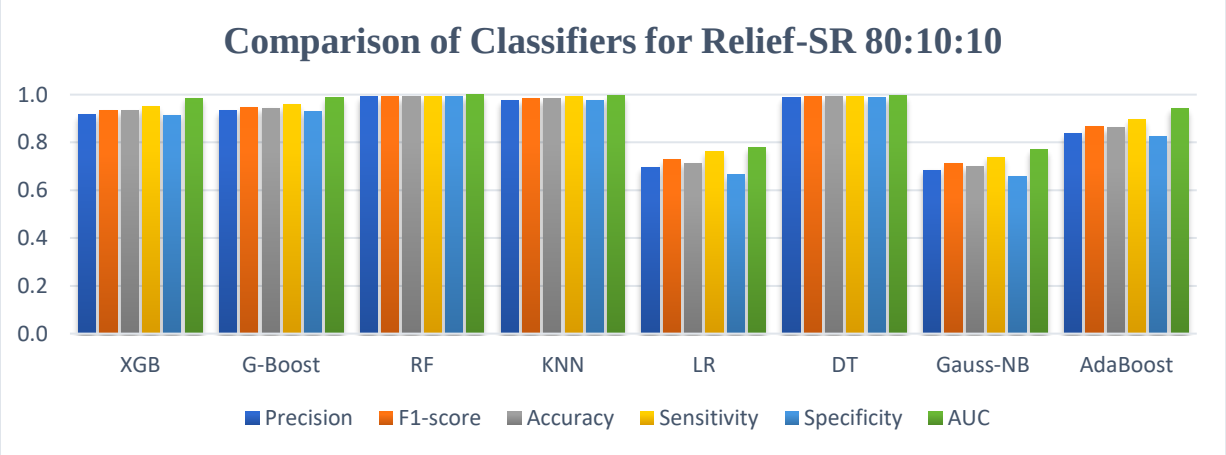


**Figure 5.12 Relief-SR 80:10:10-Models Results**

## 5.3 Discussion

### 5.3.1 Results

In the comprehensive analysis of various classifiers and feature selection methods, considering accuracy, sensitivity, and specificity as key criteria, we can conclude each feature selection method:

**1. PCA:**

- Best Accuracy: 0.991 (RF- 10-fold Cross Validation)
- Best Sensitivity: 1.000 (RF- 10-fold Cross Validation)
- Best Specificity: 0.993 (RF - SR 70:20:10), (RF - SR 80:10:10)
- Best AUC: 1.000 (RF- 10-fold Cross Validation), (RF - SR 70:20:10), (RF - SR 80:10:10)

**2. CFS:**

- Best Accuracy: 0.989 (RF - SR 70:20:10)
- Best Sensitivity: 0.998 (RF - SR 70:20:10), (DT- 10-fold Cross Validation)
- Best Specificity: 0.992 (RF - SR 70:20:10)
- Best AUC: 1.000 (RF - SR 70:20:10), (RF - SR 80:10:10)

**3. Relief feature:**

- Best Accuracy: 0.993 (RF - SR 80:10:10)
- Best Sensitivity: 0.995 (RF - SR 80:10:10)
- Best Specificity: 0.993 (RF - SR 80:10:10)
- Best AUC: 1.000 (RF- 10-fold Cross Validation), (RF - SR 70:20:10), (RF - SR 80:10:10)
  (XGB- 10-fold Cross Validation)

Based on the analysis, both CFS and Relief-feature methods demonstrate slightly better performance compared to PCA, as indicated by their highest values for accuracy, sensitivity, and specificity across various classifiers and scenarios. However, the Relief-feature method combined with Random Forest on the SR 80:10:10 splitting strategy emerges as the most consistently high-performing combination. This suggests that the Relief feature extraction method, particularly when paired with Random Forest and the specified splitting strategy, yields optimal results in terms of performance and accuracy.

Given these findings, the top ten important features identified through the Relief feature extraction method become crucial indicators for analyzing and predicting Acute Aortic Syndrome in emergency department settings. These features, including Age, Troponin levels, Temperature, Weight, Height, Respiratory Rate, Body Mass Index (BMI), Pulse Oximetry readings, Pain Score assessments, and Gender, offer valuable insights into potential risk factors and indicators associated with the syndrome. Leveraging these key features in predictive models enhances the accuracy and efficiency of diagnosis and prognosis processes within emergency departments, ultimately improving patient outcomes and resource allocation strategies.

### 5.3.2 Processing Time

One of the important item in

### 5.3.3 Sample Size

Regarding sample size, it has often been a significant limitation in many research studies in this field. Based on the comparison of all the statistical analyses and machine learning papers explored

in the literature review of this study, it is evident from Figure 5.13, that the largest sample size used was 5548 in McLatchie's [29] research. However, in this study, the integrated dataset is substantially larger than all of them, containing information from 148,707 patients. This vast dataset could serve as an impressive resource for future researchers in the field of Acute Aortic Syndrome (AAS) and other cardiovascular fields.

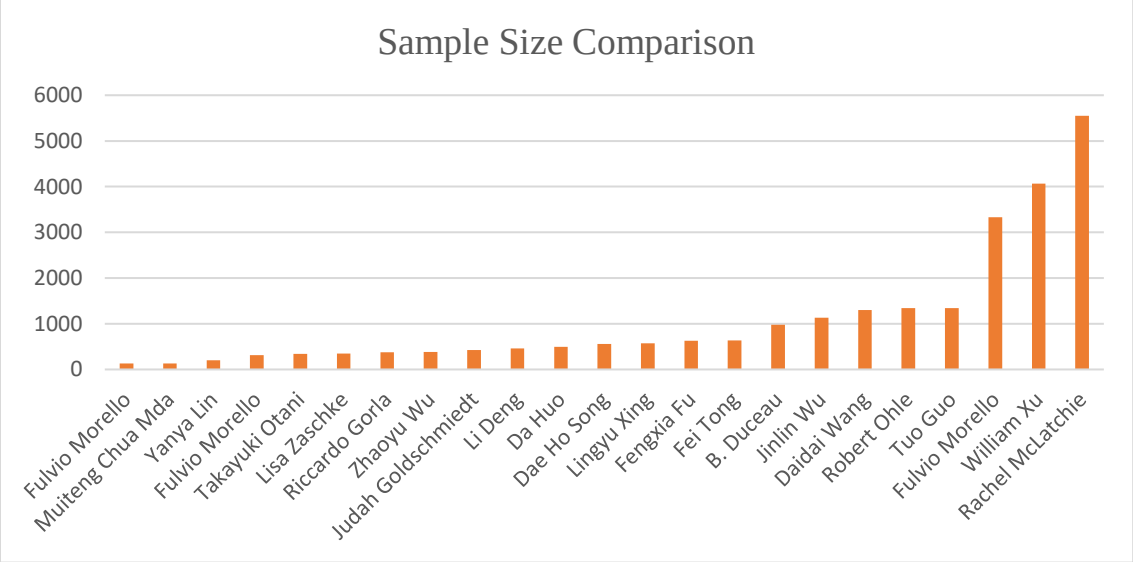


**Figure 5.13 Sample Size Comparison**

# Chapter 6
# Conclusions and Future Work

## 6.1 Conclusions

The goal of this research was to integrate and clean a large clinical dataset relevant to cardiovascular studies, with a specific focus on predicting Acute Aortic Syndrome (AAS). Missing values were addressed using the Multiple Imputations by Chained Equations (MICE) method during data cleaning and integration, proving effective for handling substantial missing data. Given the dataset's imbalance, the SMOTE method was employed to address minority class imbalances.

Three feature selection methods, namely Principal Component Analysis (PCA), Correlation-based Feature Selection (CFS), and Relief Feature, were utilized, each with its advantages and drawbacks. Hence, a comparison was drawn between the results from each method to identify the most effective approach.

Eight classification models—Gradient Boosting, Extreme Gradient Boosting (XGB), Random Forest, Logistic Regression, K-nearest Neighbors, Decision Tree, Gaussian Naive Bayes, and AdaBoost—were applied to the selected features. Evaluation metrics, including accuracy, AUC, precision, F1 score, Sensitivity, and Specificity, were used to gauge model efficacy.

Results revealed that the Relief-feature method with a Split ratio of 80-10-10 and the Random Forest classifier achieved a remarkable accuracy of 99.3%, outperforming other models. However, model performance varied based on the feature selection method employed, highlighting the

importance of choosing an appropriate combination of feature selection and classification methods for a given dataset and task.

In summary, this study integrated a novel clinical dataset comprising 148,707 patient records, applicable to various cardiovascular diseases. It also showcased the potential of machine learning in early AAS prediction in the Emergency Department. Further research is needed to optimize techniques and validate findings with different classifiers or methods. The study lays a foundation for future research in this domain, aiming to develop effective tools for early detection and management of AAS.

## 6.2 Future Work

For future work, the methodologies and findings of this study could be extended to leverage this integrated dataset for the prediction of other cardiovascular diseases and diverse datasets. The techniques employed in this research, specifically in the realms of data preprocessing, feature extraction, and machine learning model application, offer opportunities for further refinement and optimization.

Additionally, the potential for handling imbalanced data could be explored using the ADASYN (Adaptive Synthetic Sampling) technique as an alternative to SMOTE, aiming to address dataset imbalances more effectively.

Moreover, future research endeavors could explore the incorporation of advanced machine learning models, such as deep learning, to obtain more comprehensive results and potentially enhance predictive capabilities. This expansion into cutting-edge methodologies could contribute to a deeper understanding of the dataset and improve the overall effectiveness of predictive models.

Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated remarkable success in various medical procedures, including the detection and classification of cardiovascular diseases. Given the complex nature of Acute Aortic Syndrome (AAS) and the rich information contained within medical procedures data, as well as vital signs and laboratory results datasets, exploring deep learning models presents a promising avenue for improving predictive accuracy. By leveraging the hierarchical feature learning capabilities of deep learning models, it is possible to automatically extract intricate patterns and relationships from medical images and other types of medical data. This approach has the potential to enhance the performance of predictive models for AAS, providing more accurate and reliable predictions. Therefore, exploring deep learning models represents a logical next step to further improve the predictive capabilities of the developed models.

# References


[1] C. A. Nienaber and J. T. Powell, “Management of acute aortic syndromes,” *Eur. Heart J.*, vol. 33, no. 1, pp. 26–35, Jan. 2012, doi: 10.1093/eurheartj/ehr186.

[2] E. M. Isselbacher *et al.*, “2022 ACC/AHA Guideline for the Diagnosis and Management of Aortic Disease,” *J. Am. Coll. Cardiol.*, vol. 80, no. 24, pp. e223–e393, Dec. 2022, doi: 10.1016/j.jacc.2022.08.004.

[3] F. F. Mussa, J. D. Horton, R. Moridzadeh, J. Nicholson, S. Trimarchi, and K. A. Eagle, “Acute Aortic Dissection and Intramural Hematoma: A Systematic Review,” *JAMA*, vol. 316, no. 7, pp. 754–763, Aug. 2016, doi: 10.1001/jama.2016.10026.

[4] P. G. Hagan *et al.*, “The International Registry of Acute Aortic Dissection (IRAD): New Insights Into an Old Disease,” *JAMA*, vol. 283, no. 7, p. 897, Feb. 2000, doi: 10.1001/jama.283.7.897.

[5] W. D. Clouse *et al.*, “Acute Aortic Dissection: Population-Based Incidence Compared With Degenerative Aortic Aneurysm Rupture,” *Mayo Clin. Proc.*, vol. 79, no. 2, pp. 176–180, Feb. 2004, doi: 10.4065/79.2.176.

[6] R. Ohle, D. W. Savage, S. McIsaac, K. Yadav, J. Caswell, and M. Conlon, “Epidemiology, mortality and miss rate of acute aortic syndrome in Ontario, Canada: a population-based study,” *Can. J. Emerg. Med.*, vol. 25, no. 1, pp. 57–64, Jan. 2023, doi: 10.1007/s43678-022-00413-x.

[7] F. Morello, M. Santoro, A. T. Fargion, S. Grifoni, and P. Nazerian, “Diagnosis and management of acute aortic syndromes in the emergency department,” *Intern. Emerg. Med.*, vol. 16, no. 1, pp. 171–181, Jan. 2021, doi: 10.1007/s11739-020-02354-8.

[8] Robert Ohle *et al.*, “Diagnosing acute aortic syndrome: a Canadian clinical practice guideline,” *Can. Med. Assoc. J.*, vol. 192, no. 29, p. E832, Jul. 2020, doi: 10.1503/cmaj.200021.

[9] S. Subhan *et al.*, “Role of Artificial Intelligence and Machine Learning in Interventional Cardiology,” *Curr. Probl. Cardiol.*, vol. 48, no. 7, p. 101698, Jul. 2023, doi: 10.1016/j.cpcardiol.2023.101698.

[10] E. Bossone, T. M. LaBounty, and K. A. Eagle, “Acute aortic syndromes: diagnosis and management, an update,” *Eur. Heart J.*, vol. 39, no. 9, pp. 739–749d, Mar. 2018, doi: 10.1093/eurheartj/ehx319.

[11] B. D. Allen and L. L. Bergmann, “Current Concepts in Acute Aortic Syndrome,” *Adv. Clin. Radiol.*, vol. 3, pp. 139–151, Sep. 2021, doi: 10.1016/j.yacr.2021.04.012.

[12] I. Vilacosta *et al.*, “Acute Aortic Syndrome Revisited,” *J. Am. Coll. Cardiol.*, vol. 78, no. 21, pp. 2106–2125, Nov. 2021, doi: 10.1016/j.jacc.2021.09.022.

[13] I. El Naqa and M. J. Murphy, “What Is Machine Learning?,” in *Machine Learning in Radiation Oncology*, I. El Naqa, R. Li, and M. J. Murphy, Eds., Cham: Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

[14] Z. Ma *et al.*, “Diagnosis of Acute Aortic Syndromes on Non-Contrast CT Images with Radiomics-Based Machine Learning,” *Biology*, vol. 12, no. 3, p. 337, Feb. 2023, doi: 10.3390/biology12030337.

[15] T. Grist and G. D. Rubin, “Acute Aortic Syndrome: State-of-the-Art Diagnostic Imaging,” in *Diseases of the Chest and Heart 2015–2018*, J. Hodler, G. K. Von Schulthess, R. A. Kubik-Huch, and Ch. L. Zollikofer, Eds., Milano: Springer Milan, 2015, pp. 149–156. doi: 10.1007/978-88-470-5752-4_19.

[16] B. Duceau *et al.*, “Prehospital triage of acute aortic syndrome using a machine learning algorithm,” *Br. J. Surg.*, vol. 107, no. 8, pp. 995–1003, Jun. 2020, doi: 10.1002/bjs.11442.

[17] L. Liu *et al.*, “A study of aortic dissection screening method based on multiple machine learning models,” *J. Thorac. Dis.*, vol. 12, no. 3, pp. 605–614, Mar. 2020, doi: 10.21037/jtd.2019.12.119.

[18] M. Radja and A. W. R. Emanuel, “Performance Evaluation of Supervised Machine Learning Algorithms Using Different Data Set Sizes for Diabetes Prediction,” in *2019 5th International Conference on Science in Information Technology (ICSITech)*, Yogyakarta, Indonesia: IEEE, Oct. 2019, pp. 252–258. doi: 10.1109/ICSITech46713.2019.8987479.

[19] S. Safdar, S. Zafar, N. Zafar, and N. F. Khan, “Machine learning based decision support systems (DSS) for heart disease diagnosis: a review,” *Artif. Intell. Rev.*, vol. 50, no. 4, pp. 597–623, Dec. 2018, doi: 10.1007/s10462-017-9552-8.

[20] M. Chua, I. Ibrahim, X. Neo, V. Sorokin, L. Shen, and S. B. S. Ooi, “Acute aortic dissection in the ED: risk factors and predictors for missed diagnosis,” *Am. J. Emerg. Med.*, vol. 30, no. 8, pp. 1622–1626, Oct. 2012, doi: 10.1016/j.ajem.2011.11.017.

[21] T. Otani, T. Abe, T. Ichiba, K. Kashiwa, and H. Naito, “D-dimer measurement is useful irrespective of time from the onset of acute aortic syndrome symptoms,” *Am. J. Emerg. Med.*, vol. 71, pp. 7–13, Sep. 2023, doi: 10.1016/j.ajem.2023.05.044.

[22] Y. Von Kodolitsch, A. G. Schwartz, and C. A. Nienaber, “Clinical Prediction of Acute Aortic Dissection,” *Arch. Intern. Med.*, vol. 160, no. 19, p. 2977, Oct. 2000, doi: 10.1001/archinte.160.19.2977.

[23] L. Zaschke *et al.*, “Acute type A aortic dissection: Aortic Dissection Detection Risk Score in emergency care – surgical delay because of initial misdiagnosis,” *Eur. Heart J. Acute Cardiovasc. Care*, vol. 9, no. 3_suppl, pp. S40–S47, Oct. 2020, doi: 10.1177/2048872620914931.

[24] J. Goldschmiedt, J. M. Levsky, E. Y. Bellin, E. Mizrachi, D. Esses, and L. B. Haramati, "Prospective study of a non-restrictive decision rule for acute aortic syndrome," *Am. J. Emerg. Med.*, vol. 35, no. 9, pp. 1309–1313, Sep. 2017, doi: 10.1016/j.ajem.2017.04.014.

[25] W. Xu *et al.*, "Acute aortic syndrome: nationwide study of epidemiology, management, and outcomes," *Br. J. Surg.*, vol. 110, no. 9, pp. 1197–1205, Aug. 2023, doi: 10.1093/bjs/znad162.

[26] P. Bima *et al.*, "Systematic Review of Aortic Dissection Detection Risk Score Plus D-dimer for Diagnostic Rule-out Of Suspected Acute Aortic Syndromes," *Acad. Emerg. Med.*, vol. 27, no. 10, pp. 1013–1027, Oct. 2020, doi: 10.1111/acem.13969.

[27] M. S. Yellapragada, Y. Xie, B. Graf, D. Richmond, A. Krishnan, and A. Sitek, "Deep Learning Based Detection of Acute Aortic Syndrome in Contrast CT Images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA: IEEE, Apr. 2020, pp. 1474–1477. doi: 10.1109/ISBI45749.2020.9098362.

[28] F. Morello *et al.*, "Pre-Test Probability Assessment and d-Dimer Based Evaluation in Patients with Previous Acute Aortic Syndrome," *Medicina (Mex.)*, vol. 59, no. 3, p. 548, Mar. 2023, doi: 10.3390/medicina59030548.

[29] R. McLatchie *et al.*, "Diagnosis of Acute Aortic Syndrome in the Emergency Department (DAShED) study: an observational cohort study of people attending the emergency department with symptoms consistent with acute aortic syndrome," *Emerg. Med. J.*, p. emermed-2023-213266, Nov. 2023, doi: 10.1136/emermed-2023-213266.

[30] D. Wang *et al.*, "Early Prediction Model of Acute Aortic Syndrome Mortality in Emergency Departments," *Int. J. Gen. Med.*, vol. Volume 15, pp. 3779–3788, Apr. 2022, doi: 10.2147/IJGM.S357910.

[31] R. Ohle *et al.*, "Evaluation of the Canadian Clinical Practice Guidelines Risk Prediction Tool for Acute Aortic Syndrome: The RIPP Score," *Emerg. Med. Int.*, vol. 2023, pp. 1–7, May 2023, doi: 10.1155/2023/6636800.

[32] R. Gorla *et al.*, "Accuracy of a diagnostic strategy combining aortic dissection detection risk score and D-dimer levels in patients with suspected acute aortic syndrome," *Eur. Heart J. Acute Cardiovasc. Care*, vol. 6, no. 5, pp. 371–378, Aug. 2017, doi: 10.1177/2048872615594497.

[33] L. Deng *et al.*, "Aortic Dissection Detection Risk Score and D-Dimer for Acute Aortic Syndromes in the Chinese Population: Exploration of Optimal Thresholds and Integrated Diagnostic Value," *J Cardiovasc. Transl. Res.*, vol. 16, no. 4, pp. 886–895, Aug. 2023, doi: 10.1007/s12265-023-10354-0.

[34] F. Morello *et al.*, "Prospective diagnostic and prognostic study of copeptin in suspected acute aortic syndromes," *Sci. Rep.*, vol. 8, no. 1, p. 16713, Nov. 2018, doi: 10.1038/s41598-018-35016-z.

[35] D. H. Song, J. H. Choi, and J. Y. Lee, “Predicting acute aortic syndrome using aortic dissection detection risk score, D-dimer, and X-ray,” *Heliyon*, vol. 9, no. 10, p. e20578, Oct. 2023, doi: 10.1016/j.heliyon.2023.e20578.

[36] F. Fu *et al.*, “Acute aortic syndrome risk stratification score: A new risk assessment tool,” In Review, preprint, Jul. 2023. doi: 10.21203/rs.3.rs-3153344/v1.

[37] F. Tong, Y. Wang, and Z. Sun, “Development and validation of nomogram models to discriminate between acute aortic syndromes and non-ST-elevation myocardial infarction during troponin-blind period,” *Front. Cardiovasc. Med.*, vol. 10, p. 1077712, Jan. 2023, doi: 10.3389/fcvm.2023.1077712.

[38] F. Morello *et al.*, “Development and Validation of a Simplified Probability Assessment Score Integrated With Age-Adjusted D -Dimer for Diagnosis of Acute Aortic Syndromes,” *J. Am. Heart Assoc.*, vol. 10, no. 3, p. e018425, Feb. 2021, doi: 10.1161/JAHA.120.018425.

[39] L. Xing *et al.*, “Simple Death Risk Models to Predict In-hospital Outcomes in Acute Aortic Dissection in Emergency Department,” *Front. Med.*, vol. 9, p. 890567, May 2022, doi: 10.3389/fmed.2022.890567.

[40] D. Huo, B. Kou, Z. Zhou, and M. Lv, “A machine learning model to classify aortic dissection patients in the early diagnosis phase,” *Sci. Rep.*, vol. 9, no. 1, p. 2701, Feb. 2019, doi: 10.1038/s41598-019-39066-9.

[41] J. Wu *et al.*, “Predicting in-hospital rupture of type A aortic dissection using random forest,” *J. Thorac. Dis.*, vol. 11, no. 11, pp. 4634–4646, Nov. 2019, doi: 10.21037/jtd.2019.10.82.

[42] T. Guo *et al.*, “Machine Learning Models for Predicting In-Hospital Mortality in Acute Aortic Dissection Patients,” *Front. Cardiovasc. Med.*, vol. 8, p. 727773, Sep. 2021, doi: 10.3389/fcvm.2021.727773.

[43] Z. Wu *et al.*, “Prediction of preoperative in-hospital mortality rate in patients with acute aortic dissection by machine learning: a two-centre, retrospective cohort study,” *BMJ Open*, vol. 13, no. 4, p. e066782, Apr. 2023, doi: 10.1136/bmjopen-2022-066782.

[44] Y. Lin *et al.*, “Application of Logistic Regression and Artificial Intelligence in the Risk Prediction of Acute Aortic Dissection Rupture,” *J. Clin. Med.*, vol. 12, no. 1, p. 179, Dec. 2022, doi: 10.3390/jcm12010179.

[45] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, “An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases,” *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[46] J. Li *et al.*, “Machine Learning Prediction Model for Acute Renal Failure After Acute Aortic Syndrome Surgery,” *Front. Med.*, vol. 8, p. 728521, Jan. 2022, doi: 10.3389/fmed.2021.728521.

[47] J. Klaudel, B. Klaudel, M. Glaza, W. Trenkner, P. Derejko, and M. Szołkiewicz, "Forewarned Is Forearmed: Machine Learning Algorithms for the Prediction of Catheter-Induced Coronary and Aortic Injuries," *Int. J. Environ. Res. Public. Health*, vol. 19, no. 24, p. 17002, Dec. 2022, doi: 10.3390/ijerph192417002.

[48] G. Lei, G. Wang, C. Zhang, Y. Chen, and X. Yang, "Using Machine Learning to Predict Acute Kidney Injury After Aortic Arch Surgery," *J. Cardiothorac. Vasc. Anesth.*, vol. 34, no. 12, pp. 3321–3328, Dec. 2020, doi: 10.1053/j.jvca.2020.06.007.

[49] L. Zhang *et al.*, "Identification of Clinical Heterogeneity and Construction of Prediction Models for Novel Subtypes in Patients with Abdominal Aortic Aneurysm: An Unsupervised Machine Learning Study," *Ann. Vasc. Surg.*, vol. 98, pp. 75–86, Jan. 2024, doi: 10.1016/j.avsg.2023.06.013.

[50] N. P. Ostberg, M. A. Zafar, S. K. Mukherjee, B. A. Ziganshin, and J. A. Elefteriades, "A machine learning approach for predicting complications in descending and thoracoabdominal aortic aneurysms," *J. Thorac. Cardiovasc. Surg.*, vol. 166, no. 4, pp. 1011-1020.e3, Oct. 2023, doi: 10.1016/j.jtcvs.2021.12.045.

[51] Y. Wang *et al.*, "A radiomics model for predicting the outcome of endovascular abdominal aortic aneurysm repair based on machine learning," *Vascular*, vol. 31, no. 4, pp. 654–663, Aug. 2023, doi: 10.1177/17085381221091061.

[52] J. Ke *et al.*, "Machine learning-based in-hospital mortality prediction models for patients with acute coronary syndrome," *Am. J. Emerg. Med.*, vol. 53, pp. 127–134, Mar. 2022, doi: 10.1016/j.ajem.2021.12.070.

[53] J. Emakhu *et al.*, "Acute coronary syndrome prediction in emergency care: A machine learning approach," *Comput. Methods Programs Biomed.*, vol. 225, p. 107080, Oct. 2022, doi: 10.1016/j.cmpb.2022.107080.

[54] D. R. Sax *et al.*, "Use of Machine Learning to Develop a Risk-Stratification Tool for Emergency Department Patients With Acute Heart Failure," *Ann. Emerg. Med.*, vol. 77, no. 2, pp. 237–248, Feb. 2021, doi: 10.1016/j.annemergmed.2020.09.436.

[55] F. Ahmad, N. Cheshire, and M. Hamady, "Acute aortic syndrome: pathology and therapeutic strategies," *Postgrad. Med. J.*, vol. 82, no. 967, pp. 305–312, May 2006, doi: 10.1136/pgmj.2005.043083.

[56] C. Elendu, D. C. Amaechi, T. C. Elendu, J. O. Ibhiedu, A. O. Torubiri, and O. K. Okoye, "Comprehensive review of aortic aneurysms, dissections, and cardiovascular complications in connective tissue disorders," *Medicine (Baltimore)*, vol. 102, no. 48, p. e36499, Dec. 2023, doi: 10.1097/MD.0000000000036499.

[57] S. K. Gill *et al.*, "Artificial intelligence to enhance clinical value across the spectrum of cardiovascular healthcare," *Eur. Heart J.*, vol. 44, no. 9, pp. 713–725, Mar. 2023, doi: 10.1093/eurheartj/ehac758.

[58] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, and C. Ding, “Nbias: A natural language processing framework for BIAS identification in text,” *Expert Syst. Appl.*, vol. 237, p. 121542, Mar. 2024, doi: 10.1016/j.eswa.2023.121542.

[59] P. Heus *et al.*, “Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies,” *BMJ Open*, vol. 9, no. 4, p. e025611, Apr. 2019, doi: 10.1136/bmjopen-2018-025611.

[60] C. Ordonez, “Data set preprocessing and transformation in a database system,” *Intell. Data Anal.*, vol. 15, no. 4, pp. 613–631, Jun. 2011, doi: 10.3233/IDA-2011-0485.

[61] J. Leipzig, D. Nüst, C. T. Hoyt, K. Ram, and J. Greenberg, “The role of metadata in reproducible computational research,” *Patterns*, vol. 2, no. 9, p. 100322, Sep. 2021, doi: 10.1016/j.patter.2021.100322.

[62] Subasi, Abdulhamit., *Practical machine learning for data analysis using python. Academic Press, 2020.*

[63] Nelli and Fabio, *Data analysis and science using PANDAs, Matplotlib and the Python Programming Language. Apress, 2015.*

[64] K. W. Fung, Xu Julia, and O. Bodenreider, “The new International Classification of Diseases 11th edition: a comparative analysis with ICD-10 and ICD-10-CM,” *J. Am. Med. Inform. Assoc.*, vol. 27, no. 5, pp. 738–746, May 2020, doi: 10.1093/jamia/ocaa030.

[65] E. W. Steyerberg, *Clinical Prediction Models*. in Statistics for Biology and Health. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-77244-8.

[66] Little, R.J.A and Rubin, *Statistical analysis with missing data, 2nd edn*.

[67] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.

[68] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” *J. Big Data*, vol. 8, no. 1, p. 140, Oct. 2021, doi: 10.1186/s40537-021-00516-9.

[69] J. R. Carpenter and M. G. Kenward, *Multiple Imputation and its Application*, 1st ed. Wiley, 2013. doi: 10.1002/9781119942283.

[70] M. Pampaka, G. Hutcheson, and J. Williams, “Handling missing data: analysis of a challenging data set using multiple imputation,” *Int. J. Res. Method Educ.*, vol. 39, no. 1, pp. 19–37, Jan. 2016, doi: 10.1080/1743727X.2014.979146.

[71] A. Pedersen *et al.*, “Missing data and multiple imputation in clinical epidemiological research,” *Clin. Epidemiol.*, vol. Volume 9, pp. 157–166, Mar. 2017, doi: 10.2147/CLEP.S129785.

[72] S. Greenland and W. D. Finkle, “A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses,” *Am. J. Epidemiol.*, vol. 142, no. 12, pp. 1255–1264, Dec. 1995, doi: 10.1093/oxfordjournals.aje.a117592.

[73] A. R. T. Donders, G. J. M. G. Van Der Heijden, T. Stijnen, and K. G. M. Moons, “Review: A gentle introduction to imputation of missing values,” *J. Clin. Epidemiol.*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006, doi: 10.1016/j.jclinepi.2006.01.014.

[74] K. M. Lang and T. D. Little, “Principled Missing Data Treatments,” *Prev. Sci.*, vol. 19, no. 3, pp. 284–294, Apr. 2018, doi: 10.1007/s11121-016-0644-5.

[75] Schafer, J.L and Chapman & Hall, *Analysis of incomplete multivariate data*.

[76] Van Buuren. and Chapman and Hall/CRC, *Flexible imputation of missing dataFlexible imputation of missing data*.

[77] S. J. Gibson *et al.*, “Using Multivariate Imputation by Chained Equations to Predict Redshifts of Active Galactic Nuclei,” *Front. Astron. Space Sci.*, vol. 9, p. 836215, Mar. 2022, doi: 10.3389/fspas.2022.836215.

[78] D. Zhang, L. Zou, X. Zhou, and F. He, “Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer,” *IEEE Access*, vol. 6, pp. 28936–28944, 2018, doi: 10.1109/ACCESS.2018.2837654.

[79] S. Mishra *et al.*, “Principal Component Analysis,” *Int. J. Livest. Res.*, p. 1, 2017, doi: 10.5455/ijlr.20170415115235.

[80] S. Ibrahim, S. Nazir, and S. A. Velastin, “Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis,” *J. Imaging*, vol. 7, no. 11, p. 225, Oct. 2021, doi: 10.3390/jimaging7110225.

[81] K. Dissanayake and M. G. Md Johar, “Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms,” *Appl. Comput. Intell. Soft Comput.*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/5581806.

[82] P. Palimkar, R. N. Shaw, and A. Ghosh, “Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach,” in *Advanced Computing and Intelligent Technologies*, vol. 218, M. Bianchini, V. Piuri, S. Das, and R. N. Shaw, Eds., in Lecture Notes in Networks and Systems, vol. 218. , Singapore: Springer Singapore, 2022, pp. 219–244. doi: 10.1007/978-981-16-2164-2_19.

[83] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[84] K. Yongcharoenchaiyasit, S. Arwatchananukul, P. Temdee, and R. Prasad, “Gradient Boosting Based Model for Elderly Heart Failure, Aortic Stenosis, and Dementia

Classification," *IEEE Access*, vol. 11, pp. 48677–48696, 2023, doi: 10.1109/ACCESS.2023.3276468.

[85] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *J. Data Anal. Inf. Process.*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[86] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[87] M. V. Anand, B. KiranBala, S. R. Srividhya, K. C., M. Younus, and M. H. Rahman, "Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer," *Mob. Inf. Syst.*, vol. 2022, pp. 1–7, Jun. 2022, doi: 10.1155/2022/2436946.

[88] C. Bemando, E. Miranda, and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, Pekan, Malaysia: IEEE, Aug. 2021, pp. 232–237. doi: 10.1109/ICSECS52883.2021.00049.

[89] W. Wang and D. Sun, "The improved AdaBoost algorithms for imbalanced data classification," *Inf. Sci.*, vol. 563, pp. 358–374, Jul. 2021, doi: 10.1016/j.ins.2021.03.042.