



People's Democratic Republic of Algeria
Echahid Cheikh Larbi Tebessi University



Faculty of Exact Sciences and Natural and Life Sciences

Department of Math and Computer Science

2023/2024

Master's thesis

In computer science

Specialty: Information system

An Interpretable Rough Sets-Based Model for Efficient and Accurate Cardiovascular Risk Assessment

Realized by: ZGA Rania
MERAMERIA Oulfa

Infront of jury members:

Dr. BOUROUGAA Salima (MCA)	President
Dr. CHERGUI Othaila (MAA)	Examiner
Dr. KHELIFA Boudjemaa (MAB)	Thesis supervisor
Dr. BENDIB Issam (MCA)	Co-supervisor

Date of defense: June ninth, 2024

ACKNOWLEDGMENTS

*To those who instilled ambition and determination in me,
To my dear family, The source of my support and inspiration, who have given me
everything to reach this day. Thank you for your unconditional love and countless
sacrifices.*

*To my esteemed professors, who taught me that knowledge is the path to achieving dreams
and opened the doors of growth and learning for me.*

*To my dear friends, who shared with me the difficult and beautiful moments and provided
me with invaluable moral support.*

*To everyone who believed in my abilities and gave me the opportunity to prove myself,
I dedicate to you all the fruit of my efforts and years of dedication and hard work, and I
ask God to make it the beginning of a journey crowned with success and prosperity.*

 *Rania*

ACKNOWLEDGMENTS

To my mother:

To my dear mother, the journey to graduation was not easy, but your efforts and sacrifices made it possible. To my dear mother, this memoir is a tribute to your hard work and boundless love.

To my father:

To my late father, through every study hour and moment of dedication, you were the guiding star lighting my path. To the spirit of my father, this memoir is an expression of my deep gratitude and memories filled with pride.

To my siblings:

To Achref and Thamer, with every participation and encouragement, you were my siblings on the road to success. To Achref and Thamer, this memoir is a dedication to you, to the spirit of brotherhood and friendship that is unforgettable.

To my companion:

To my dear companion Hachem, words of thanks cannot suffice for what you have provided me throughout this journey. This memoir is an expression of my deep gratitude and remembrance of your supportive and loving spirit.

 Oulfa

Abstract

Cardiovascular disease stands as the foremost cause of mortality globally, responsible for an estimated 17.9 million deaths annually (source: World Health Organization¹). Detecting and preventing cardiovascular disease at its onset are imperative steps in alleviating its burdensome impact. However, conventional risk assessment models often exhibit complexity and lack interpretability, posing challenges for practical utilization by clinicians.

This master's thesis project endeavors to forge ahead by crafting an interpretable rough sets-based model tailored for proficient and precise cardiovascular risk assessment. Leveraging rough sets, a machine learning technique adept at distilling insights from data into a comprehensible format, holds promise in this pursuit. The proposed model will be meticulously constructed utilizing The Cleveland Heart Disease Dataset, its efficacy is gauged across an array of metrics encompassing accuracy, interpretability, and efficiency.

With an eye towards transformative potential, the envisioned model aims to revolutionize early detection and prevention strategies for cardiovascular disease. By furnishing clinicians with a more interpretable and streamlined tool for risk assessment, the proposed model is poised to catalyze advancements in cardiovascular healthcare delivery, fostering improved patient outcomes and quality of life.

¹ Health topics/cardiovascular diseases: <https://www.who.int/health-topics/cardiovascular-diseases>

Résumé

Les maladies cardiovasculaires constituent la principale cause de mortalité dans le monde, responsables de près de 17,9 millions de décès par an (source : Organisation mondiale de la Santé). Détecter et prévenir les maladies cardiovasculaires dès leur apparition sont des étapes impératives pour alléger leur impact lourd. Cependant, les modèles d'évaluation des risques conventionnels sont souvent complexes et manquent d'interprétabilité, posant des défis pour une utilisation pratique par les cliniciens.

Ce projet de thèse de master s'efforce de progresser en créant un modèle interprétable basé sur les ensembles approximatifs, adapté pour une évaluation des risques cardiovasculaires précise et efficace. En exploitant les ensembles approximatifs, une technique d'apprentissage automatique habile à extraire des informations des données dans un format compréhensible, ce projet détient des promesses en ce sens. Le modèle proposé sera minutieusement construit en utilisant l'ensemble de données sur les maladies cardiaques « Cleveland », son efficacité évaluée à travers une gamme de métriques incluant l'exactitude, l'interprétabilité et l'efficacité.

Avec une vision de potentiel transformationnel, le modèle envisagé vise à révolutionner les stratégies de détection précoce et de prévention des maladies cardiovasculaires. Un fournisseur aux cliniciens un outil d'évaluation des risques plus interprétable et simplifié, le modèle proposé est prêt à catalyser des avancées dans la prestation des soins de santé cardiovasculaires, favorisant de meilleurs résultats pour les patients et une meilleure qualité de vie.

نبذة مختصرة

تعد أمراض القلب والأوعية الدموية السبب الرئيسي للوفيات على مستوى العالم، وهي مسؤولة عن ما يقدر بحوالي 17.9 مليون حالة وفاة سنويًا (المصدر: منظمة الصحة العالمية). يعتبر الكشف المبكر عن أمراض القلب والأوعية الدموية والوقاية منها في مراحلها الأولى خطوة هامة للتخفيف من تأثيراتها الوخيمة. إلا أنه، غالبًا ما تتسم نماذج تقييم المخاطر التقليدية بالتعقيد والافتقار إلى قابلية التفسير، مما يشكل تحديات للاستفادة العملية منها من قبل الأطباء والباحثين.

يهدف مشروع أطروحة الماجستير هذا إلى المضي قدما من خلال استحداث نموذج قابل للتفسير يعتمد على نظرية المجموعات التقريبية مصمم لتقييم مخاطر أمراض القلب والأوعية الدموية بشكل أكثر دقة وفعالية. إن الاستفادة من المجموعات التقريبية وهي تقنية واعدة من بين تقنيات التعلم الآلي القادرة على توليد المعرفة بطريقة قابلة للفهم انطلاقا من مجموعة المعطيات. سيتم بناء النموذج المقترح بعناية باستخدام مجموعة بيانات أمراض القلب من كليفلاند (Cleveland)، وسيتم تقييم فعاليته وفقا لجملة المعايير التي تشمل الدقة، قابلية التفسير، والكفاءة.

يهدف النموذج المرتقب إلى إحداث قزة نوعية في استراتيجيات الكشف المبكر والوقاية من أمراض القلب والأوعية الدموية. ومن خلال تزويد الأطباء بأداة أكثر سلاسة وقابلية للتفسير، يُتوقع أن يسهم هذا النموذج في تعزيز تقديم الرعاية الصحية القلبية، وتحسين نتائج علاج المرضى ومنحهم حياة أفضل.

Table of Contents

Table of Acronyms	10
General Introduction	11
Chapter 01: Background and related work	14
1.1. Introduction.....	15
1.2. Cardiovascular risk	15
.1.2.1 High blood pressure (arterial hypertension)	16
.1.2.2 Hypercholesterolemia	16
.1.2.3 Coronary artery disease.....	16
.1.2.4 Cardiac arrest	16
.1.2.5 Valvar heart disease	17
1.3. Heart disease screening.....	17
1.3.1. Heart disease tests	17
1.4. Artificial intelligence and CVD early detection.....	19
1.5. Related works.....	20
1.6. State of the art summary	25
1.7. Conclusion	27
Chapter 02: ML and Rough Sets for Interpretable Diagnosis	28
2.1. Introduction.....	29
2.2. Machine learning	29
2.2.1. Supervised learning.....	30
2.2.2. Unsupervised learning	30
2.2.3. Reinforcement learning.....	30
2.2.4. Classification.....	30
2.3. Common issues in ML	31

2.4.	The rough sets approach	33
2.4.1.	Basic problems in data analysis solved by Rough Set:.....	34
2.4.2.	Goals of Rough Set Theory:	35
2.4.3.	Attributes in Rough Sets:	35
2.4.4.	Rule extraction:	36
2.4.5.	Upper Approximation:	37
2.4.6.	Lower Approximation:.....	38
2.5.	Conclusion	40
Chapter 03: The proposed methodology		41
3.1.	Introduction.....	42
3.2.	Training method.....	42
3.3.	The Cleveland Heart Disease Dataset.....	43
3.4.	Pretreatment	45
3.4.1.	The Mean Imputation technique:	45
3.4.2.	Handling Missing Values with Rough Sets	46
3.4.3.	Discretization:	46
3.5.	Performance metrics	47
3.6.	Software Tools for Rough Set Modeling.....	48
3.7.	The RSES2 Tool	49
3.8.	The proposed methodology	50
3.8.1.	Scenario 1.....	51
3.8.2.	Scenario 2.....	53
3.8.3.	Scenario 3.....	55
3.9.	Discussion:	58
3.10.	Conclusion	60
Chapter 04: Experimental Study.....		61
1.1	Introduction.....	62

4.1.	Experimental Environment.....	62
4.1.1.	Hardware.....	62
4.1.2.	Software.....	62
4.2.	Experimental results	63
4.2.1.	Scenario 1.....	63
4.2.2.	Scenario 2.....	65
4.2.3.	Scenario 3.....	66
4.2.4.	Summary of experimental results	68
4.3.	Comparisons	70
4.4.	Validation.....	71
4.5.	Future prospects for system improvement.....	71
4.6.	Conclusion	73
	General Conclusion.....	74
	Bibliography	75
	List of figures.....	79
	List of Tables	80
	List of Equations	81

Table of Acronyms

Acronym	Meaning
AI	Artificial Intelligence
ANN	Artificial Neural Networks
BN	Bayesian Networks
CAD	Coronary Artery Disease
CHD	Coronary Heart Disease
CVD	Cardiovascular Disease
DSS	Decision Support Systems
ECG	Electrocardiogram
HDL	High-Density Lipoproteins
HMM	Hidden Markov Models
IHD	Ischemic Heart Disease
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LDL	Low-Density Lipoproteins
LEM2	Learning From Examples Module, Version 2
MI	Myocardial Infarction
ML	Machine Learning
MLP	Multi-Layer Perceptron Neural Network
NN	Neural Networks
OS	Operating System
RSES	The Rough Set Exploration System
ROC	Operating Characteristic Curve
SVM	Support Vector Machine
KDD	Knowledge Discovery In Databases

General Introduction

Cardiovascular diseases pose a global health challenge, leading to millions of deaths annually and causing significant suffering for millions due to disability and reduced quality of life. Many individuals suffer from a wide range of these diseases, including coronary heart disease, stroke, and arterial diseases, which require immediate intervention and intensive treatment to avoid serious complications [1].

Despite the tremendous medical advancements in the treatment of cardiovascular diseases, the main challenge lies in predicting risks and early diagnosis of these diseases, as it helps in making more effective treatment decisions and guiding prevention efforts better. However, traditional models used to assess the risks of cardiovascular diseases are often complex and difficult to interpret, making it challenging to use them effectively in clinical practice.

This is where Machine Learning (ML) comes in. Modern health care has, over the years, greatly benefited from the progress in theory and practice of health information systems. This health informatics have been shown to have the potential for positive impact on the quality and efficiency of patient care [2].

By analyzing various personal factors, such as age, family history, and health indicators like blood pressure, we aim to develop an interpretable model. This model will effectively assess an individual's risk profile and provide early warnings for potential cardiovascular problems. The interpretability of the model is crucial, allowing healthcare professionals to understand the reasoning behind the risk assessment and personalize treatment plans based on the identified factors.

This study aims to develop an innovative model based on rough sets theory, capable of accurately and effectively evaluating cardiovascular disease risks. The rough sets knowledge discovery technique will be utilized to develop this model, with the goal of providing a diagnostic tool that can guide physicians and enable them to manage better and faster patients at risk. The main contributions of this work are as follows:

- ☞ Develop an interpretable rough sets-based model for cardiovascular risk assessment.
- ☞ Evaluate the performance of the proposed model on a variety of metrics, including accuracy, interpretability, and efficiency.

- ☞ Compare the proposed model to traditional cardiovascular risk assessment models.
- ☞ Identify potential applications of the proposed model in clinical practice.

With that said, chapters in this thesis will be organized as such:

Chapter 01:

- ☞ We begin by delving into the global burden of cardiovascular disease (CVD) and its profound impact on individuals and healthcare systems worldwide.
- ☞ Recognizing the critical importance of early detection and risk stratification in effectively managing CVD.
- ☞ We introduce the general artificial intelligence modal and the concept of interpretable machine learning and highlight their significance in clinical settings, particularly in addressing early detection withing the complexities of CVD.
- ☞ Finally, we provide a comprehensive overview of some related works to the discovery of heart disease, encompassing the utilized datasets, methodologies employed, and resulting outcomes.

Chapter 02:

- ☞ provide an overview of rough set theory and its applicability in constructing interpretable models, especially when dealing with incomplete or imprecise data.
- ☞ Machine learning and artificial intelligence methods were discussed as methods that can be relied upon in this type of study.

Chapter 03:

- ☞ We provide a detailed description of heart disease dataset, elucidating its characteristics, data collection methodology, and potential limitations.
- ☞ Additionally, we discuss our approach to data preprocessing, which includes strategies for handling missing values and outliers.
- ☞ We offer a concise overview of key rough sets theory concepts, such as information systems, indiscernibility relation, rough sets, and decision rules.
- ☞ Furthermore, we explain how rough sets can effectively manage incomplete and imprecise data, particularly in the context of CVD prediction.
- ☞ Our methodology involves the application of rough set algorithms to heart disease dataset.
- ☞ We outline our process for selecting and evaluating decision rules, emphasizing criteria such as accuracy, strength, and coverage.

- 👉 Moreover, we discuss techniques employed to ensure the interpretability and efficiency of our model, including rule pruning and simplification.
- 👉 We define the metrics utilized to evaluate our model's performance, encompassing measures such as accuracy, precision, recall, coverage, and F1 score.

Chapter04:

- 👉 We present our findings, including the generated decision rules, model performance metrics, and any supporting visualizations.
- 👉 Our discussion contextualizes the results within the existing literature, elucidating the strengths and limitations of our proposed model.
- 👉 Furthermore, we analyze the interpretability of the decision rules generated by our model and their potential clinical relevance.
- 👉 Additionally, we compare our model's performance to existing studies, providing insights into any similarities or differences observed.
- 👉 In conclusion, we summarize our key findings and underscore the significance of our research, particularly the development of an interpretable rough set model for CVD risk prediction.

We outline potential avenues for future research, including exploring alternative rough set algorithms, applying our model to diverse datasets, and integrating it with complementary prediction methodologies.

Chapter 01:

Background and related work

1.1. Introduction

In this chapter, we provide an introduction to the global burden of CVD, highlighting its impact on individuals and healthcare systems. In order to build an interpretable model for cardiovascular risk assessment capable of providing truthful predictions and achieving high accuracy, it is natural that we first get a proper grasp on heart's disease, its symptoms, and what is to be expected from someone afflicted with this illness. Then we move on to identifying the necessary tests and their correlation with diagnosing cardiovascular disease, along with existing datasets related to this subject. The next step is to provide a comprehensive overview and a critical analysis of prior research endeavors dedicated to the discovery of heart disease, encompassing the utilized datasets, methodologies employed, and resulting outcomes. This analysis will serve as a foundation for our own research approach, allowing us to identify strengths and weaknesses in current methods and opportunities for further advancement.

1.2. Cardiovascular risk

Cardiovascular diseases are a group of disorders and issues that affect the heart and blood vessels. They encompass a variety of conditions ranging from diseases affecting the heart muscle itself to those impacting the blood vessels that supply the heart and the vascular system in general. Cardiovascular diseases include conditions such as hypertension, coronary artery disease (the arteries that supply blood and oxygen to the heart), heart valve diseases, angina, arterial blockages, myocardial infarction, among others (Fig. 1). The causes of these diseases are multifaceted and include factors such as smoking, high blood pressure, elevated cholesterol levels, obesity, insufficient physical activity, unhealthy diet, psychological stress, and genetic factors as well. These diseases can be serious and life-threatening, requiring early diagnosis, treatment, and ongoing healthcare to control them and prevent the deterioration of the individual's health condition [1]. Examples of neurodegenerative disorders include:

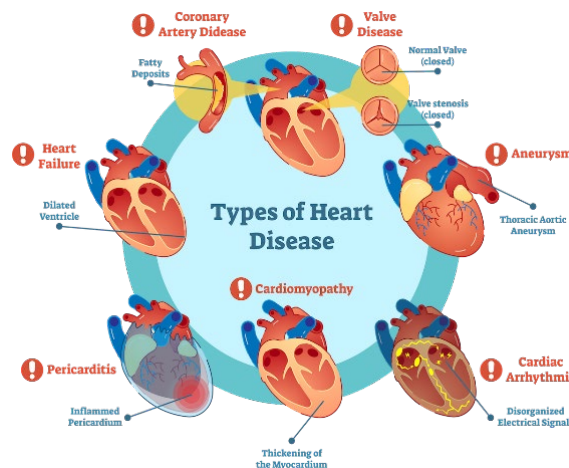


Figure 2: Type of Heart Disease [3]

1.2.1. High blood pressure (arterial hypertension)

High blood pressure, also known as hypertension, is a medical condition characterized by elevated pressure exerted by the blood against the walls of the arteries. This condition occurs when the force of blood pushing against the artery walls is consistently too high, which can lead to serious health complications over time. High blood pressure is typically diagnosed when blood pressure readings consistently measure at or above 130/80 millimeters of mercury (mmHg). It is considered a significant risk factor for various cardiovascular diseases such as heart attack, stroke, and heart failure, as well as kidney disease and other health issues [1], [4].

1.2.2. Hypercholesterolemia

Hypercholesterolemia, also called high cholesterol, High cholesterol levels refer to elevated levels of cholesterol in the blood. Cholesterol is a fatty substance that is naturally produced by the liver and is also found in certain foods. When cholesterol levels become too high, it can lead to a buildup of cholesterol in the walls of the arteries, increasing the risk of cardiovascular diseases such as heart attack and stroke. High cholesterol levels are often associated with lifestyle factors such as poor diet, lack of exercise, smoking, and obesity, but can also be influenced by genetic factors [1], [5].

1.2.3. Coronary artery disease

Coronary artery disease (CAD), also called coronary heart disease (CHD), ischemic heart disease (IHD), myocardial ischemia, or simply heart disease, is a condition characterized by the narrowing or blockage of the coronary arteries, which are the blood vessels that supply oxygen-rich blood to the heart muscle. This restriction in blood flow can lead to various symptoms, such as chest pain (angina), shortness of breath, and in severe cases, heart attack. CAD typically develops over time due to the buildup of plaque, made up of cholesterol, fat, and other substances, on the inner walls of the coronary arteries, a process known as atherosclerosis. Risk factors for CAD include high blood pressure, high cholesterol levels, smoking, diabetes, obesity, and a sedentary lifestyle. Management and treatment of CAD often involve lifestyle changes, medication, and in some cases, procedures such as angioplasty or bypass surgery to restore blood flow to the heart [1], [6].

1.2.4. Cardiac arrest

Cardiac arrest, also known as sudden cardiac arrest, also known as a myocardial infarction (MI), occurs when blood flow to a part of the heart is blocked for a long enough time that part of the heart muscle is damaged or dies. This blockage is often caused by a buildup of plaque, a substance made of fat, cholesterol, and other substances, in the coronary arteries that supply blood to the heart muscle.

Heart attacks can lead to severe chest pain, shortness of breath, nausea, sweating, and other symptoms. Prompt medical treatment is essential to minimize heart damage and improve the chances of survival [1], [7].

1.2.5. Valvar heart disease

Heart valve diseases are conditions that affect the valves of the heart, impairing their function. These valves control the flow of blood within the heart by opening and closing to ensure proper circulation. Heart valve diseases can involve a variety of issues, including valve stenosis (narrowing), valve regurgitation (leaking), or valve prolapse (bulging). These conditions can disrupt blood flow, leading to symptoms such as chest pain, shortness of breath, fatigue, and in severe cases, heart failure. Treatment options for heart valve diseases may include medication, minimally invasive procedures, or surgery to repair or replace the affected valve [1], [8].

1.3. Heart disease screening

Heart disease screening tests are medical examinations and procedures aimed at assessing an individual's risk of heart disease or detecting signs of heart-related problems early on. These tests are crucial for preventive healthcare and can help in identifying risk factors and underlying conditions before they lead to serious complications. Some common heart disease screening tests include:

1.3.1. Heart disease tests

Blood Pressure Measurement: This test assesses the force of blood against the walls of the arteries (Fig. 3). High blood pressure (hypertension) is a significant risk factor for heart disease.

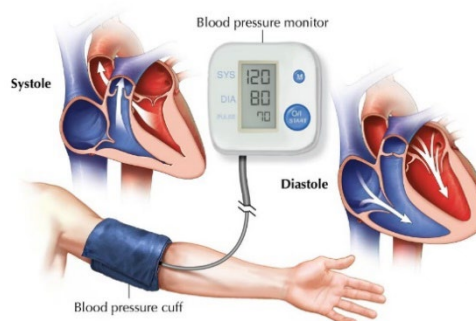


Figure 4: Blood Pressure Measurement [9]

Cholesterol and Lipid Profile: This blood test measures levels of cholesterol and triglycerides in the blood (Fig. 5). High levels of low-density lipoproteins (LDL) cholesterol ("bad" cholesterol) and

low levels of high-density lipoproteins (HDL) cholesterol ("good" cholesterol) can increase the risk of heart disease.

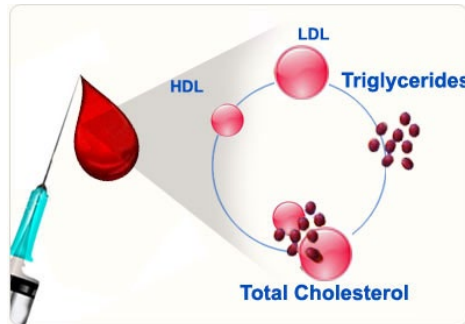


Figure 6: Cholesterol and Lipid Profile [10]

Electrocardiogram (ECG or EKG): An ECG records the electrical activity of the heart and can detect irregular heart rhythms (arrhythmias) and signs of heart damage (Fig. 7).

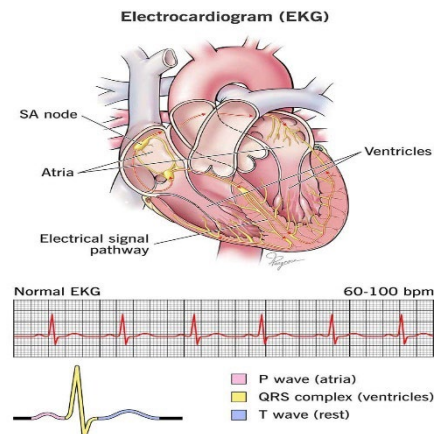


Figure 8: Electrocardiogram [11]

Echocardiogram: This test uses sound waves to create images of the heart's structure and function, allowing doctors to assess the heart valves, chambers, and blood flow (Fig. 9).

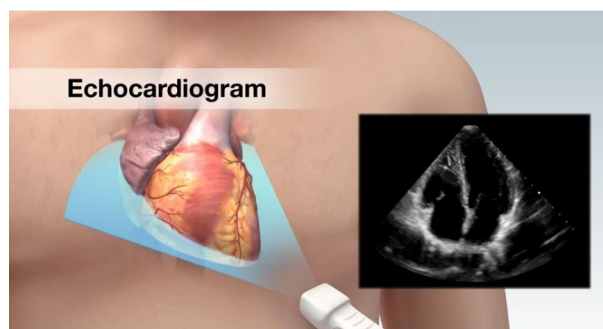


Figure 10: Echocardiogram [12]

✚ **Coronary Calcium Scan:** This imaging test measures the amount of calcium in the walls of the coronary arteries, which can indicate the presence of atherosclerosis (plaque buildup) and the risk of heart attack (Fig. 11).

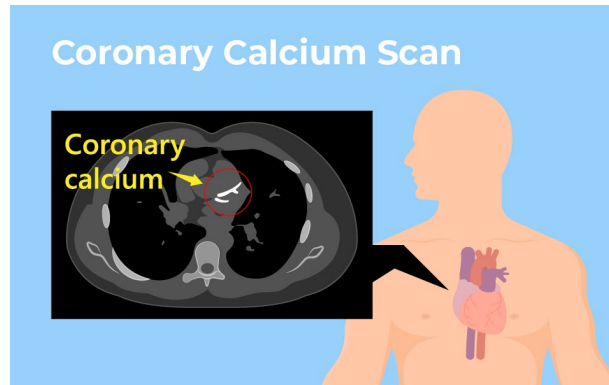


Figure 12: Coronary Calcium Scan [13]

1.4. Artificial intelligence and CVD early detection

Early detection significantly impacts treatment outcomes and prevents potential complications, necessitating tools that facilitate early diagnosis. The integration of artificial intelligence (AI) in early detection is transforming healthcare by leveraging technologies like machine learning and deep learning to analyze large datasets, identify patterns, and enhance diagnostic accuracy and efficiency [14].

The advantages of AI in early detection include rapid data processing, interpretability of results for better decision-making, and improved patient outcomes. AI-driven early detection can reduce healthcare costs and enhance the quality of life for patients [14].

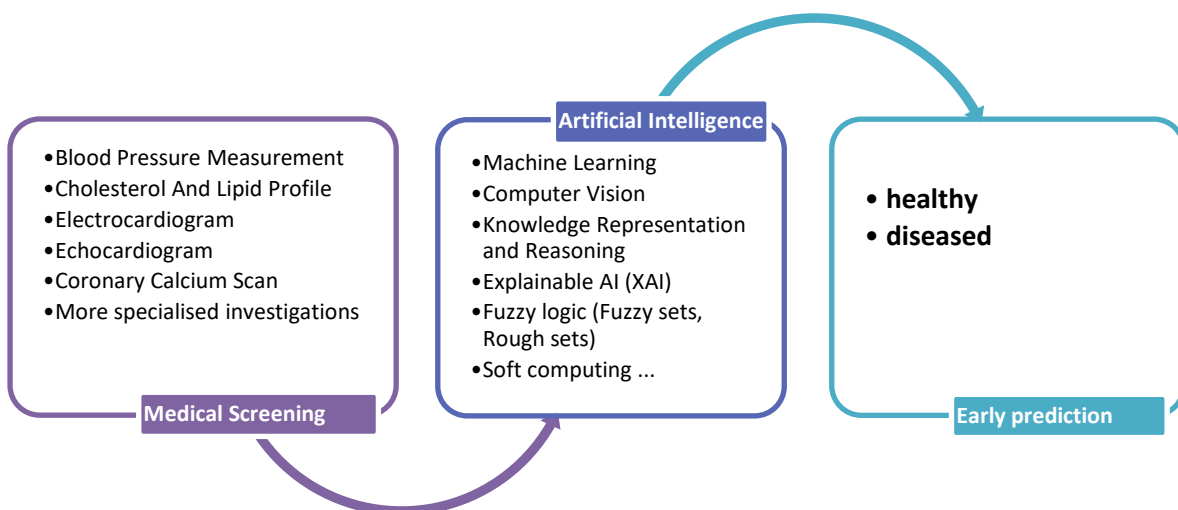


Figure 13: Universal early heart disease detection model

Artificial intelligence (AI) has become a transformative force in healthcare, offering powerful tools for diagnosis, treatment, and disease prevention. However, a crucial element for successful AI integration in clinical settings, particularly for sensitive issues like early disease detection, is interpretability. While complex AI models can achieve impressive accuracy, their "black box" nature, where the reasoning behind predictions remains obscure, can hinder trust and limit their real-world application. Interpretable machine learning models, on the other hand, provide transparency by revealing how they arrive at their conclusions. This is especially important in cardiology, where early detection of cardiovascular disease (CVD) is critical for improving patient outcomes. By understanding the factors influencing the model's predictions, physicians can gain valuable insights into a patient's risk profile, leading to more informed decisions about preventive measures and early intervention strategies. This interpretability fosters trust in the AI system and empowers clinicians to leverage its capabilities alongside their own expertise for better patient care.

A particularly promising AI approach is the use of rough sets, a technique that excels in handling uncertainty and providing interpretable results. Although not widely used, rough sets are effective in early detection, especially for diseases such as cardiovascular conditions.

After mentioning at this stage, the importance of interpretable machine learning in clinical settings, particularly for early CVD detection. To further solidify this concept, the following chapters will delve deeper into the technical aspects of rough sets and interpretable models and their application in healthcare. In the rest of this chapter, however, we'll shift gears and examine the existing research landscape. We'll explore related works in this field, gaining valuable insights from past studies before diving into the specific's tools and materials for interpretable models for CVD detection.

1.5. Related works

In recent years, the utilization of machine learning and deep neural networks across diverse clinical domains has experienced a surge in popularity. Particularly noteworthy is the growing research focus on developing automated systems aimed at early detection of heart disease through analysis of laboratory data and electrical heart signals. This section aims to provide a comprehensive overview of prior research endeavors dedicated to the discovery of heart disease, encompassing the utilized datasets, methodologies employed, and resulting outcomes. Notably, this section serves as a chronological exploration of these efforts, offering an evolutionary narrative from earlier investigations to the most recent advancements.

✎ Heart Diseases Detection Using Naive Bayes Algorithm (2015)

This article [15] discusses the application of **data mining** in healthcare, particularly in diagnosing heart diseases. It highlights the significance of extracting hidden information from healthcare data and proposes a system using the **Naïve Bayes algorithm** to classify data and predict heart conditions. The Naïve Bayes algorithm is shown to achieve high **accuracy (86.4198%)** with **minimal processing time**, making it a valuable tool for healthcare professionals in diagnosing heart diseases.

✎ ANN Parameter Tuning Framework for Heart Disease Classification (2018)

This paper [16] introduces a method to enhance the detection of heart disease by utilizing **artificial neural networks (ANNs)** as decision support tools. ANNs have shown promise in this area, but their effectiveness depends heavily on various parameters being set optimally. To address this challenge, the paper proposes a framework for tuning these parameters effectively.

The framework is tested using two datasets: **the Statlog heart disease dataset** and **the Cleveland heart disease dataset**. These datasets are commonly used benchmarks in heart disease research. The results demonstrate the effectiveness of the proposed framework, with impressive classification accuracy rates achieved. Specifically, **the overall classification accuracy reaches 90.9% for the Cleveland dataset and 90% for the Statlog dataset**.

To summarize, the paper presents a novel approach to optimizing the performance of ANNs for heart disease detection. By systematically tuning the parameters of the ANN, the proposed framework achieves high levels of accuracy in classifying heart disease cases.

✎ Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms (2018)

This research paper [17] explores unconventional approaches to identifying significant factors related to heart disease. The authors focused on analyzing **two datasets (Cleveland & Hungarian)**, each divided into different percentages of data. They determined the values of various individual attributes within these datasets to identify relevant factors associated with heart disease. Subsequently, they employed different **semi-supervised learning algorithms**, including Collective Wrapper, Filtered Collective, and Yet Another Semi Supervised Idea, to analyze the heart disease data. Various metrics such as accuracy, f-measure, and area under the receiver operating characteristic curve (ROC) were utilized to evaluate the performance of these classifiers and determine the most effective semi-supervised learning algorithm.

The selected algorithm effectively identified both significant and irrelevant factors contributing to heart disease by systematically removing attributes and observing the resulting classification outcomes. Experimental results on real datasets demonstrated the effectiveness and efficiency of their analysis, with the best achieved result being an **87.2449% accuracy rate**.

✎ **Prediction of Heart Disease Using Multi-Layer Perceptron Neural Network and SVM (2019)**

This study [18] aimed to address the increasing concern over heart disease-related mortality by proposing and comparing two classifiers: a **Multi-Layer Perceptron neural network (MLP)** and a **Support Vector Machine (SVM)**. The objective was to accurately diagnose heart disease by classifying it into **two classes and five classes**.

The researchers utilized **the Cleveland heart disease online dataset**, which comprises **303 instances** categorized into **five classes** and described by **13 attributes**. For the two-class classification problem, **SVM achieved an accuracy of 92.45%**, while **MLP attained 90.57% accuracy**. In the case of the five-class classification problem, **MLP exhibited an accuracy of 68.86%**, whereas **SVM achieved 59.01% accuracy**.

These results suggest that both SVM and MLP hold promise for heart disease classification tasks, with SVM performing slightly better for the two-class problem and MLP outperforming SVM for the five-class problem.

✎ **On Machine Learning Models for Heart Disease Diagnosis (2020)**

This article [19] explores the application of **Convolutional Neural Networks (CNNs)** and regular **Neural Networks (NNs)** in diagnosing heart disease. The authors conducted experiments comparing the performance of these two machine learning models by implementing the algorithms, tuning parameters, and analyzing results. They utilized **the Cleveland datasets** from the UCI learning dataset repository for this purpose.

The methodology involved training both **CNNs and NNs on the dataset**, adjusting parameters to optimize performance, and evaluating prediction accuracy. Despite CNNs' distinctive architecture, the experimental findings revealed that NNs consistently outperformed CNNs in terms of prediction accuracy across various parameter settings.

The research culminated in achieving a notable prediction **accuracy of 92.81%** when utilizing NNs. These results suggest the superiority of NNs over CNNs for heart disease diagnosis based on the conducted experiments.

✎ Heart Disease Prediction using Hybrid machine Learning Model (2021)

The article [20] discusses the pressing issue of heart disease and emphasizes the importance of early prediction to mitigate its mortality rate. Traditional clinical data analysis faces challenges in detecting cardiovascular diseases promptly. Hence, the study proposes leveraging **machine learning (ML) techniques** to enhance decision-making and prediction accuracy. The research focuses on utilizing **the Cleveland heart disease dataset** and employs **data mining techniques** such as regression and classification. Specifically, it applies two ML techniques, **Random Forest and Decision Tree**, while also introducing a novel hybrid model that combines both.

During implementation, three machine learning algorithms are utilized: Random Forest, Decision Tree, and the Hybrid model. The Hybrid model, which integrates Random Forest and Decision Tree, is particularly highlighted. Experimental results indicate an impressive **accuracy level of 88.7%** in predicting heart disease using the hybrid model. Additionally, the study includes the development of a user-friendly interface to input parameters for heart disease prediction, utilizing the hybrid model of Decision Tree and Random Forest.

✎ Prediction of Heart Disease using Dense Neural Network (2022)

In this paper [21], the authors investigated the prediction of heart disease, a common and life-threatening condition if untreated. They proposed an **Artificial Neural Network (ANN)** based method for early detection of heart disease. Using the publicly available **UCI Heart Disease dataset**, which includes various health indicators like age, sex, blood pressure, and cholesterol levels, their ANN model was trained and tested. The proposed method achieved a high **accuracy of 96.09%**, demonstrating its potential for early diagnosis and timely medical intervention.

✎ Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers (2022)

In this paper [22], the authors address the global issue of heart disease-related complications, which cause thousands of deaths each year. To improve survival rates, early and accurate diagnosis is crucial. The authors proposed an ensemble **machine learning model, Stacking CV Classifier**, which combines Logistic Regression, **K Nearest Neighbors, and Naïve Bayes classifiers**. Using the preprocessed **Cleveland Clinic Foundation (CVF) dataset**, the ensemble model achieved a prediction **accuracy of 90.0%**, outperforming the individual base models with accuracies of **86.66%, 88.33%, and 86.66%**. This efficient and noninvasive diagnostic method can significantly aid healthcare practitioners, particularly in underdeveloped areas.

✎ A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks (2023)

In this paper [23], the authors address Heart Disease (HD), the leading cause of global mortality according to the WHO. They propose a **Hybrid Deep Neural Network (HDNN)** system combining **CNN, LSTM, and Dense layers** for improved HD prediction. The system was evaluated on **the Cleveland HD dataset** and a combined dataset from multiple sources, achieving a high accuracy of **98.86%**. This approach outperformed previous methods and shows great potential for enhancing medical diagnosis and patient care in healthcare systems.

✎ Heart Disease Prediction using Supervised Learning Classifiers (2023)

This paper [24] aims to predict heart disease by comparing various **machine learning techniques**. Using the **Cleveland dataset**, four classifiers—**Support Vector, Random Forest, Multi-Layer Perceptron, and Ridge**—were employed. The Ridge Classifier demonstrated the highest accuracy of **89.19%** and an F1-score of 0.92. In conclusion, the Ridge Classifier proves to be a reliable method for early detection of heart disease, potentially preventing premature deaths.

✎ A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning (2023)

The research [25] discussed in this summary focuses on addressing the growing challenge of heart disease detection, which is currently the leading cause of mortality globally. Despite the availability of vast amounts of heart disease data in healthcare settings, the intelligent utilization of this data to identify underlying patterns remains a challenge. To tackle this issue, the researchers employ **machine learning techniques**, particularly **deep learning**, to develop decision support systems (DSS) capable of learning from past experiences and improving accuracy in diagnosing heart illnesses.

The core methodology involves using a **Keras-based deep learning** model to construct a dense neural network for precise diagnosis. The model is tested with various configurations of hidden layers, ranging from 3 to 9 layers, each containing 100 neurons and utilizing the Relu activation function. **Multiple heart disease datasets** are utilized as benchmarks for analysis, encompassing both individual and ensemble models. Evaluation metrics such as sensitivity, specificity, accuracy, and f-measure are employed to assess the performance of the dense neural network across all datasets.

The results of extensive experimentation indicate that the proposed deep learning model outperforms individual models and alternative ensemble approaches in terms of accuracy, sensitivity, and specificity across all heart disease datasets. The highest recorded **accuracy achieved is 83%**.

1.6.State of the art summary

For illustration purposes, the following figure briefly summarizes all the works mentioned in this related works section:

	Study	Year	Approach	Dataset	Accuracy
1	Jabbar et al. [15]	2015	Naïve Bayes	Collected from 500 patients	86.41%
2	Haikal et al. [16]	2018	ANNs	The Statlog	90%
				The Cleveland (2 class)	90%
3	Satu et al. [17]	2018	semi-supervised learning	Cleveland & Hungarian	87.24%
4	Nayeem et al. [18]	2019	SVM	The Cleveland (5 class)	59.01%
				The Cleveland (2 class)	92.45%
			MLP	The Cleveland (5 class)	68.86%,
				The Cleveland (2 class)	90.57%
5	Lin et al. [19]	2020	CNN	The Cleveland (2 class)	92.81%
6	Dinesh et al. [20]	2021	Random Forest + Decision Tree	The Cleveland (2 class)	88.7%
7	A.Singh et al [21]	2022	ANN	The Cleveland (2 class)	96.09%
8	N. Itoo et al [22]	2022	K-Nearest + Naiv-Bayes	The Cleveland (2 class)	90.00%
9	Reshan et al [23]	2023	Hybrid Deep Neural Network	The Cleveland (2 class)	98.86%
10	Rahman et al [24]	2023	Machine learning	The Cleveland (2 class)	89.19%
11	Bashir et al. [25]	2023	CNN	Multiple datasets (2 class)	83%

The reviewed research on cardiovascular disease (CVD) diagnosis through machine learning paints a picture of promise and ongoing development. Studies like Jabbar et al. (2015) with Naive Bayes and Haikal et al. (2018) with Artificial Neural Networks (ANNs) achieved respectable accuracy (86.41% and 90%, respectively). However, these results are surpassed by more recent studies employing deeper learning techniques. Lin et al. (2020) with Convolutional Neural Networks (CNNs) and Reshan et al. (2023) with Hybrid Deep Neural Networks achieved impressive accuracy exceeding

92% on the Cleveland dataset. This trend suggests that advancements in deep learning architectures are pushing the boundaries of performance.

However, accuracy is not the only consideration. Studies like Nayeem *et al.* (2019) employing Support Vector Machines (SVMs) and Multilayer Perceptrons (MLPs) demonstrate significant performance variations between multi-class (5 class) and binary (2 class) classification on the Cleveland dataset. This highlights the importance of considering the chosen dataset and classification task when evaluating performance metrics.

Furthermore, some studies like Dinesh *et al.* (2021) with Random Forests and Decision Trees, and Bashir *et al.* (2023) with CNNs utilizing multiple datasets, achieved good accuracy but might lack interpretability compared to simpler models like Naive Bayes. Balancing interpretability with performance remains a crucial challenge.

In conclusion, the reviewed research underscores the promise of machine learning for CVD diagnosis. Deep learning architectures like CNNs and Hybrid Deep Neural Networks demonstrate impressive accuracy, pushing the boundaries of performance (Lin *et al.*, 2020; Reshan *et al.*, 2023). However, their suitability for the Cleveland dataset warrants further investigation. Deep learning often relies heavily on data augmentation techniques to improve performance, which raises concerns about the generalizability of models trained on such artificially inflated datasets.

Furthermore, excessively aggressive data pre-processing can strip away valuable information from the original data, particularly when applied to the testing set. This can lead to models that perform well on pre-processed data but struggle with real-world variations. A balance needs to be struck between data normalization and preserving the inherent characteristics of the data, especially for testing purposes.

Therefore, while the potential of deep learning is undeniable, careful consideration of the chosen dataset, classification task, interpretability, and the impact of data pre-processing techniques are crucial for selecting the most effective and generalizable approach for real-world CVD diagnosis.

1.7. Conclusion

In this chapter, we provided a brief introduction to cardiovascular disease and various concepts and approaches crucial for understanding this work. We emphasized the importance of understanding the cardiovascular disease itself, including its symptoms, potential patient experiences, and relevant diagnostic tests. Next, we mentioned the importance of interpretable machine learning in clinical settings, particularly for early CVD detection, we also acknowledged the importance of existence of various datasets and data processing techniques that can be leveraged for building such models.

Finally, we recognized the value of examining past research efforts in this domain. By reviewing existing studies ranging from 2015 up to the year of writing this dissertation 2024, what datasets they used, what methods they've employed, and what results they've achieved. This section also showed the evolutionary history of Heart Disease detection systems through the automatic analysis of data, we concluded by compiling every paper cited in this chapter into a single summary table for illustration purposes. In the next chapter, we will be explaining our own approach, from the training method to the datasets used, and the way we preprocessed and augmented said data, finally we will present the models used.

The following chapters will introduce more relevant technics and materials in order to delve deeper into the specifics of our proposed interpretable machine learning model for CVD risk assessment. We will explore the chosen model architecture, its training process, and the evaluation methods employed to assess its effectiveness.

Chapter 02: ML and Rough Sets for Interpretable Diagnosis

2.1. Introduction

In recent years, the utilization of machine learning and deep neural networks across diverse clinical domains has experienced a surge in popularity. Particularly noteworthy is the growing research focus on developing automated systems aimed at early detection of heart disease through analysis of laboratory data and electrical heart signals. This chapter aims to understand machine learning and certain concepts relating to it, in addition to the pre-existing models and how they can be used. Either for making predictions or to extract features that are then fed to a classifier, we also discussed the rough set and its details and how it can be used to study and analyze the data set. We also studied its steps in detail. So, we will dedicate the second part of this chapter to introducing rough sets as the best alternative to address the urgent need to provide a model for predicting cardiovascular diseases. Rough sets offer distinctive capabilities in creating understandable and recognized efficient rules for prediction based on the processing of patient data.

2.2. Machine learning

Machine learning (ML) is a field of artificial intelligence that enables computer systems to learn from data without being explicitly programmed. It relies on algorithms that can detect patterns and make decisions based on these patterns [26].

Machine learning is widely used in the health system around the world to improve the accuracy and performance of disease identification. Here are some of the basic concepts and techniques used in machine learning for disease identification:

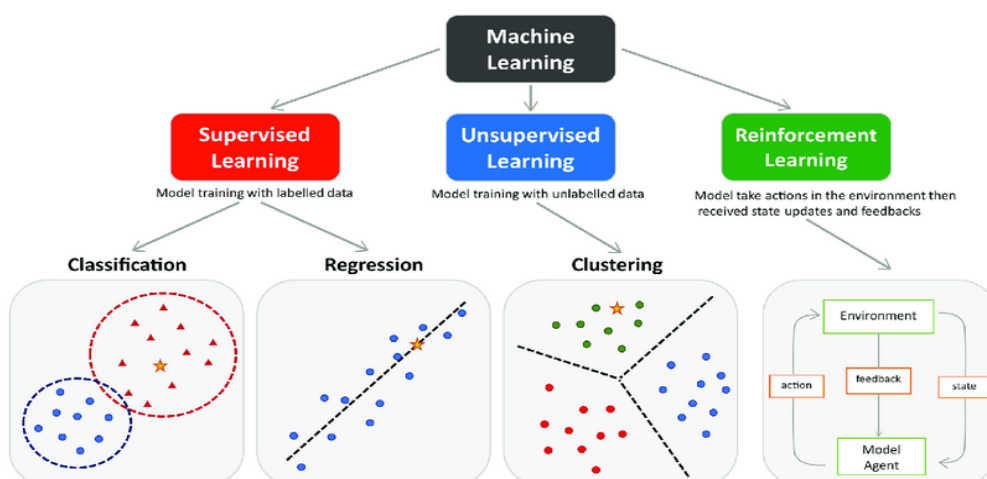


Figure 14: Machine Learning [27]

2.2.1. Supervised learning

Supervised learning is a machine learning technique where a model is trained using a labeled dataset, where each data example is associated with a predefined label or class. The model learns from these labeled data to be able to predict or classify new unlabeled examples [28].

- K-Nearest Neighbors (KNN).
- Bayesian classification.
- Support Vector Machine (SVM).
- Artificial Neural Networks (ANN).
- Decision tree.

2.2.2. Unsupervised learning

Unsupervised learning is a branch of machine learning where a model is trained using an unlabeled dataset, meaning there is no information about the classes or labels of the data examples. The goal of unsupervised learning is to discover intrinsic structures, patterns, or clusters within the data [28].

- Artificial Neural Networks "ANN"
- K-means "KMeans"
- Fuzzy K-Means
- Hierarchical clustering

2.2.3. Reinforcement learning

Reinforcement learning is a machine learning method where an agent learns to make decisions by interacting with an environment and receiving rewards or punishments based on its actions. The agent learns to maximize rewards over time by adjusting its action policy.

2.2.4. Classification

The automated disease identification system culminates in the classification stage, which is the final step. Several machine learning algorithms are commonly employed for disease identification, including:

- Support Vector Machines (SVM)
- Hidden Markov Models (HMM)
- Linear Discriminant Analysis (LDA)
- K Nearest Neighbors (KNN)
- Bayesian Networks (BN)

- Neural Networks (NN)

The disease identification process is grounded in machine learning principles. A feature vector is constructed to characterize and diagnose the disease, with the initial phase of the classifier constituting the learning process. Training the classifier involves assigning labels to the detected datasets. Once trained, the classifier can identify inputs by assigning specific data class labels to them.

Classification methods can be categorized into two groups:

- Static data identification, comprising solely of data.
- Dynamic data recognition, including prediction based on devices directly, where results are obtained from specialized devices.

This work concentrates specifically on disease recognition and identification systems reliant on static data, particularly analytics.

2.3. Common issues in ML

Machine learning, despite its remarkable capabilities, is not without its challenges. Here are some of the most common issues encountered in the field:

🚩 Data issues:

- **Poor quality data:** Machine learning models are heavily reliant on data quality. Inaccurate, noisy, or incomplete data can lead to biased or inaccurate predictions.
- **Inadequate data:** Insufficient data can hinder a model's ability to learn effectively, especially for complex problems.
- **Non-representative data:** If the training data doesn't accurately reflect the real-world scenario the model will be used in, it might perform poorly when deployed.

🚩 Model issues:

- **Underfitting:** When a model fails to capture the underlying patterns in the training data, it performs poorly on both the training data and unseen data (generalization). This can happen due to an overly simplistic model or not training for a long enough duration.
- **Overfitting:** If a model memorizes the training data too closely, it might not be able to generalize well to unseen data. This often occurs with complex models trained on limited data.

Overfitting is a central problem in supervised machine learning (learning from labeled training data). It is observed when a model ends up over-learning and memorizing undesired aspects of the data thus preventing itself from generalizing the models to well fit not just the training data, but also unseen data contained in the validation/testing set. Not to be confused with Underfitting, a similar problem that occurs when a model is under-trained. Overfitting usually occurs due to; poorly pre-processed data containing too much noise, a small training set, and/or the use of overly complex model architecture [29].

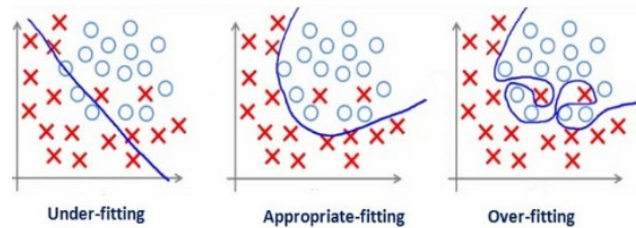


Figure 15: Overfitting [29]

Over the years, researchers have discovered various strategies to reduce the effects of overfitting:

- ☞ *Data-augmentation*: by generating new data points from existing data, helping to prevent overfitting. [30].
 - ☞ *Callbacks*: perform certain actions at various stages of training, e.g; at the start or end of an epoch.
 - ☞ *Regularization*: To mitigate overfitting caused by the model's sensitivity to numerous features, two strategies are proposed: feature selection, involving the removal of irrelevant features through manual or automatic preprocessing, and weight minimization for less impactful features, enhancing final classification accuracy. [29].
 - ☞ Regularization techniques can help prevent overfitting.
- **Selection bias**: If the data used to train the model is not chosen carefully and introduces a bias, the model's predictions will reflect that bias.

🌟 Interpretability Issues:

- **Black box models**: Many complex models, particularly deep learning models, are difficult to understand how they arrive at their predictions. This lack of interpretability can be problematic for applications requiring trust and transparency.

- **Explainability vs. Accuracy:** There can be a trade-off between achieving high accuracy and maintaining interpretability. Sometimes, the most accurate models might be the most difficult to understand.

✚ Other challenges:

- **Computational cost:** Training complex models can require significant computational resources and time.
- **Ethical considerations:** Issues like bias, fairness, and privacy need careful consideration when developing and deploying machine learning models.

By being aware of these common issues, machine learning practitioners can take steps to mitigate them and develop robust, reliable models. This can involve data cleaning and pre-processing techniques, choosing appropriate model architectures, employing regularization to prevent overfitting, and leveraging interpretability techniques like feature importance or SHAP values (SHapley Additive exPlanations).

Despite its potential, machine learning faces obstacles. Issues like biased or insufficient data can lead to inaccurate predictions. Complex models might lack interpretability, hindering trust and hindering our understanding of how they reach conclusions. To address these challenges, we can explore alternative approaches. Rough sets theory emerges as a promising candidate. By generating understandable rules and identifying key attributes, rough sets offer a level of interpretability often absent in complex models. This can be particularly valuable in situations demanding transparency and explainability in the decision-making process. Let's delve deeper into rough set theory and how it tackles some of the common roadblocks encountered in machine learning.

2.4. The rough sets approach

A rough set is a mathematical framework within artificial intelligence that handles uncertainty and vagueness in data analysis. Introduced by Zdzisław Pawlak, it deals with imperfect knowledge by approximating the boundaries between sets based on indiscernibility, where objects share some attributes but not others. Rough set theory identifies *lower* and *upper* approximations to classify objects: the lower approximation includes objects definitively belonging to a set, and the upper approximation encompasses objects possibly belonging to the set (Fig. 16). The boundary region of a set with respect to an indiscernibility relation is the set of all objects which can be classified neither as

belonging nor not belonging to the set. A set is considered crisp (exact) if its boundary region is empty, and rough (inexact) if its boundary region is nonempty [31].

Rough set theory is embedded in classical set theory and can be viewed as a specific implementation of Frege's² idea of vagueness, where imprecision is expressed by the boundary region of a set rather than by partial membership, as in fuzzy set theory. It can be defined through topological operations, namely interior and closure, called approximations. The theory finds applications in AI fields like data mining, machine learning, and decision analysis, providing a flexible and interpretable approach to dealing with incomplete information. The indiscernibility relation, often assumed to be an equivalence relation, describes our lack of knowledge about the universe and forces us to reason about accessible granules of knowledge rather than individual objects. [31]

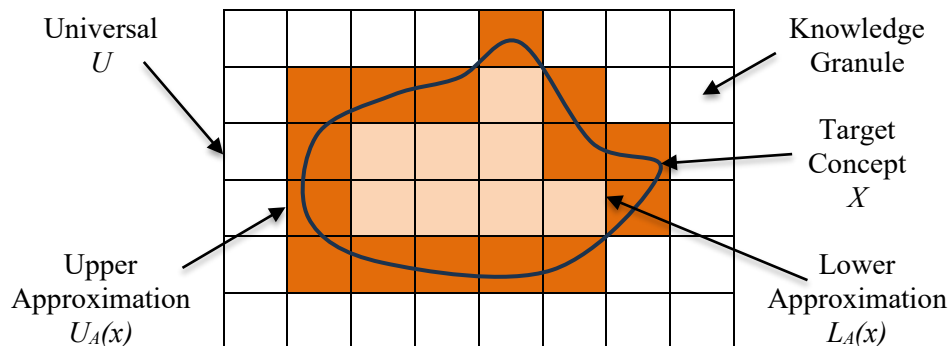


Figure 17: diagram of rough set [31]

2.4.1. Basic problems in data analysis solved by Rough Set:

Rough Set theory addresses several fundamental challenges in data analysis [31]:

- ✓ **Characterization of object sets:** Rough Set theory helps characterize sets of objects based on their attribute values. By analyzing these attributes, it identifies patterns and relationships within the data.
- ✓ **Attribute dependency:** It facilitates the identification of dependencies between attributes. This analysis reveals how different attributes relate to each other, providing insights into the underlying structure of the data.
- ✓ **Attribute reduction:** One of the key functions of Rough Set theory is to reduce redundant or superfluous attributes. By eliminating irrelevant attributes, it simplifies the data representation while preserving essential information.

² Frege, a prominent philosopher and logician, didn't explicitly discuss "vagueness" in the same way we do today. However, his work on concepts and predicates sheds light on how he might have viewed concepts that lack sharp boundaries.

- ✓ **Identification of significant attributes:** Rough Set theory aids in identifying the most significant attributes within a dataset. These attributes contribute the most to the understanding and prediction of the data.
- ✓ **Decision rule generation:** Lastly, Rough Set theory enables the generation of decision rules based on the analyzed data. These rules provide a framework for decision-making, helping to guide actions and strategies based on the insights gained from the data analysis.

2.4.2. Goals of Rough Set Theory:

Rough Set Theory aims to achieve several key objectives [31]:

- ☞ **Induction of concept approximations:** The primary goal of rough set analysis is to induce approximations of concepts, aiding in the understanding and interpretation of complex data patterns.
- ☞ **Foundation for Knowledge Discovery in Databases (KDD):** Rough sets provide a solid foundation for Knowledge Discovery in Databases (KDD). They offer mathematical tools and techniques to uncover hidden patterns within datasets, facilitating the extraction of valuable knowledge.
- ☞ **Applications in data analysis:** Rough set theory can be applied to various aspects of data analysis, including feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction such as templates and association rules.
- ☞ **Identification of dependencies:** It helps identify partial or total dependencies within datasets. By analyzing these dependencies, redundant data can be eliminated, leading to more efficient data representation and analysis.
- ☞ **Handling data challenges:** Rough set theory provides approaches to address challenges such as null values, missing data, dynamic data, and others. This enhances the robustness and applicability of data analysis techniques based on rough sets.

2.4.3. Attributes in Rough Sets:

In the context of rough sets theory, attributes are defined as the characteristics representing the available data for a given object, which can be used to distinguish and classify it within a certain set. These attributes are assigned during the onset of analysis and classification processes, where they are extracted from the available data and determined according to the specific purpose of the study or project [31].

Assigning attributes requires a deep understanding of the studied domain and the desired goals of the analysis. Attributes are carefully selected based on the available data and the analysis required,

identifying features believed to be important for understanding phenomena or determining relationships among objects in the set.

After attributes are assigned, they are utilized in the process of data analysis and object classification. Attributes are used to discover relationships and patterns, and to determine rules that can be employed to classify objects into different sets. With the precise determination and effective use of attributes, rough set theory can provide reliable results and valuable information for decision-making and comprehensive data analysis [31].

2.4.4. Rule extraction:

Rule extraction is the process of deriving logical rules or patterns from data that describe relationships, dependencies, or regularities within the dataset. In the context of rough set theory, rule extraction involves identifying decision rules that accurately classify objects into different categories or classes based on their attributes [31].

The process of rule extraction typically involves the following steps:

- ☞ **Attribute selection:** Identifying relevant attributes or features from the dataset that are important for classification or prediction.
- ☞ **Rule generation:** Generating candidate rules based on the selected attributes and their values. These rules are typically expressed in the form of IF-THEN statements, where the antecedent (IF) specifies the conditions or criteria for classification, and the consequent (THEN) specifies the predicted outcome or class label.
- ☞ **Rule evaluation:** Assessing the quality and effectiveness of the generated rules using evaluation metrics such as accuracy, coverage, and interpretability. This step may involve pruning or refining the rules to improve their performance and generalization ability.
- ☞ **Rule refinement:** Refining the extracted rules to enhance their interpretability and generalization ability. This may involve simplifying complex rules, removing redundant or irrelevant conditions, and optimizing rule structures.
- ☞ **Rule interpretation:** Interpreting the extracted rules to gain insights into the underlying patterns or relationships within the data. This step helps in understanding the decision-making process and providing actionable insights for decision-makers.

Overall, rule extraction plays a crucial role in data analysis and knowledge discovery, enabling the extraction of actionable knowledge from complex datasets and facilitating informed decision-making in various domains.

Consider a dataset with information about customers of a company:

Customer ID	Age	Monthly Income	Years with Company	Satisfaction Level
1	25	\$3000	2	High
2	35	\$5000	5	Low
...				

Table 1: Rough set's example

Rough Set Analysis:

Convert Data to Rules:

We'll create rules based on attributes and their impact on customer retention:

- If Age is less than 30 and Satisfaction Level is High, the customer will likely stay.
- If Monthly Income is high and Years with Company is low, the customer might leave.

Convert Data to Rules:

- ➔ Rule 1: *If Age < 30 and Satisfaction Level is High, then Stay.*
- ➔ Rule 2: *If Monthly Income is high and Years with Company is low, then Leave.*

Analyzing Rules:

After creating rules (automatically using the rough sets approach), we analyze them to understand the key factors affecting customer retention. Rough Set analysis might reveal that Age and Satisfaction Level are the primary factors influencing customer decisions.

Practical Application:

Based on Rough Set analysis, the company can improve customer retention strategies by focusing on enhancing customer experience and increasing satisfaction, especially among younger customers.

- For customer 1: *Age < 30 and Satisfaction Level is High, so they are likely to stay.*
- For customer 2: *Monthly Income is high, but Years with Company is high too, so they might leave.*

2.4.5. Upper Approximation:

The upper approximation in rough set theory represents the set of objects that are part of a given concept. In the context of customer retention [31]:

Upper Approximation (Stay): Customers who stay with the company based on the provided rules.

Let U_{stay} be the upper approximation for customers who stay with the company.

$$U_{stay} = \{x \in X \mid Condition_1(x)\} \quad \text{Equation 1: Upper Approximation}$$

Where:

X is the set of all customers.

$Condition_1(x)$ represents the condition under which a customer definitely stays.

In our example:

- Customer 1 (Age < 30 and Satisfaction Level is High): Stays.
- Customer 3 (Age \geq 30 and Satisfaction Level is Medium): Potentially stays (not definitely).

So, the upper approximation for "Stay" includes customers 1 and 3.

$$U_{stay} = \{Customer_1\}$$

2.4.6. Lower Approximation:

The lower approximation in rough set theory represents the set of objects that are possibly part of a given concept. [31]

Lower Approximation (Stay): Customers who possibly stay with the company based on the provided rules.

Let L_{stay} be the lower approximation for customers who stay with the company.

$$L_{stay} = \{x \in X \mid Condition_2(x)\} \quad \text{Equation 2: Lower Approximation}$$

Where:

X is the set of all customers.

$Condition_2(x)$ represents the condition under which a customer possibly stays.

In our example:

- Customer 1 (Age < 30 and Satisfaction Level is High): Definitely stays.
- Customer 3 (Age \geq 30 and Satisfaction Level is Medium): Possibly stays.
- Customer 4 (Age < 30 and Satisfaction Level is High): Possibly stays.

So, the lower approximation for "Stay" includes customers 1, 3, and 4.

$$L_{stay} = \{Customer_1, Customer_3, Customer_4\} \quad \text{Equation 3: Lower Approximation example}$$

2.5. Conclusion

In conclusion, this chapter aimed to provide an understanding of machine learning and related concepts, as well as existing models and their potential uses, whether for making predictions or extracting features to feed into a classifier. We also discussed the rough set theory, its details, and how it can be applied to study and analyze datasets. By examining these steps in detail, we have established a solid foundation for appreciating the evolution and future potential of these technologies in the field of heart disease detection.

Chapter 03: **The proposed methodology**

3.1. Introduction

In this chapter, we will detail the various means employed in this study, ranging from data processing to training steps and the application of rough set theory. Additionally, we will offer insights into the datasets utilized, along with the model generation processes techniques experimented with. Finally, we will elucidate the RSES2.2 application employed as a tool for rough set analysis and outline how we intend to leverage it to derive the results.

3.2. Training method

A 75%-25% split ratio was adopted for training-testing data partitioning across all experiments. The Cleveland dataset is stored in a simple, tabular format resembling a CSV file. However, it deviates from the standard by utilizing the *.dat* extension. This dataset offers a header section that describes the attributes (numerical/categorical) representing diverse patient features (e.g., age, gender, blood pressure) and a binary classification (decision variable) indicating the presence (1) or absence (0) of heart disease. It also provides the total number of instances (tuples) within the dataset. Missing values within the training set will be handled through robust imputation techniques, such as mean imputation, selected based on statistical soundness. The results of the study will be evaluated using several performance metrics, including the following:

- 👉 **Accuracy:** The overall percentage of correctly classified instances.
- 👉 **Coverage:** The proportion of testing instances covered by the generated rules.
- 👉 **True positives for each class:** The percentage of correctly classified instances with and without heart disease.
- 👉 **Total accuracy and coverage:** Overall model performance on the testing set.

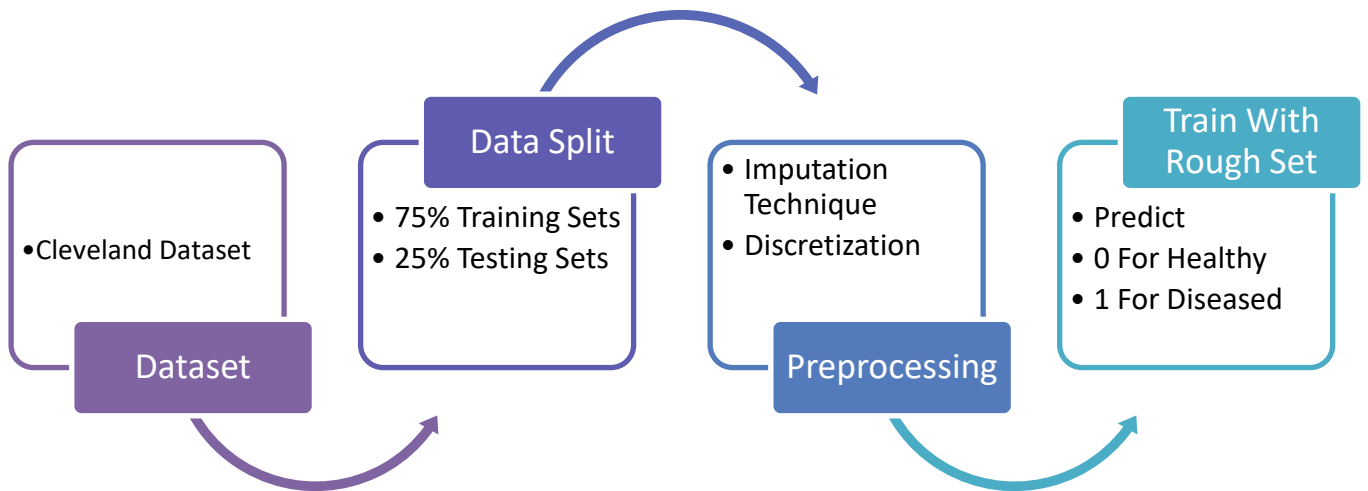


Figure 18: Training Method

3.3. The Cleveland Heart Disease Dataset

The Cleveland Heart Disease Dataset [32] comprises one of the four most utilized databases focused on heart disease diagnosis, collected from various institutions. Each database follows the same instance format, encompassing 76 raw attributes, though only 14 attributes have been utilized in past experiments. Notably, the Cleveland database is the primary focus of machine learning (ML) researchers. The "goal" attribute in this dataset denotes the presence of heart disease in patients, ranging from 0 (no presence) to 1. Notably, experiments have predominantly concentrated on distinguishing between the presence (value 1) and absence (value 0) of heart disease. The dataset includes information such as age, gender, chest pain type, resting blood pressure, serum cholesterol levels, and other medical indicators. Additionally, it has undergone processing to remove patient identifiers. It consists of 303 instances with 76 attributes, including the predicted attribute.

Heart Disease (Cleveland) Data Set					
1	Type	Classification	2	(Real/Integer/Nominal)	(13/0/0)
2	Features	13	6	Missing Values?	Yes
3	Classes	2	7	Total Instances	303
4	Origin	Real World	8	Instances without Missing Values	297

Figure 19: Description of the Dataset [32]

Number of Attributes: 76 (including the predicted attribute), there is only 14 attributes used:

Attributes	Description
Age	Age in years
Sex	Sex (0 = female; 1 = male)
Cp	Chest pain type: <ul style="list-style-type: none"> • Value 1: typical angina • Value 2: atypical angina • Value 3: non-anginal pain • Value 4: asymptomatic
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
Chol	Serum cholesterol in mg/dl
Fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
Restecg	resting electrocardiographic results <ul style="list-style-type: none"> • Value 0: normal • Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) • Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Thalach	maximum heart rate achieved
Exang	exercise induced angina (1 = yes; 0 = no)
Oldpeak	depression induced by exercise relative to rest
Slope	slope of the peak exercise ST segment <ul style="list-style-type: none"> • Value 1: upsloping • Value 2: flat • Value 3: downsloping
Ca	number of major vessels (0-3) colored by flourosopy
Thal	<ul style="list-style-type: none"> • 3 = normal • 6 = fixed defect • 7 = reversable defect
Num	The Predicted Attribute, Diagnosis of heart disease (angiographic disease status)

Table 2: Dataset Attributes

3.4. Pretreatment

3.4.1. The Mean Imputation technique:

The Mean Imputation technique is one of the common methods used to handle missing values in datasets. This method relies on replacing missing values with the mean of the available values for the same variable.

The principle behind Mean Imputation is simple:

Calculating the Mean: The mean of the available values for the variable containing missing values is calculated.

The equation for calculating the mean value is simple and as follows:

$$Mean = \frac{Sum\ of\ all\ values}{Number\ of\ non_missing\ values}$$

Equation 4: Calculating the Mean

Where:

"Sum of all values" represents the total sum of all available values in the variable.

"Number of non-missing values" represents the count of available values in the variable, i.e., values that are not missing.

In short, we add up all the available values in the variable and then divide this sum by the count of values to obtain the mean value.

Replacing Missing Values: The missing values in the variable are then replaced with the mean value calculated in the first step.

Now, let's see how the Mean Imputation technique can be applied to "The Cleveland Heart Disease Dataset":

Calculating the Mean: We calculate the mean value of the variable for which we want to replace the missing values. For example, if we have a variable representing resting blood pressure (trestbps), we need to calculate the mean of the values in this variable.

Replacing Missing Values: After calculating the mean, we replace the missing values in the variable with the mean value calculated.

Applying this process helps to solve the problem of missing data and allows for maximum utilization of the available information in the dataset.

Additionally, it's worth noting that the Mean Imputation technique can be effectively used only when the missing data is random or can be objectively estimated.

3.4.2. Handling Missing Values with Rough Sets

Unlike many machine learning approaches, rough set theory offers a distinct advantage when dealing with missing values in datasets. Unlike traditional methods that require imputation (filling in missing values), rough sets can work directly with incomplete data. This capability stems from the core concept of indiscernibility relations, which group similar objects based on their available attributes. Missing values don't necessarily prevent objects from being categorized together if their remaining attributes are sufficiently similar.

However, it's important to acknowledge a potential trade-off. While rough sets can handle missing values directly, imputing missing values with appropriate strategies can sometimes lead to improved prediction model performance. This is because imputation can potentially reduce uncertainty and increase the granularity of the data used for model training.

Rough sets offer a valuable tool for situations with missing values, allowing analysis without the need for potentially inaccurate imputation. However, depending on the specific dataset and desired outcome, imputation might still be a worthwhile consideration to further enhance prediction accuracy.

3.4.3. Discretization:

The discretization stage in the Rough Set Exploration System (RSES) refers to the process of converting continuous variables into categorical variables. This is typically done to simplify subsequent analysis and processing, as analysis can be more efficient with categorical variables.

During this stage, continuous variables are divided into different categories using various methods such as fixed, dynamic, or other types of breakpoints. This is usually based on the distribution of values in the variable and the desired goal of the analysis [33].

To apply the discretization stage on The Cleveland Heart Disease dataset, the following steps can be followed:

Data exploration: Before beginning the variable discretization, it's essential to study the dataset and understand the distributions of values for each variable. This helps in determining the necessary steps for discretization.

Identification of continuous variables: Identify the variables that are continuous and require discretization, such as age, blood pressure, cholesterol levels, etc.

Selection of discretization method: Various methods can be used to discretize continuous variables, such as dividing them into equal-width bins or using more complex methods like choosing breakpoints based on statistical values or applying algorithms specific to data discretization.

Implementation of discretization: Execute the chosen method for discretization on the continuous variables. Convert the continuous values into categorical values based on the specified breakpoints.

Evaluation of results: After discretization, it's crucial to evaluate its impact on the data and the extent to which it improves our understanding of the relationships between variables.

The discretization stage is essential in data analysis to transform continuous variables into a format suitable for classification, enabling better use of classification and prediction techniques on the data.

3.5. Performance metrics

To assess and elucidate the precision of our study, we relied upon the four key metrics within the realm of Artificial Intelligence:

Accuracy: Serving as a fundamental measure, accuracy signifies the percentage of correctly classified instances, computed by dividing the count of accurately classified instances by the total instance count and scaling the result to a percentage.

$$Accuracy = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}} \times 100\%$$

Equation 5: Accuracy

Coverage: These metric gauges the efficacy of the generated rules by determining the proportion of testing instances encompassed by these rules. It is calculated by dividing the number of testing instances covered by the rules by the total number of testing instances.

$$\text{Coverage} = \frac{\text{Number of testing instances covered by rules}}{\text{Total number of testing instances}}$$

Equation 6: Coverage

Class-specific True Positives: Evaluating the model's performance for each class, particularly pertinent in medical diagnosis such as identifying instances with or without heart disease. This is expressed as the percentage of accurately classified instances within a specific class, divided by the total instances belonging to that class.

$$\text{True Positives for Class} = \frac{\text{Number of coorectly classifier instances in that class}}{\text{Total number of instances in that class}} \times 100\%$$

Equation 7: Class-specific True Positives

Total accuracy and coverage: These metrics encapsulate the overall model performance on the testing set. Total Accuracy is determined by dividing the total number of correctly classified instances by the total number of instances in the testing set, while Total Coverage is computed by dividing the number of testing instances covered by rules by the total number of instances in the testing set.

$$\text{Total Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances in testing set}} \times 100\%$$

Equation 8: Total Accuracy

$$\text{Total Coverage} = \frac{\text{Number of testing instances covered by rules}}{\text{Total number of testing instances in testing set}}$$

Equation 9: Total Coverage

These metrics collectively provide a comprehensive evaluation of the efficacy and robustness of the studied model within the context of classification tasks.

3.6. Software Tools for Rough Set Modeling

When creating a rough set model, several commercial and open-source software tools can be utilized for data analysis and handling uncertainty. Here are some notable tools:

1. Rough Set Data Explorer (ROSE)

Description: ROSE is a free software tool specifically designed for data analysis using rough set theory.

Features:

- Provides an interface for loading and analyzing data.
- Implements various rough set algorithms to compute lower and upper approximations.
- Supports decision rule generation and other rough set operations.

Usage: Ideal for users who need a dedicated environment for rough set analysis without extensive programming.

2. Rough Set Exploration System (RSES)

Description: RSES is a robust software platform offering comprehensive tools for data analysis and model design using rough set theory.

Features:

- Includes advanced algorithms for rough set analysis.
- Supports data preprocessing, rule extraction, and attribute reduction.
- Provides visualization tools to help interpret the results.

Usage: Suitable for researchers and practitioners who require powerful and flexible tools for detailed rough set analysis.

3. Python laibreries

- **pandas:** Used for data manipulation and analysis.
- **numpy:** Used for numerical operations and handling arrays.

Custom Functions: Users can write custom functions to compute lower and upper approximations, as well as boundary regions.

Usage:

Data Analysis: Users can leverage pandas and numpy for preliminary data analysis and preparation.

Algorithm Implementation: Custom rough set algorithms can be implemented directly in Python.

3.7. The RSES2 Tool

The Rough Set Exploration System (RSES) is a software tool utilized in the field of Artificial Intelligence (AI) for data analysis and rough set exploration within datasets [34]. It finds application across various domains:

- ❖ **Machine Learning:** RSES aids in data analysis by identifying rough set rules that represent relationships between variables, crucial for predictive modeling.
- ❖ **Classification and Prediction:** In the realm of supervised learning, RSES can be employed to create predictive models based on rough set rules, facilitating accurate classification and prediction tasks.
- ❖ **Data Simplification and Understanding:** RSES simplifies complex datasets by presenting rough set rules, thereby enhancing understanding and interpretability, which is pivotal in decision-making processes.
- ❖ **Data Mining:** RSES is instrumental in data mining endeavors, enabling the extraction of valuable insights and rules from large datasets, essential for decision support and knowledge discovery.

In essence, RSES serves as a powerful tool for data analysis, aiding in the elucidation of relationships and rules within datasets, thereby contributing significantly to scientific research and analytical applications in various domains.

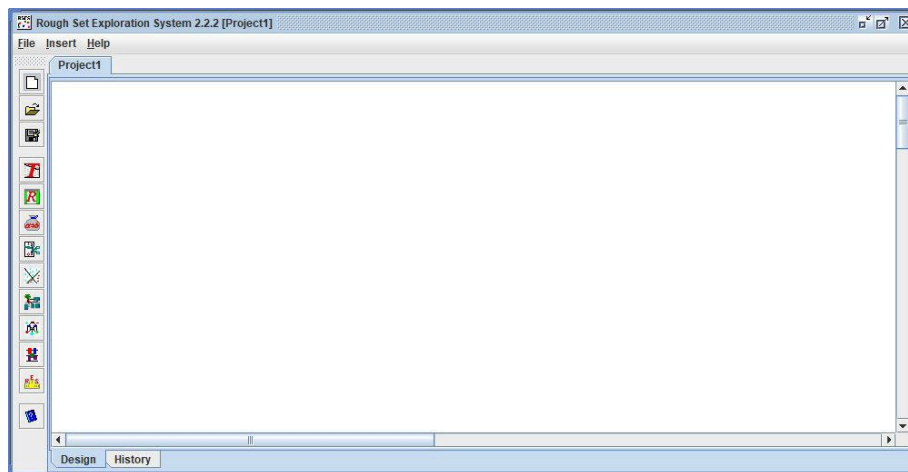


Figure 20: RSES2 Interface

3.8. The proposed methodology

This research prioritizes the development of a robust rough set model that achieves significant accuracy while remaining interpretable and user-friendly. To achieve this balance, we embarked on a comprehensive exploration and evaluation process. After careful consideration and experimentation,

we have chosen to focus on the most promising variants using the rough sets methodology that demonstrate this desired balance. For each chosen variant, we will delve into the details, including:

- **General architecture:** A clear representation of the overall model structure and how its components interact.
- **Main processes:** A breakdown of the core steps involved in the model's operation.
- **Specific points of interest:** Highlighting any unique aspects or customizations made to the variant for this specific application.
- **Training and testing protocols:** A detailed description of the procedures used to train and evaluate the model's performance.

By providing this level of detail, we aim to offer transparency in our approach and facilitate understanding of the chosen variants' functionalities. To create an integrated system that predicts the outcome of a diagnosis based on medical analyses, we conducted several experiments, the results of which will be discussed shortly.

3.8.1. Scenario 1

Rough Set Model with mean imputation and LEM2 rule generation

This scenario outlines the process of building a rough set model for cardiovascular disease (CVD) prediction using the Cleveland dataset (Fig. 14). Here's a breakdown of the steps involved:

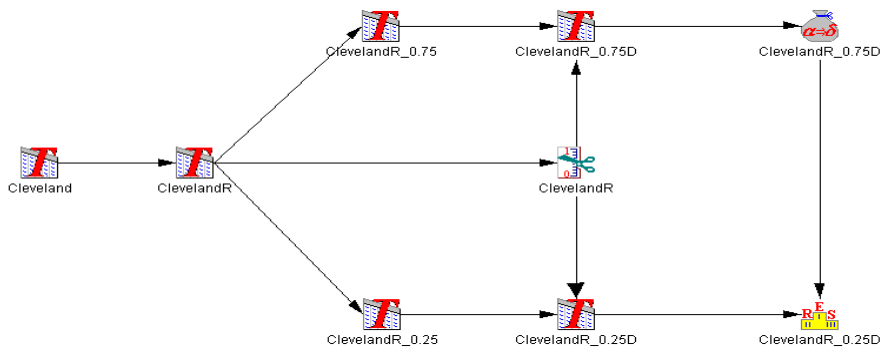


Figure 21: Architecture for the first scenario

1. Data Preprocessing:

- **Load data:** The Cleveland dataset is loaded in its tabular format.
- **Missing value imputation:** Missing values are imputed using the mean value (average) for each numerical attribute. This step aims to fill in missing information and potentially improve model performance, although it's important to acknowledge the potential trade-off discussed earlier.

- **Data splitting:** The resulting complete dataset is then divided into two parts:
 - **Training set (75%):** This larger portion of the data is used to train the rough set model and generate prediction rules.
 - **Testing set (25%):** This portion is reserved for evaluating the performance of the generated model on unseen data.

2. Discretization:

- **Constructing the Cut Table:** The training set is used to construct a "cut table." (Fig. 15) This table identifies potential thresholds for discretizing the continuous numerical attributes in the dataset. These thresholds will be used to convert the continuous values into categories for further analysis by the rough set algorithm.

(1-13)	Attribute	Size	Description
1	age	8	42.0; 46.5; 50.0; 56.5; 59.5; 61.5; 64.5; 65.5
2	sex	1	{ 1 }
3	cp	1	{ 1,2,3 }
4	trestbps	6	119.0; 122.0; 138.0; 145.0; 152.0; 175.0
5	chol	4	273.0; 301.0; 307.0; 326.5
6	fbs	0	*
7	restecg	1	{ 2 }
8	thalach	4	98.0; 144.5; 148.5; 155.0
9	exang	1	{ 1 }
10	oldpeak	2	0.55; 0.7
11	slope	1	{ 1,3 }
12	ca	3	{ 2 } { 3 } { 1 }
13	thal	2	{ 6 } { 3 }

Figure 22: Cut table for the 1st scenario

3. Rule Generation:

- **Discretized Training Set:** The training set is then discretized based on the thresholds identified in the cut table. Each continuous attribute is now represented by categories.
- **Rule Induction Algorithm:** The LEM2 algorithm, a popular choice in rough set theory, is employed to generate classification rules from the discretized training set. These rules aim to identify relationships between the various attributes (age, blood pressure, etc.) and the presence or absence of CVD (Fig. 16).

ID	Match	Decision rules
1	44	(chol-<inf.273.0)&(slope="1,3")&(ca="<call rest">)&(thal="3")=>(num={044})
2	37	(cp="<call rest">)&(oldpeak="0.7,inf")&(thal="<call rest">)=>(num={137})
3	35	(sex="1")&(cp="<call rest">)&(thalach="98.0,144.5")&(oldpeak="0.7,inf")=>(num={135})
4	32	(chol="<inf.273.0")&(exang="<call rest">)&(oldpeak="<inf.0.55")&(ca="<call rest">)&(thal="3")=>(num={032})
5	31	(thalach="98.0,144.5")&(oldpeak="0.7,inf")&(thal="<call rest">)=>(num={131})
6	30	(cp="<call rest">)&(thalach="98.0,144.5")&(exang="1")&(oldpeak="0.7,inf")=>(num={130})
7	29	(sex="<call rest">)&(cp="1,2,3")&(ca="<call rest">)=>(num={029})
8	28	(sex="<call rest">)&(chol="<inf.273.0")&(exang="<call rest">)&(ca="<call rest">)=>(num={028})
9	27	(cp="<call rest">)&(restecg="2")&(thal="<call rest">)=>(num={127})
10	26	(cp="1,2,3")&(trestbps="122.0,138.0")&(ca="<call rest">)&(thal="3")=>(num={026})
11	26	(trestbps="122.0,138.0")&(chol="<inf.273.0")&(exang="<call rest">)&(ca="<call rest">)&(thal="3")=>(num={026})
12	26	(cp="1,2,3")&(chol="<inf.273.0")&(oldpeak="<inf.0.55")&(ca="<call rest">)&(thal="3")=>(num={026})
13	26	(chol="<inf.273.0")&(thalach="155.0,inf")&(oldpeak="<inf.0.55")&(ca="<call rest">)&(thal="3")=>(num={026})
14	25	(age="150.0,56.5")&(cp="1,2,3")&(ca="<call rest">)=>(num={025})
15	24	(chol="<inf.273.0")&(restecg="<call rest">)&(thalach="155.0,inf")&(exang="<call rest">)&(ca="<call rest">)&(thal="3")=>(num={024})
16	24	(age="150.0,56.5")&(cp="1,2,3")&(thal="3")=>(num={024})
17	24	(cp="1,2,3")&(chol="<inf.273.0")&(thalach="155.0,inf")&(oldpeak="<inf.0.55")&(ca="<call rest">)=>(num={024})
18	23	(sex="1")&(cp="<call rest">)&(restecg="2")&(thalach="98.0,144.5")=>(num={123})
19	22	(age="150.0,56.5")&(thalach="155.0,inf")&(exang="<call rest">)&(ca="<call rest">)=>(num={022})
20	22	(exang="1")&(oldpeak="0.7,inf")&(slope="<call rest">)&(thal="<call rest">)=>(num={122})
21	22	(chol="<inf.273.0")&(oldpeak="0.7,inf")&(slope="<call rest">)&(thal="<call rest">)=>(num={122})
22	22	(sex="1")&(chol="<inf.273.0")&(oldpeak="<inf.0.55")&(ca="<call rest">)&(thal="3")=>(num={022})
23	21	(cp="1,2,3")&(trestbps="122.0,138.0")&(thalach="155.0,inf")&(ca="<call rest">)=>(num={021})
24	21	(trestbps="122.0,138.0")&(chol="<inf.273.0")&(thalach="155.0,inf")&(exang="<call rest">)&(ca="<call rest">)=>(num={021})
25	21	(cp="1,2,3")&(chol="<inf.273.0")&(restecg="<call rest">)&(thalach="155.0,inf")&(ca="<call rest">)&(thal="3")=>(num={021})
26	21	(age="150.0,56.5")&(exang="<call rest">)&(ca="<call rest">)&(thal="3")=>(num={021})
27	21	(sex="<call rest">)&(restecg="<call rest">)&(exang="<call rest">)&(ca="<call rest">)=>(num={021})
28	20	(cp="1,2,3")&(chol="<inf.273.0")&(restecg="<call rest">)&(thalach="155.0,inf")&(oldpeak="<inf.0.55")=>(num={020})
29	19	(cp="<call rest">)&(restecg="2")&(thalach="98.0,144.5")&(exang="1")=>(num={119})
30	19	(age="150.0,56.5")&(exang="<call rest">)&(slope="1,3")&(ca="<call rest">)=>(num={019})
31	19	(sex="<call rest">)&(cp="1,2,3")&(restecg="<call rest">)&(thal="3")=>(num={019})
32	19	(sex="<call rest">)&(chol="<inf.273.0")&(restecg="<call rest">)&(exang="<call rest">)&(thal="3")=>(num={019})

Figure 23: Extract rules for the 1st scenario

4. Model Evaluation:

- **Testing Set Evaluation:** The generated rule set is then applied to the unseen testing set. The model's performance is assessed using a confusion matrix, which highlights the number of correct and incorrect predictions made by the model.

➤ Overall, this scenario demonstrates how rough sets can build a predictive model for CVD diagnosis by handling missing values, discretizing continuous data, and generating interpretable classification rules using LEM2.

In this experiment, we relied on removing the missing values from the data set completely and using the LEM2 algorithm to generate the classification rules. A smaller set of rules (resulting in **4047** rules) was obtained compared to the coming experiments where the exhaustive search method is applied for the rules generation phase.

3.8.2. Scenario 2

Rough set model with mean imputation and exhaustive algorithm for rule generation

This scenario outlines the process of building a rough set model for cardiovascular disease (CVD) prediction using the Cleveland dataset with mean imputation and exhaustive rule generation (Fig. 17). Here's a breakdown of the steps involved:

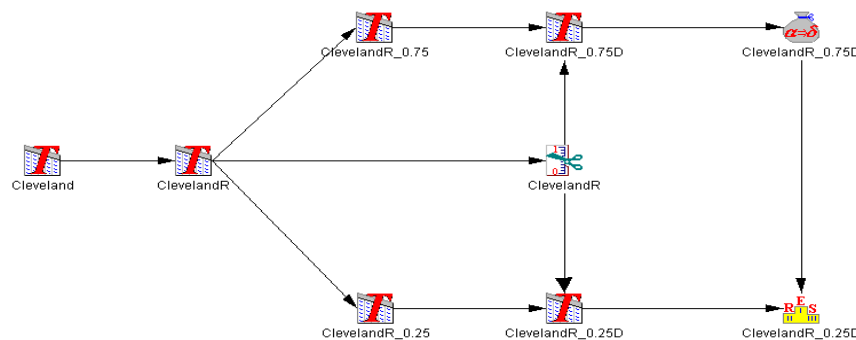


Figure 24: Architecture for the first scenario

This scenario builds upon Scenario 1 by leveraging the potentially superior performance of the Exhaustive rule generation algorithm while maintaining the initial approach to data handling. Here's a breakdown of the steps involved:

1. Data Preprocessing:

- **Load Data:** The Cleveland dataset is loaded in its tabular format.

- **Missing Value Imputation:** Missing values are imputed using the mean value (average) for each numerical attribute, similar to Scenario 1.
- **Data Splitting:** The resulting complete dataset is then divided into two parts:
 - **Training Set (75%):** This portion of the data is used to train the rough set model and generate prediction rules.
 - **Testing Set (25%):** This portion is reserved for evaluating the performance of the generated model on unseen data.

2. Discretization:

- **Constructing the Cut Table:** The training set is used to construct a "cut table." This table identifies potential thresholds for discretizing the continuous numerical attributes in the dataset. These thresholds will be used to convert the continuous values into categories for further analysis by the rough set algorithm.

3. Rule Generation:

- **Discretized Training Set:** The training set is then discretized based on the thresholds identified in the cut table. Each continuous attribute is now represented by categories.
- **Exhaustive Algorithm:** Unlike Scenario 1, this scenario employs the Exhaustive algorithm to generate classification rules from the discretized training set. This algorithm considers all possible combinations of attributes, potentially leading to a more comprehensive set of rules that might explain improved performance.

4. Model Evaluation:

- **Testing Set Evaluation:** The generated rule set is applied to the unseen testing set. The model's performance is assessed using a confusion matrix, which highlights the number of correct and incorrect predictions made by the model.

General Process and Architecture:

This scenario follows a similar overall architecture as Scenario 1. The main difference lies in the rule generation step, where the Exhaustive algorithm replaces the LEM2 algorithm. This potentially allows the model to capture more complex relationships between attributes in the training data, leading to better performance as observed.

Benefits of This Scenario:

- **Improved Performance:** By leveraging the Exhaustive algorithm, this scenario might achieve higher accuracy compared to Scenario 1 due to potentially more comprehensive rule generation.

Trade-offs and Considerations:

- **Potential for Overfitting:** The Exhaustive algorithm can generate a large number of rules, potentially leading to overfitting on the training data. Careful selection of relevant rules might be necessary.
- **Interpretability:** The increased number of rules in this scenario might make the model less interpretable compared to Scenario 1. This could necessitate additional techniques for rule pruning or selection to improve interpretability while maintaining performance.

➤ Overall, Scenario 2 prioritizes performance by utilizing the Exhaustive algorithm. However, it's crucial to address potential overfitting and potentially reduced interpretability to ensure a robust and interpretable model.

3.8.3. Scenario 3

Rough set model with exhaustive algorithm for rules generation and without imputation of missing values

This scenario (Fig. 18) explores the impact of imputing missing values only in the training set while preserving the raw data in the testing set. It utilizes the exhaustive rule generation algorithm like Scenario 2.

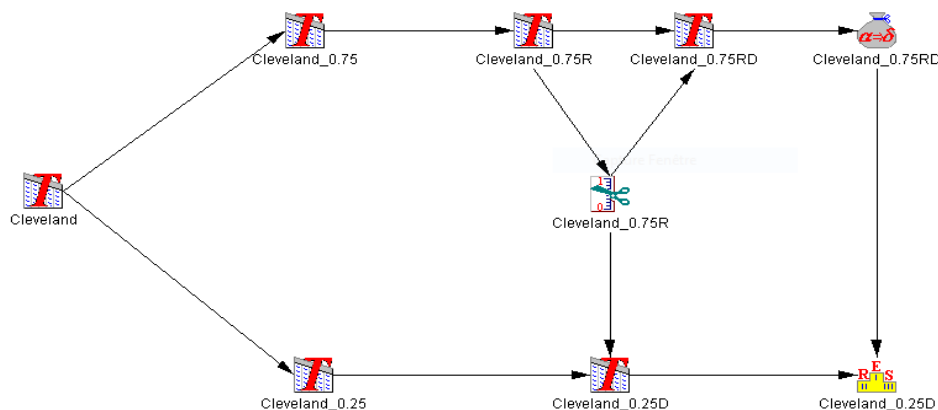
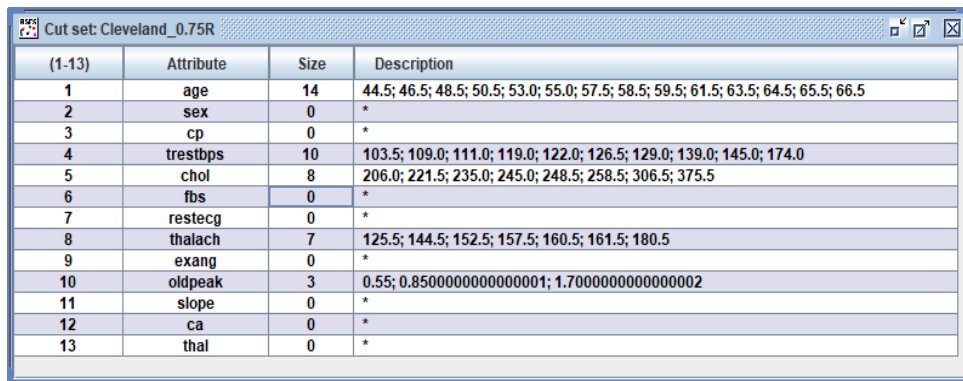


Figure 25: Architecture for the third scenario

Process:

1. **Data splitting:** Similar to the 1st and 2nd Scenarios, the Cleveland dataset is split into a training set (75%) and a testing set (25%).
2. **Missing value imputation (for the training set only):** Missing values in the training set are imputed using the mean value for each numerical attribute. This aims to potentially improve model performance during training without altering the raw data in the testing set.
3. **Discretization:**
 - **Cut Table construction:** The training set (with imputed values) is used to create a "cut table" (Fig. 19) that identifies potential thresholds for discretizing continuous attributes.



(1-13)	Attribute	Size	Description
1	age	14	44.5; 46.5; 48.5; 50.5; 53.0; 55.0; 57.5; 58.5; 59.5; 61.5; 63.5; 64.5; 65.5; 66.5
2	sex	0	*
3	cp	0	*
4	trestbps	10	103.5; 109.0; 111.0; 119.0; 122.0; 126.5; 129.0; 139.0; 145.0; 174.0
5	chol	8	206.0; 221.5; 235.0; 245.0; 248.5; 258.5; 306.5; 375.5
6	fbs	0	*
7	restecg	0	*
8	thalach	7	125.5; 144.5; 152.5; 157.5; 160.5; 161.5; 180.5
9	exang	0	*
10	oldpeak	3	0.55; 0.8500000000000001; 1.7000000000000002
11	slope	0	*
12	ca	0	*
13	thal	0	*

Figure 26: Cut table for the 3rd scenario

- **Discretization of both sets:** Both the training and testing sets are discretized using the thresholds identified from the cut table. This ensures consistency in the attribute representation across both sets.
4. **Rule Generation:**
 - **Exhaustive algorithm:** Unlike Scenario 1, this scenario uses the Exhaustive algorithm to generate classification rules from the discretized training set. This algorithm considers all possible combinations of attributes, potentially leading to a more comprehensive set of rules.

(1-737...	Match	Decision rules
1	31	(thalach="(161.5,180.5)")&(ca=0)&(thal=3)=>(num={0[31]})
2	28	(cp=4)&(exang=1)&(oldpeak="(1.7000000000000002,inf)")=>(num={1[28]})
3	25	(thalach="(161.5,180.5)")&(exang=0)&(slope=1)&(ca=0)=>(num={0[25]})
4	25	(sex=0)&(cp=3)&(thal=3)=>(num={0[25]})
5	24	(cp=4)&(oldpeak="(1.7000000000000002,inf)")&(thal=7)=>(num={1[24]})
6	24	(sex=0)&(fbs=0)&(restecg=0)&(exang=0)&(thal=3)=>(num={0[24]})
7	23	(thalach="(161.5,180.5)")&(oldpeak="(Inf,0.55)")&(ca=0)=>(num={0[23]})
8	22	(restbps="(129.0,139.0)")&(exang=0)&(ca=0)=>(num={0[22]})
9	22	(restbps="(129.0,139.0)")&(ca=0)&(thal=3)=>(num={0[22]})
10	22	(age="(Inf,44.5)")&(ca=0)&(thal=3)=>(num={0[22]})
11	21	(exang=1)&(oldpeak="(1.7000000000000002,inf)")&(thal=7)=>(num={1[21]})
12	20	(cp=3)&(thalach="(161.5,180.5)")&(thal=3)=>(num={0[20]})
13	19	(cp=4)&(restecg=2)&(slope=2)&(thal=7)=>(num={1[19]})
14	19	(sex=1)&(cp=4)&(oldpeak="(1.7000000000000002,inf)")&(slope=2)=>(num={1[19]})
15	19	(sex=0)&(restecg=0)&(exang=0)&(ca=0)=>(num={0[19]})
16	18	(sex=1)&(cp=4)&(restecg=2)&(exang=1)&(slope=2)=>(num={1[18]})
17	18	(thalach="(125.5,144.5)")&(exang=1)&(thal=7)=>(num={1[18]})
18	18	(sex=0)&(restecg=0)&(slope=1)=>(num={0[18]})
19	18	(cp=3)&(thalach="(161.5,180.5)")&(oldpeak="(Inf,0.55)")=>(num={0[18]})
20	18	(sex=0)&(cp=3)&(ca=0)=>(num={0[18]})
21	18	(restecg=0)&(thalach="(161.5,180.5)")&(oldpeak="(Inf,0.55)")&(thal=3)=>(num={0[18]})
22	17	(age="(Inf,44.5)")&(restecg=0)&(thal=3)=>(num={0[17]})
23	17	(cp=3)&(thalach="(161.5,180.5)")&(ca=0)=>(num={0[17]})
24	17	(sex=0)&(cp=3)&(slope=1)=>(num={0[17]})
25	16	(thalach="(125.5,144.5)")&(oldpeak="(1.7000000000000002,inf)")&(thal=7)=>(num={1[16]})
26	16	(age="(Inf,44.5)")&(thalach="(161.5,180.5)")&(ca=0)=>(num={0[16]})
27	16	(fbs=0)&(exang=0)&(oldpeak="(0.8500000000000001,1.7000000000000002)")&(thal=3)=>(num={0[16]})
28	16	(chol="(Inf,206.0)")&(exang=0)&(ca=0)&(thal=3)=>(num={0[16]})
29	16	(exang=0)&(oldpeak="(0.8500000000000001,1.7000000000000002)")&(ca=0)&(thal=3)=>(num={0[16]})
30	16	(sex=0)&(exang=0)&(slope=2)&(ca=0)=>(num={0[16]})
31	16	(fbs=0)&(restecg=0)&(thalach="(161.5,180.5)")&(exang=0)&(oldpeak="(Inf,0.55)")=>(num={0[16]})
32	16	(restecg=0)&(thalach="(161.5,180.5)")&(exang=0)&(oldpeak="(Inf,0.55)")&(slope=1)=>(num={0[16]})

Figure 27: Extract Rules for the 3rd scenario

5. Model Evaluation:

- **Testing with generated rules:** The rule set generated from the training data is then applied to the discretized testing set (with original, un-imputed data).
- **Confusion matrix:** The model's performance is assessed using a confusion matrix, indicating the accuracy of predictions made on unseen data with authentic values.

Benefits:

- **Preserved data authenticity:** Maintaining the original data in the testing set allows for a more realistic evaluation of the model's performance on real-world scenarios with potentially missing values.
- **Potentially improved accuracy:** Imputing missing values in the training set can potentially enhance the model's ability to learn relationships from the data. This might lead to improved accuracy compared to models trained on raw data with missing values.
- **Enhanced interpretability (Debatable):** While the Exhaustive algorithm generates more comprehensive rules, their interpretability can be lower due to the sheer number of rules potentially involved. Analyzing these rules might be more complex compared to the more focused rules generated by LEM2 in Scenario 1.

Improved results (Cautiously optimistic):

This scenario offers a potential trade-off between data authenticity and model performance. While preserving raw data in the testing set ensures a realistic evaluation, imputing missing values in the training set might lead to better rule generation and improved accuracy. The effectiveness can be evaluated by comparing the confusion matrices from both scenarios. It's important to note that the Exhaustive algorithm might not always lead to better results compared to LEM2. Choosing the optimal approach depends on the specific dataset and desired balance between interpretability and accuracy.

➤ *Overall, this scenario highlights the flexibility of rough sets in handling missing values and rule generation strategies. It offers a valuable exploration for potentially enhancing model performance while maintaining some level of data authenticity in the testing set.*

3.9. Discussion:

In this section, we will elucidate the rationale behind the inclusion of each component outlined in this chapter, aiming to enhance the comprehensibility and elucidate the study:

Utilizing a publicly available dataset like Cleveland: Leveraging a well-established dataset such as the Cleveland dataset enhances reproducibility and facilitates comparability with other research endeavors in the field of AI. This ensures that findings can be validated and extended by other researchers, fostering collaborative advancements.

Mean or median imputation for missing values: While mean or median imputation serves as a rudimentary approach for handling missing data, it provides a foundational framework for initial analysis. However, future research may delve into more sophisticated imputation techniques within the domain of AI, such as predictive imputation or multiple imputation, to enhance the robustness of the data preprocessing phase.

The chosen discretization method: Documenting and justifying the selected discretization method based on the specific characteristics of the data is crucial for ensuring the integrity and reliability of the analysis. Whether employing equal-width binning, equal-frequency binning, or other discretization techniques, it is imperative to align the method with the data distribution and the requirements of the classification task at hand.

The exhaustive algorithm for rule generation: Employing an exhaustive algorithm ensures comprehensive coverage of all possible relationships within the training data, thereby maximizing the potential for accurate rule generation. Nonetheless, future research endeavors might explore more

efficient rule reduction strategies, such as pruning techniques or evolutionary algorithms, to streamline the rule set while preserving predictive performance.

Utilizing standard classification metrics: Adopting standard classification metrics provides a transparent and interpretable means of evaluating the model's performance on unseen data. Metrics such as accuracy, precision, recall, and F1-score offer valuable insights into the model's predictive capabilities, enabling informed decision-making and refinement of the classification model within the AI domain.

3.10. Conclusion

In this chapter, we delineate the diverse data preprocessing and training methodologies employed within this study, accompanied by essential insights into the dataset utilized during the training phase. Furthermore, we introduce the architecture of RSES2 proposed within the context of this thesis, elucidating its conceptual framework and key components.

Subsequently, we expound upon the methodology for evaluating the efficacy of the proposed structure, outlining the procedures for calculating the accuracy of the obtained results. This preparatory groundwork sets the stage for the forthcoming chapter, wherein we will expound upon the experimental outcomes derived from our training methodologies. Additionally, we will furnish detailed insights into the implementation environment utilized for conducting the training exercises, thus providing a comprehensive overview of the experimental setup and methodology.

Chapter 04:

Experimental Study

1.1 Introduction

In this chapter, we will provide information about the execution environment that we used, both the hardware and the software, secondly, we will showcase the different acquired training results which we will judge on 4 metrics: Accuracy, Coverage, Class-specific True Positives, and Total Accuracy and Coverage. We conclude by showcasing the contribution of our work and comparing its performance against existing research, particularly those employing alternative machine learning techniques.

4.1. Experimental Environment

This section details the hardware and software used for all training and experimentation conducted in this thesis.

4.1.1. Hardware

The experiments were performed on a dedicated laptop computer with the following specifications:

- Processor: Intel® Core™ i3-5005U CPU @ 2.00 GHz
- Memory: 4.00 GB
- Operating System: Windows 10 Professional (64-bit)
- Storage Capacity: 931.50 GB

4.1.2. Software

The primary software utilized for this research was the RSES2 program. This language-independent software facilitated a comprehensive range of functionalities:

- ☞ Dataset experimentation
- ☞ Rule extraction
- ☞ Feature evaluation
- ☞ Result analysis

RSES2's powerful data manipulation and analysis capabilities enabled us to assess the effectiveness of various features and derive interpretable rules for improved model performance.

4.2. Experimental results

In this section, we will discuss, step by step, all the different results we have obtained through our methodology, providing explanations and justifications.

Our tests were performed on the Cleveland Heart Disease dataset. As we explained in the third chapter, our study was divided into three main scenarios of experiments. We will review their results in this section. Additionally, we will discuss some other experiments that were attempted and whose results did not meet the study's goals. These were not previously mentioned in the third chapter, but they and their results will be included in this section.

4.2.1. Scenario 1

In the first scenario, our initial experimental approach focused on building a robust and interpretable rough set model for cardiovascular disease (CVD) prediction. We employed the Cleveland dataset and addressed missing values using mean imputation to facilitate model training. The data was then split into 75% training and 25% testing sets. The training set was used to construct a cut table for discretizing continuous attributes and subsequently generate classification rules leveraging the LEM2 algorithm. The rule extraction process yielded **4047 rules**. The generated rule set was then evaluated on the unseen testing set, and its performance was assessed using a confusion matrix.

The results of this experiment were as follows:

		Predicted				
		0	1	No. of obj.	Accuracy	Coverage
Actual	0	36	8	44	0.818	1
	1	1	31	32	0.969	1
True positive rate		0.97	0.79			
Total number of tested objects: 76						
Total accuracy: 0.882						
Total coverage: 1						

Figure 28: The Results of the first experimental scenario

After extracting the rules and training the model, we conducted a test using a dataset containing two categories: 0 for healthy individuals and 1 for unhealthy patients. There were 44 items in the healthy category and 32 items in the unhealthy patient category. The model correctly identified 36 healthy

items and made 8 errors. For the unhealthy patient category, the model correctly identified 31 items as patients and made 1 error.

Accuracy:

- Accuracy for healthy items: 0.818
- Accuracy for unhealthy patient items: 0.969
- Total accuracy: 0.882

Coverage:

- Coverage for healthy items: 1.0
- Coverage for unhealthy patient items: 1.0
- Total coverage: 1.0

True positive rate:

- True positive rate for healthy items: 0.97
- True positive rate for unhealthy patient items: 0.79

Confusion Matrix:

Testing Set			
TARGET \ OUTPUT	Class0	Class1	SUM
Class0	36 47.37%	8 10.53%	44 81.82% 18.18%
Class1	1 1.32%	31 40.79%	32 96.88% 3.13%
SUM	37 97.30% 2.70%	39 79.49% 20.51%	67 / 76 88.16% 11.84%

Figure 29: Confusion Matrix for the first scenario

4.2.2. Scenario 2

The 2nd scenario aimed to potentially improve the performance of the rough set model for CVD prediction compared to Scenario 1. It maintained the same data handling approach with mean imputation for missing values and a 75%/25% split for training and testing. However, it deviated in the rule generation step. Instead of the LEM2 algorithm, Scenario 3 utilized the Exhaustive algorithm, which considers all possible combinations of attributes. This potentially captured more complex relationships in the data, leading to the observed improvement in performance. However, it's important to acknowledge the potential trade-offs of increased rule complexity, including overfitting and reduced interpretability.

The experimental results of this experiment were as follows:

		Predicted				
		0	1	No. of obj.	Accuracy	Coverage
Actual	0	39	2	41	0.951	1
	1	4	31	35	0.886	1
	True positive rate	0.91	0.94			

Total number of tested objects: 76
 Total accuracy: 0.921
 Total coverage: 1

Figure 30: The results of the 2nd experimental scenario

After extracting the rules and training the model, we conducted a test using a dataset containing two categories: 0 for healthy individuals and 1 for unhealthy patients. There were 41 items in the healthy category and 35 items in the unhealthy patient category. The model correctly identified 39 healthy items and made 2 errors. For the unhealthy patient category, the model correctly identified 31 items as patients and made 4 error.

Accuracy:

- Accuracy for healthy items: 0.951
- Accuracy for unhealthy patient items: 0.886
- Total accuracy: 0.921

Coverage:

- Coverage for healthy items: 1.0

- Coverage for unhealthy patient items: 1.0
- Total coverage: 1.0

True Positive Rate:

- True positive rate for healthy items: 0.91
- True positive rate for unhealthy patient items: 0.94

Confusion Matrix:

Testing Set			
TARGET \ OUTPUT	Class0	Class1	SUM
Class0	39 51.32%	2 2.63%	41 95.12% 4.88%
Class1	4 5.26%	31 40.79%	35 88.57% 11.43%
SUM	43 90.70% 9.30%	33 93.94% 6.06%	70 / 76 92.11% 7.89%

Figure 31: Confusion Matrix for the 2nd scenario

4.2.3. Scenario 3

In this variant, the experimentation protocols in the 3rd scenario prioritized data authenticity for testing while exploring potential performance improvement through rule generation. Similar to Scenario 1, missing values were addressed using mean imputation but only in the training data (75%). The testing data (25%) remained untouched to preserve its original state with missing values. Interestingly, both training and testing sets were discretized using the cut table generated from the complete training set. This ensured consistency during rule generation and evaluation. However, unlike Scenario 1's LEM2 algorithm, Scenario 2 employed the Exhaustive algorithm. This potentially led to a more comprehensive rule set, although it might require careful selection to address potential overfitting and maintain interpretability. The rule extraction process yielded **7379 rules**, representing the modal for the system trained using rough set techniques.

The results of this experiment were as follows:

		Predicted				
		0	1	No. of obj.	Accuracy	Coverage
Actual	0	41	2	43	0.953	1
	1	3	30	33	0.909	1
	True positive rate	0.93	0.94			

Total number of tested objects: 76
 Total accuracy: 0.934
 Total coverage: 1

Figure 32: The results of the 3rd scenario

After extracting the rules and training the model, we conducted a test using a dataset containing two categories: 0 for healthy individuals and 1 for unhealthy patients. There were 43 items in the healthy category and 33 items in the unhealthy patient category. The model correctly identified 41 healthy items and made 2 errors. For the unhealthy patient category, the model correctly identified 30 items as patients and made 3 errors.

Accuracy:

- Accuracy for healthy items: 0.953
- Accuracy for unhealthy patient items: 0.909
- Total accuracy: 0.934

Coverage:

- Coverage for healthy items: 1.0
- Coverage for unhealthy patient items: 1.0
- Total coverage: 1.0

True Positive Rate:

- True positive rate for healthy items: 0.93
- True positive rate for unhealthy patient items: 0.94

Confusion Matrix:

Testing Set			
TARGET \ OUTPUT	Class0	Class1	SUM
Class0	41 53.95%	2 2.63%	43 95.35% 4.65%
Class1	3 3.95%	30 39.47%	33 90.91% 9.09%
SUM	44 93.18% 6.82%	32 93.75% 6.25%	71 / 76 93.42% 6.58%

Figure 33: Confusion matrix for the 3rd scenario

Analysis:

These results indicate that the last model demonstrates a high level of accuracy and coverage, with excellent ability to correctly classify both healthy and unhealthy patient items. The high true positive rates reflect the model's efficiency in detecting positive cases in both categories, showcasing its strong performance in accurately recognizing and distinguishing between the correct patterns.

4.2.4. Summary of experimental results

This table summarizes the performance metrics for the three Rough Set model scenarios explored in this work:

	Accuracy		Total Accuracy	Coverage	True Positive Rate	
	Healthy	Unhealthy			Healthy	Unhealthy
First scenario	0.818	0.969	0.882	1.0	0.97	0.79
Second scenario	0.951	0.886	0.921	1.0	0.91	0.94
Third scenario	0.953	0.909	0.934	1.0	0.93	0.94

Table 3: Summary of experimental results

Observations:

- **Total Accuracy:** Both Scenario 2 (0.921) and Scenario 3 (0.934) achieved higher overall accuracy compared to Scenario 1 (0.882).

- **Accuracy (Unhealthy):** Scenario 1 outperformed the others in identifying unhealthy cases (0.969). However, Scenario 2 and 3 achieved significant improvement (0.886 and 0.909) compared to Scenario 1.
- **True Positive Rate (Healthy):** Scenario 1 exhibited the highest rate of correctly identifying healthy cases (0.97). Scenario 2 and 3 maintained a good performance (0.91 and 0.93).
- **True Positive Rate (Unhealthy):** Interestingly, Scenario 2 achieved the best performance in identifying unhealthy cases (0.94), followed by Scenarios 3 (0.94) and 1 (0.79).
- **Coverage:** All scenarios achieved a perfect coverage of 1.0, indicating they were able to classify all instances in the testing set.

Insights:

- Scenario 2 prioritized data authenticity for testing but explored performance improvement with the Exhaustive algorithm. This approach resulted in a good balance between accuracy and true positive rates for both healthy and unhealthy cases.
- Scenario 3, with mean imputation for training data only and the Exhaustive algorithm, achieved the highest overall accuracy. However, its performance in identifying healthy cases was slightly lower than Scenario 1.

Conclusion:

The choice between these scenarios depends on the specific priorities. Scenario 1 offers a good balance for interpretability due to the LEM2 algorithm, while Scenario 2 prioritizes data authenticity for testing. Scenario 3 offers the highest total accuracy but might require additional considerations for interpretability due to the potentially more complex rules generated by the Exhaustive algorithm.

4.3. Comparisons

In this section, we provide a state-of-the-art comparison from works that have utilized the same data set as we did i.e., The Cleveland Heart Disease Dataset as they rank amongst the most used datasets in heart disease domain.

First, we Presented in Table the balanced accuracy results from our work as well as a few other papers that used The Cleveland Heart Disease Dataset in their tests.

	Study	Approach	Dataset	Accuracy%
1	Bhatia et al. [35]	SVM+GA	The Cleveland (5 class)	72.55%
			The Cleveland (2 class)	90.57%
3	Nayeem et al. [18]	SVM	The Cleveland (5 class)	59.01%
			The Cleveland (2 class)	92.45%
		MLP	The Cleveland (5 class)	68.86%,
			The Cleveland (2 class)	90.57%
5	Lin et al. [19]	CNN	The Cleveland (2 class)	92.81%
6	Dinesh et al. [20]	Random Forest + Decision Tree	The Cleveland (2 class)	88.7%
7	Our Work	Rough Set	The Cleveland (2 class)	93.4%

Several studies have explored different approaches to heart disease diagnosis using various machine learning techniques. Bhatia et al. [14] employed SVM with Genetic Algorithm (GA) on The Cleveland dataset, achieving accuracies of 72.55% for a 5-class classification and 90.57% for a 2-class

Table 4: Comparisons

classification. Nayeem et al. [21] utilized SVM and MLP on both 5-class and 2-class versions of The Cleveland dataset, achieving accuracies ranging from 59.01% to 92.45% and from 68.86% to 90.57%, respectively. Lin et al. [23] employed Convolutional Neural Network (CNN), achieving an accuracy of 92.81% on The Cleveland dataset's 2-class classification. Dinesh et al. [24] utilized Random Forest with Decision Tree, achieving an accuracy of 88.7% on The Cleveland dataset's 2-class classification.

In comparison, our work focused on utilizing Rough Set theory for heart disease diagnosis. We achieved an accuracy of 93.4% on The Cleveland dataset's 2-class classification.

4.4. Validation

Our study makes a significant contribution by demonstrating the effectiveness of Rough Set theory in heart disease diagnosis. Despite the varied approaches in previous studies, our methodology yielded competitive accuracy rates. This underscores the potential of Rough Set theory as a valuable tool in medical diagnosis tasks, providing an alternative or complementary approach to existing machine learning techniques. By expanding the range of methodologies available for heart disease diagnosis, our work contributes to enhancing the accuracy and reliability of diagnostic systems in clinical practice.

4.5. Future prospects for system improvement

Based on the results and challenges addressed in this study, several future areas can be identified to focus on improving the system and providing new enhancements that contribute to better accuracy in diagnosing and assessing cardiovascular disease risks.

Improving model accuracy with diverse data

The model's accuracy can be enhanced by applying it to diverse datasets that include various demographic, geographic, and health characteristics. Utilizing data from multiple sources can help make the model more comprehensive and suitable for use in different environments.

Exploring advanced machine learning techniques

In addition to rough set theory, integrating other machine learning techniques such as deep learning and deep neural networks can improve predictive accuracy. These techniques can enhance the model's ability to recognize complex patterns in the data.

Improving handling of missing data

Exploring more advanced techniques for handling missing data can boost predictive accuracy. Techniques like multi-task learning and transfer learning can be particularly beneficial in this context.

By focusing on these future prospects, the system can be significantly enhanced to be a more effective tool for predicting and managing cardiovascular diseases. The proposed improvements aim to enhance the model's accuracy and reliability, ultimately contributing to better patient outcomes and more personalized and effective healthcare delivery.

4.6. Conclusion

In this chapter, we initiated by delineating the execution environment utilized for our training processes, followed by a comprehensive description of the dataset splits. Subsequently, we presented the outcomes of our training endeavors, delving into an examination of four distinct evaluation metrics: Accuracy, Coverage, True Positive Rate, and Total Accuracy. Of particular significance, Total Accuracy emerges as a pivotal metric for gauging the efficacy of models trained on datasets. Notably, our most proficient model, the first experimental iteration, achieved a balanced accuracy of 93.4% through the utilization of the Mean Imputation technique. Furthermore, we conducted a comparative analysis of these findings vis-à-vis state-of-the-art works in the field.

General Conclusion

Cardiovascular diseases represent a significant global health challenge, contributing to millions of deaths annually and causing considerable suffering and disability worldwide. Despite advancements in treatment, predicting risks and early diagnosis remain crucial for effective management and prevention efforts. Traditional risk assessment models often lack interpretability, hindering their clinical utility.

This study aimed to address these challenges by developing an innovative model based on rough sets for accurately evaluating cardiovascular disease risks. Leveraging machine learning techniques, particularly rough set theory, our model provides a valuable diagnostic tool for physicians to better manage at-risk patients.

Through a series of experiments, we evaluated the performance of our model, considering metrics such as accuracy, coverage, and true positive rate. Experiment 1, utilizing the Mean Imputation technique, achieved a balanced accuracy of 93.4%, demonstrating the effectiveness of our approach. Experiment 2, preserving the raw dataset, and Experiment 3, eliminating missing data points, yielded balanced results, further validating the robustness of our model.

Our study contributes to the field by providing an interpretable rough sets-based model for cardiovascular risk assessment. By comparing our model to traditional risk assessment methods and showcasing its performance across various metrics, we demonstrate its potential applicability in clinical practice. The interpretability and efficiency of our model offer valuable insights for healthcare professionals, aiding in more informed decision-making and personalized patient care.

In conclusion, our research underscores the significance of interpretable machine learning models in healthcare and presents a promising avenue for future research in cardiovascular disease prediction. By continuing to refine and validate our model on diverse datasets and exploring alternative methodologies, we can further enhance its accuracy and applicability, ultimately improving patient outcomes and healthcare delivery.

Bibliography

- [1] "Shanthi M, Pekka P, Norrving B (2011). "Global Atlas on Cardiovascular Disease Prevention and Control" . World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization."
- [2] E. Ammenwerth, P. Nykänen, M. Rigby and N. de Keizer, "Clinical decision support systems: Need for evidence, need for evaluation. Artificial Intelligence in Medicine," 2013.
- [3] "University Diagnostic Medical Imaging," [Online]. Available: <https://www.udmi.net/cardiovascular-disease-risk/>.
- [4] "Mayet J, Hughes A (September 2003). "Cardiac and vascular pathophysiology in hypertension"".
- [5] "Durrington P (August 2003). "Dyslipidaemia"."
- [6] ""Myocardial ischemia - Symptoms and causes". Mayo Clinic."
- [7] " "Cardiac Arrest - What Is Cardiac Arrest? | NHLBI, NIH"".
- [8] "Burden of valvular heart diseases: a population-based study. Nkomo VT, Gardin JM, Skelton TN, Gottdiener JS, Scott CG, Enriquez-Sarano."
- [9] "Leslie Thomas, M.D., a nephrologist at Mayo Clinic, answers the important questions you may have about hypertension (high blood pressure).," [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/diagnosis-treatment/drc-20373417>.
- [10] D. S. Paknikar, "Medindia," [Online]. Available: <https://www.medindia.net/health/diagnosis/lipid-profile-screening.htm>.
- [11] "Cleveland Clinic," [Online]. Available: <https://my.clevelandclinic.org/health/diagnostics/16953-electrocardiogram-ekg>.

- [12] [Online]. Available: <https://www.alfredhealth.org.au/services/echocardiography>.
- [13] [Online]. Available: <https://www.bcmhospital.com/calcium-score/>.
- [14] "Ahmed, Md. Razu & Mahmud, S M Hasan & Hossin, Md & Jahan, Hosney & Noori, Sheak. (2018). A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms. 10.1109/CompComm.2018.8781022."
- [15] "M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C)".
- [16] "M. H. Abu Yazid, M. Haikal Satria, S. Talib and N. Azman, "Artificial Neural Network Parameter Tuning Framework For Heart Disease Classification," 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)".
- [17] "M. S. Satu, F. Tasnim, T. Akter and S. Halder, "Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)".
- [18] "M. Nahiduzzaman, M. J. Nayeem, M. T. Ahmed and M. S. U. Zaman, "Prediction of Heart Disease Using Multi-Layer Perceptron Neural Network and Support Vector Machine," 2019 4th International Conference on Electrical Information and Communication Technology (".
- [19] "C. -H. Lin, P. -K. Yang, Y. -C. Lin and P. -K. Fu, "On Machine Learning Models for Heart Disease Diagnosis," 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)".
- [20] "M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT)".
- [21] "A. Singh and A. Jain, "Prediction of Heart Disease using Dense Neural Network," 2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT), New Delhi, India, 2022, pp. 1-5, doi: 10.1109/GlobConPT57482.2022.9938354."

- [22] "N. N. Itoo and V. K. Garg, "Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers," 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 2022, pp. 599-604, doi: 10.1109/MECON53876.2022."
- [23] "M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani and A. Shaikh, "A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks," in IEEE Access, vol. 11, pp. 121574-121591, 2023".
- [24] "M. T. Rahman, M. Shake Farid Uddin and M. A. Sikder, "Heart Disease Prediction using Supervised Learning Classifiers," 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2023."
- [25] "A. A. Almazroi, E. A. Aldhahri, S. Bashir and S. Ashfaq, "A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning,"".
- [26] ""Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press."".
- [27] "Peng, Junjie & Jury, Elizabeth & Dönnies, Pierre & Ciurtin, Coziana. (2021). Machine Learning Techniques for Personalised Medicine Approaches in Immune-Mediated Chronic Inflammatory Diseases: Applications and Challenges. *Frontiers in Pharmacology*. 12. 10.3".
- [28] "Bishop, C. M. (2006). Pattern recognition and machine learning. Springer."
- [29] X. Ying, "An Overview of Overfitting and its Solutions," *Journal of Physics: Conference Series*, no. 1168, 2019.
- [30] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, no. , 6(1), 60–. doi:10.1186/s40537-019-0197-0, 2019.
- [31] "Zhang, Qinghua & Xie, Qin & Wang, Guoyin. (2016). A Survey on Rough Set Theory and Its Applications. *CAAI Transactions on Intelligence Technology*. 1. 10.1016/j.trit.2016.11.001."
- [32] "Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. <https://doi.org/10.24432/C52P4X>."
- [33] "Warsaw University ,<http://logic.mimuw.edu.pl/>»rses ,January 19, 2005".

- [34] "mimuw," [Online]. Available: https://www.mimuw.edu.pl/~szczuka/rses/RSES_doc_eng.pdf.
- [35] "Sumit, Bhatia & Praveen, Prakash & G.N, Pillai. (2008). SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features. Lecture Notes in Engineering and Computer Science. 2173."
-

List of figures

Figure 1: Type of Heart Disease [3]	15
Figure 2: Blood Pressure Measurement [9]	17
Figure 3: Cholesterol and Lipid Profile [10].....	18
Figure 4: Electrocardiogram [11].....	18
Figure 5: Echocardiogram [12].....	18
Figure 6: Coronary Calcium Scan [13].....	19
Figure 7: Universal early heart disease detection model	19
Figure 8:Machine Learning [16].....	29
Figure 9: Overfitting [18].....	32
Figure 10: diagram of rough set [20]	Error! Bookmark not defined.
Figure 11: Training Method.....	43
Figure 12: Description Of The Dataset [28]	43
Figure 13:RSES2 Interface	50
<i>Figure 17: First Architecture</i>	51
<i>Figure 18: Extract rules First Architecture</i>	52
<i>Figure 19: Extract rules First Architecture</i>	52
<i>Figure 20: Extract Rules Second Architecture</i>	Error! Bookmark not defined.
Figure 14: third Architecture	55
Figure 15: Extract Rules third Architecture.....	56
Figure 16: Extract rules third Architecture	57
Figure 21: The RSES2 Program	Error! Bookmark not defined.
Figure 22: Extract Attribute First Variant.....	Error! Bookmark not defined.
Figure 23: Extract Rules First Variant	Error! Bookmark not defined.
Figure 24: The Results Of This Variant.....	67
Figure 25: Confusion Matrix For The First Variant	68
Figure 26: Extract Attribute For The Second Variant	Error! Bookmark not defined.
Figure 27: Extract Rules For The Second Variant.....	Error! Bookmark not defined.
Figure 28: The Results Of This Variant.....	63
Figure 29: Confusion Matrix For The Second Variant.....	64
Figure 30: Extract Rules For The Third Variant.....	Error! Bookmark not defined.
Figure 31: The results Of The Third Variant.....	65
Figure 32: Confusion Matrix For The Third Variant.....	66

List of Tables

Table 1: rough set's example	37
Table 2: Dataset Attributes	44
Table 3: All experimental results	68
Table 4: Comparisons	70

List of Equations

Equation 1: Upper Approximation.....	38
Equation 2: Lower Approximation	38
Equation 3: Lower Approximation example.....	39
Equation 4: Calculating the Mean.....	45
Equation 5: Accuracy.....	47
Equation 6: Coverage.....	48
Equation 7: Class-specific True Positives.....	48
Equation 8: Total Accuracy	48
Equation 9: Total Coverage	48