PREDICTING OUTCOME OF CARDIAC REHABILITATION USING MACHINE LEARNING TECHNIQUE

MUHAMMAD ASYRAFF BIN PONAN

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2023

PREDICTING OUTCOME OF CARDIAC REHABILITATION USING MACHINE LEARNING TECHNIQUE

MUHAMMAD ASYRAFF BIN PONAN

RESEARCH PAPER SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE MASTER OF DATA SCIENCE (COURSEWORK)

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2023

*(Please delete this part) Depending on the medium of thesis/dissertation, pick either the English or Bahasa Malaysia version and delete the other.

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Muhammad Asyraff bin Ponan (I.C/Passport No: 981017-01-5331)

Matric No: S2128251/1

Name of Master: Master of Data Science (Coursework)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Predicting Outcome of Cardiac Rehabilitation using Machine Learning Technique

Field of Study: Healthcare

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature  Date: 2/9/2023

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

UNIVERSITI MALAYA

PERAKUAN KEASLIAN PENULISAN

Nama: Muhammad Asyraff bin Ponan (No. K.P/Pasport: 981017-01-5331)

No. Matrik: S2128251/1

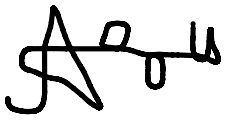
Nama Master: Master Sains Data (Kerja Khusus)

Tajuk Kertas Projek/Laporan Penyelidikan/Disertasi/Tesis ("Hasil Kerja ini"):

Meramalkan Hasil Pemulihan Jantung menggunakan Teknik Pembelajaran Mesin
Bidang Penyelidikan: Kesihatan

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

(1) Saya adalah satu-satunya pengarang/penulis Hasil Kerja ini;
(2) Hasil Kerja ini adalah asli;
(3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan dengan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerja tersebut dan pengarang/penulisnya telah dilakukan di dalam Hasil Kerja ini;
(4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
(5) Saya dengan ini menyerahkan kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya ("UM") yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalam Hasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan dalam apa jua bentuk atau dengan apa juga cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis dari UM;
(6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain sama ada dengan niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan Calon  Tarikh:2/9/2023

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan Saksi Tarikh:

Nama:

Jawatan:

PREDICTING OUTCOME OF CARDIAC REHABILITATION USING MACHINE LEARNING TECHNIQUE

ABSTRACT

A critical stage in the treatment of patients recovering from cardiac surgeries or incidents is cardiac rehabilitation (CR). The possible benefits of improving CR outcomes include improved patient wellbeing and reduced cardiovascular disease burden. To accomplish this, this study makes use of machine learning techniques. Our research is driven by three main goals: finding significant features, creating prediction models, and assessing model performance. The research finds features essential to CR prediction by careful feature selection. Predictive models are built using machine learning techniques such as Random Forest, Support Vector Machine, and XGBoost. Key performance measures are used to carefully evaluate these models. By enabling personalized therapies and enhancing patient outcomes, the study's findings have the potential to revolutionize CR programs.

Keywords: cardiac rehabilitation, machine learning

MERAMALKAN HASIL PEMULIHAN JANTUNG MENGGUNAKAN TEKNIK PEMBELAJARAN MESIN

ABSTRAK

Peringkat kritikal dalam rawatan pesakit yang pulih daripada pembedahan atau insiden jantung ialah pemulihan jantung (CR). Faedah yang mungkin untuk meningkatkan hasil CR termasuk kesejahteraan pesakit yang lebih baik dan mengurangkan beban penyakit kardiovaskular. Untuk mencapai matlamat ini, kajian ini menggunakan teknik pembelajaran mesin. Penyelidikan kami didorong oleh tiga matlamat utama: mencari ciri penting, mencipta model ramalan dan menilai prestasi model. Penyelidikan mendapati ciri penting untuk ramalan CR melalui pemilihan ciri yang teliti. Model ramalan dibina menggunakan teknik pembelajaran mesin seperti Random Forest, Mesin Vektor Sokongan dan XGBoost. Ukuran prestasi utama digunakan untuk menilai model ini dengan teliti. Dengan mendayakan terapi yang diperibadikan dan meningkatkan hasil pesakit, penemuan kajian ini berpotensi untuk merevolusikan program CR.

Kata kunci: pemulihan jantung, pembelajaran mesin

ACKNOWLEDGEMENTS

I acknowledged this research to my supervisor, Dr Kasturi Dewi Varathan, my master by coursework's lecturer and my lovely parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

For examples:

ML	:	Machine Learning
RF	:	Random Forest
XGBoost	:	Extreme Gradient Boosting
SVM	:	Support Vector Machine
MLR	:	Multiple Logistic Regression
CVD	:	CardioVascular Disease
CR	:	Cardiac Rehabilitation
ROC	:	Receiving Operating Characteristic
AUC	:	Area Under Curve
CVD	:	Cardiovascular Disease

CHAPTER 1: INTRODUCTION

The major cause of death worldwide for many years has been cardiovascular disease (CVD), which presents a serious health problem for all countries. This burden on world health is made worse by the long-term trends of an ageing population and rising CVD risk factor prevalence (Khan et al., 2022). The story of cardiovascular disease in Malaysia is like the worldwide story, albeit with local details and difficulties. Like many other developing nations, Malaysia has struggled for years with the widespread impact of CVD. Since the 1980s, when the Malaysian Ministry of Health began meticulously monitoring the development of health conditions in the nation, their reports have continuously identified CVD as the top cause of mortality.

The Malaysian Ministry of Health's figures are disheartening. A shocking 15.0% of the 109,164 medically certified fatalities in 2019 alone were related to coronary artery disease (CAD), a subtype of cardiovascular disease. When compared to all cancer-related deaths combined, which were only a small portion of CAD-related deaths, this number is especially concerning. Such information highlights the significant influence of CVD on the death rates of the Malaysian population (Khan et al., 2022). According to Khan et al. (2022), a concentrated effort is required to fully address this public health concern due to the high prevalence of CVD in Malaysia. Beyond the immediate effects on health, the hospitalization, medication, and rehabilitation expenditures connected with CVD care place a heavy financial load on the healthcare system and the overall economy of the nation.

Cardiovascular diseases (CVD) pose a significant threat to world health due to their complex signs and symptoms as well as complex etiology. A comprehensive strategy is required for the care of CVD, one that goes beyond medical interventions and includes preventive measures, lifestyle changes, and extensive rehabilitation programs. The

recovery and general well-being of people who have had cardiac events or have undergone cardiac surgeries are significantly improved by cardiac rehabilitation (CR), a crucial element of modern CVD care. We set out on a journey to investigate the field of cardiac rehabilitation, highlighting its development, difficulties, and the potential for innovations to increase its effectiveness.

Cardiac rehabilitation (CR) is an important aspect of modern healthcare and is intended to improve the health and functional capacities of those who have undergone cardiac surgeries or incidents. This extensive method of post-cardiac treatment aims to boost physical function, reduce cardiac risks, address return to social responsibilities (Davis, 2020). Recent developments in the integration of data science and healthcare, notably in cardiac rehabilitation, offer the possibility of revolutionizing patient care. We can create precise predictive models by utilizing the strength of machine learning (ML) techniques, which has the potential to revolutionize and personalize cardiac rehabilitation program.

Cardiac rehabilitation (CR), which serves patients who have undergone cardiac surgeries or experienced cardiac events, is an important part of the healthcare continuity. Optimizing the afterwards or post-event recovery process is the main objective of CR, which aims to help patients recover their physical and mental health. This all-encompassing strategy includes a variety of interventions, such as risk factor reduction, psychological support, and education on heart-healthy living (American Heart Association, 2018).

Cardiovascular rehabilitation's core goals include enhancing physical function, lowering the risk of future cardiac events, and addressing the psychosocial components of recovery. Patients improve their physical strength, cardiovascular fitness, and body endurance through specialized exercise program. This physical development is

crucial for lowering cardiac risks and boosting the living standard for those with heart failure. Additionally, CR program offer vital counselling and education to provide patients with the knowledge and skills they need to effectively manage their heart problems (Cardiac Rehabilitation - Mayo Clinic, 2023). In the end, effective CR program execution can result in a decreased mortality rate and greater longevity for cardiac patients (American Heart Association, 2018).

Despite its acknowledged significance, CR has a huge range of obstacles in pursuing its objectives. The decision to enroll a patient in a cardiac rehabilitation program has historically been a medical one, mostly based on the patient's health status and perceived ability for physical activity, as highlighted by Lofaro et al. (2016). Unexpectedly, there are no global standards creating standard guidelines for cardiac rehabilitation, leaving possibilities for variation in the strategy used by various healthcare organizations and providers.

According to Lofaro et al. (2016), Phase I (inpatient cardiac rehabilitation), Phase II (outpatient cardiac rehabilitation), and Phase III (maintenance) are the usual phases of cardiac rehabilitation. An inpatient hospital-based program is the first step. Phase I programs tend to be limited to early mobilization and educating about lifestyle changes and do not include an exercise training component due to short hospital stays and time-consuming assessments. Hospital-based outpatient programs last for two to four months (Lofaro et al., 2016). Phase II cardiac rehabilitation can take many different forms depending on the facility, but it normally consists of individual or group exercise, education, and counselling. Phase III is to preserve the goals set forth in Phase II, reconnect the patient with everyday activities, and promote healthy lifestyle practices (secondary prevention) (Torres et al., 2023).

While there is no doubt that CR has advantages, there are a number of obstacles that prevent patient engagement and program completion, thus its impact is still not entirely optimal. According to De Cannière et al. (2020), problems with referral, enrolment, and inadequate completion rates prevent the majority of patients who are eligible for CR from engaging fully. This problem calls for a strategic revision of the CR strategy, researching cutting-edge approaches to improve participation and long-term adherence.

In order to obtain the best rehabilitation for cardiovascular patients, the majority of whom are located within the health center where the program is executed, Torres et al. (2023) state that a multidisciplinary team is necessary. However, numerous centers have been forced to close areas or restrict non-emergency activities that are not linked to respiratory infection in order to contain the emergence of the coronavirus disease 2019 (COVID-19) pandemic recently. This has had an impact on CVR since the number of sessions has been restricted to minimize contact due to confinements and the risk of infection. Therefore, remote monitoring is an option for maintaining the continuity of rehabilitation programs without raising the danger of COVID-19 infection in patients and staff.

According to Shen et al. (2022), exercise-based cardiac rehabilitation can increase therapeutic results, promote health, and lower the mortality rate of individuals with coronary heart disease (CHD). However, excessive, or wrong exercise may increase your risk of cardiovascular disease. Gas monitoring technology and bicycle ergometer or treadmill exercise technology are both used in cardiopulmonary exercise testing (CPET). The real-time changes in oxygen consumption, ventilation effectiveness, heart rate, and other parameters can be detected by CPET (Shen et al., 2022).

According to Pervaiz et al. (2018), any muscular activity that uses up more energy than when the body is at rest is referred to as physical exercise (PE). PE is a crucial tool

for treating and preventing several non-communicable diseases, including diabetes, cancer, stroke, and cardiovascular disease, according to the World Health Organization. Therefore, PE has been introduced into various rehabilitation programs to aid patients and medical staff in achieving specific rehabilitation goals. The primary goal of PE in rehabilitation is to improve the patients' health-related physical fitness state, which refers to the elements needed to lead a healthy life (Pervaiz et al., 2018). Also highlighted by Pervaiz et al. (2018), for the cardiorespiratory components, aerobic exercises including walking, jogging, and riding a bike are used. Last but not least, vigorous short-duration activities (such sit-to-stand, vertical leaps, or short distance jogging) enhance aerobic capability. As a result, different activities can be used to accomplish different goals. Despite the advantages of PE, there are a few things to keep in mind when using it in rehabilitation because subjecting patients to intense activity and high levels of tiredness could result in physical or physiological issues. Given this, it is currently necessary to create a personalized exercise plan to use PE as a clinical tool and fulfil the various goals of each rehabilitation program.

Adoption of home-based cardiac rehabilitation programs is one strategy that shows a lot of promise in overcoming these obstacles according to De Cannière et al. (2020). These programs make use of remote coaching and monitoring techniques, enabling people to take part in CR from the convenience of their own homes. Such home-based programs, according to De Cannière et al. (2020), can besides increase participation rates but also be a less costly option for traditional in-hospital rehabilitation. Technology becomes a powerful element in shaping CR in this context. By continuously monitoring a patient's disease status at home, remote monitoring strategies, along with digital health solutions, have the potential to improve preventive cardiology practices. Real-time monitoring has the ability to identify early signs of

deterioration, allowing for immediate action and possibly lowering medical costs (De Cannière et al., 2020).

However, the complex nature of CVD makes it extremely difficult to apply effective remote monitoring and engagement approaches. Designing solutions that can effectively address these complex features is essential because cardiovascular diseases cover a wide range of factors, from physiological characteristics to psychological components (Claes et al., 2019).

Additionally, the CR journey involves more than just the clinical aspects. In assessing the patient's participation and adherence to the program, patient motivation is crucial. Jahandideh et al. (2021) stressed the need of in-hospital interventions that encourage people who have had cardiovascular events to start outpatient CR. The success of a program can be greatly impacted by grasping the intentions of these patients and developing treatment accordingly.

To effectively manage the complexities of CR, creative solutions are necessary. Context-aware techniques, as mentioned by Ogbuabor et al. (2020), give a platform to aggregate and correlate different key variables, giving patients and healthcare professionals individualized and useful information. These methods take into account the patient's context, which includes their location, time, identity, and activity, in order provide timely and contextually appropriate therapies.

According to De Cannière et al. (2020), huge volumes of data can be captured thanks to wearable sensor technology, which may then be utilized to tailor healthcare plans to individual patients' requirements as well as the shifting socioeconomic system. The difficulty of accurately interpreting this vast amount of data in a trustworthy and clinically significant way continues. The emergence of data science and its applications have resulted in a significant impact on the healthcare environment recently. The growth of wearable technology, medical imaging data, electronic health records (EHRs), and other sources of healthcare-related data has fueled the change. These extensive and varied data sources have opened new avenues for patient care, diagnosis, and therapy. Data science in healthcare have shown to have a remarkable achievement (Tschuggnall et al., 2021). Recently, data science especially data mining and machine learning shown to have significance contribution towards healthcare sector by cultivating the insights from tons of healthcare data. In the decade preceding, the use of machine learning and deep learning methods in healthcare has increased significantly.

Nevertheless, when making decisions or planning treatments, doctors frequently continue to use conventional techniques (Tschuggnall et al., 2021). Artificial intelligence is already utilized successfully in many areas of medicine as highlighted by Tschuggnall et al. (2021) including prediction of dementia, assessment of mortality risk, identification of Alzheimer’s disease. According to Lofaro et al. (2016), numerous earlier studies investigated the use of data mining techniques for the prediction of cardiac rehabilitation outcome in terms of physical performance as well as length of hospital stay after acute cardiovascular events, but fewer reports are focused on using predictive models to support clinicians in the selection of a patient-specific rehabilitative treatment path.

Data science's power to use massive volumes of data to generate predictions, discover patterns, and enhance decision-making is one of its most promising applications in healthcare. A highly transformational aspect of integration has been the use of machine learning, a branch of data science. Algorithms in machine learning can learn from previous patient data to predict outcomes, detect high-risk patients, and suggest individualized treatment plans. This capability of data-driven decision support has the potential to improve patient outcomes and reduce medical expenses (Obermeyer & Emanuel, 2016).

The convergence of machine learning and healthcare has created groundbreaking possibilities for personalized and predictive treatments. Rajkomar et al. (2018) stated that predictive models that can help physicians make well-informed decisions can be developed using the vast and complex datasets produced in healthcare settings. The use of machine learning models in the diagnosis, prognosis, and planning of treatments for a variety of diseases has been shown successful. These models can identify minor disease indicators, analyze complex patterns in patient data, and accurately predicted patient outcomes. Predictive models can also be continually updated and improved as new data becomes available, ensuring that they remain relevant and efficient over time (Rajkomar et al., 2018).

The use of machine learning in the context of cardiac rehabilitation has a lot of potential. We can create exact models that allow us to personalize CR program to the specific needs of patients by using the power of predictive modelling (Lofaro et al., 2016). These models can take into account a wide range of factors, such as patient characteristics, medical history, and physiological data. These prediction models have a wide range of potential advantages where they can help medical professionals recognize patients who are at greater risk of complications allowing for early interventions

and individualized treatment programs. In addition, by using these models to optimize diet plans and changes in diet and exercise, CR program can be made to be both efficient and secure for certain patients (De Cannière et al., 2020). Additionally, predictive models can support the ongoing monitoring of patient progress, allowing for timely adjustments to the rehabilitation plan as needed. Despite rising evidence that CR has positive health effects, little attention has been paid to quantifying patient performance during rehabilitation and how this may influence outcomes of the cardiac rehabilitation program (Naami et al., 2022).

Corporate IT strategies are paying more and more attention to processing massive amounts of data to assist decision-making processes. Data science is the field that uses mathematical and analytical models and tools to extract important insights from data. The use of project management and process techniques can be beneficial for data science projects. These approaches function as success factors. However, data science teams may find it difficult to adhere rigorously to a project methodology. Process models such as CRISP-DM should be helpful and can be improved by agile approaches.

The integration of machine learning, data science, and cardiac rehabilitation offers an exceptional potential to transform the way we approach post-cardiac care. We can improve the provision of cardiac rehabilitation services, improve patient outcomes, and contribute to the continued development of personalized medicine in the field of cardiovascular health by creating precise predictive models that take a variety of patient-specific parameters into account. This research's motivation is to create a model that can forecast cardiac rehabilitation using the benefits of a machine learning algorithm. Moreover, there will be several machine learning models that can be used in this research such as support vector machine (SVM), XGBoost and Random Forest.

As we learn more about the field of cardiac rehabilitation, it becomes clear that the standard approaches are changing and making way for a time when machine learning and artificial intelligence (AI), along with technology, individualization, and creative techniques, will converge to optimize outcomes. To fully implement this transition, it is necessary to have an extensive understanding of the difficulties, possibilities, and prospective innovations in the area of CR prediction utilizing machine learning and artificial intelligence. We explore these aspects to enlighten the way towards a landscape for cardiac rehabilitation that is more efficient, encompassing, and data driven.

CHAPTER 2: PROBLEM STATEMENT

Cardiovascular diseases (CVD) are a diverse group of medical conditions that continue to present major healthcare challenges. These medical conditions cover a wide variety of clinical symptoms and have a unique challenge because there is no single quantitative measure adequately captures the complicated nature of these diseases (De Cannière et al., 2020). In this context, remote monitoring and outpatient cardiac rehabilitation (CR) therapy stand out as crucial components of contemporary healthcare, addressing the rehabilitation requirements of those undergoing cardiac surgeries or having cardiac events. It is clearly established that CR programs are effective in reducing morbidity, mortality, and healthcare expenses. Optimizing CR outcomes, however, is still a difficult task that requires the use of precise predictive models that can distinguish and predict these outcomes.

When predicting the results of cardiac rehabilitation programs, the multifactorial complexity of CVD presents challenges. Contrary to some medical conditions that have distinct and well-defined metrics, CVDs have a complex clinical landscape where a variety of factors interact to affect how the disease develops (De Cannière et al., 2020) which cause the prediction of cardiac rehabilitation a challenging endeavor.

High dimensionality and complexity are two prominent characteristics of the data produced in the setting of CVDs. These databases cover a wide range of elements, such as demographic, clinical, physiological, and lifestyle-related characteristics. No one quantitative metric can fully represent the complexity of the data required to identify the state of CVDs given their multidimensional nature. Analyzing these high-dimensional datasets presents a huge challenge for researchers and engineers in the domains of machine learning and data mining.

Despite the crucial need for reliable predictive models in CR, this field has received notably little attention from researchers. To the best of our knowledge, cardiac rehabilitation prediction has only been the subject of one research project, carried out by Yuan et al. (2022). The goal of Yuan et al. (2022) was to forecast return to work (RTW) following cardiac rehabilitation, focusing on a particular component of CR outcomes. Even while this study makes a significant contribution, it only provides a limited overview of the possible outcomes of CR, does not address the whole range of rehabilitation success, and does not examine the complicated structure of the process of rehabilitation.

Our grasp of how to accurately forecast and optimize CR outcomes is significantly lacking due to the lack of research in the field of cardiac rehabilitation prediction. It is crucial to understand that CR outcomes cover a wide range of outcomes in addition to RTW, such as increases in physical fitness, decreased cardiovascular risk factors, and improved quality of life (American Heart Association, 2018). Therefore, in order to effectively advance with patient care, a comprehensive strategy for CR outcome prediction one that takes into account the complexity of CVDs and CR is needed.

Most of the current research in the field of cardiac rehabilitation focuses on adherence and motivation concerns, cardiovascular disease prediction, and monitoring patient health status (Torres et al., 2023). These are unquestionably important areas, but the research community is yet to grasp the full potential of predictive models for enhancing CR outcomes. Due to the predominance of this focus, the whole range of CR outcomes, including increases in functional capacity, psychological well-being, and cardiovascular health, are not sufficiently addressed.

Another issue is that majority of the research don't place enough emphasis on feature analysis and selection. Predictive models may contain irrelevant and unimportant

features if thorough feature analysis and selection techniques are not used. In high-dimensional data processing, Cai et al. (2018) emphasize the crucial importance of feature selection because it improves learning efficiency and gets rid of redundant and irrelevant qualities. According to Cai et al. (2018), feature selection refers to the process of obtaining a subset from an original set of features in the datasets in accordance with a specific feature selection criterion, which chooses which features are relevant from the dataset. In order to improve predictive models, systematic feature analysis and selection procedures must be included.

A considerable proportion of research in cardiac rehabilitation choose to not disclose the datasets used and the code employed in the modelling process. This limitation inhibits the replication of study results and restricts the scientific community's capacity to make improvements on prior research. To encourage collaboration and increase accuracy in predictions, it is crucial to ensure transparency in the exchange of data and code.

Datasets for cardiac rehabilitation tend to contain a large number of features, which can include both significant and insignificant information. A large dimensionality of features can result in issues including overfitting and higher computational complexity, which have been addressed in the literature, as mentioned by Cai et al. (2018). The existence of duplicate and unnecessary information might make it more difficult to interpret models, hide crucial patterns, and ultimately prevent the generation of accurate prediction models.

From our view, the lack of utilization or exploration of potentially significant variables and features in the prediction of cardiac rehabilitation outcomes is a notable problem in addition to the issue of feature overload. A prevalent problem is that certain features of clinical or physiological value are left out of prediction models, leaving us

with a limited understanding of the variables affecting the outcome of rehabilitation. Due to the potential for undiscovered vital predictive signals, this limitation could negatively impact the predictive model's accuracy.

In order to improve the prediction accuracy of cardiac rehabilitation models, it is crucial to thoroughly examine and take account of every feature that can have an impact on patient outcomes (Cai et al., 2018). This requires a careful analysis of the dataset and an in-depth comprehension of the clinical applicability of distinct factors. The ability of predictive models to produce accurate and useful predictions can be improved by incorporating these important aspects, which will eventually help patients and healthcare practitioners.

Current cardiac rehabilitation predictive models, in particular XGBoost, have performance metrics that may not be as accurate as intended. A moderate level of predictive ability is shown by the reported performance of r = 0.760 and MAPE = 0.212 (Torres et al., 2023). The correlation coefficient, or r-value, implies a fair relationship between predicted and observed values, although it might not be sufficient to make highly accurate predictions in medical situations. Furthermore, the model's predictions can deviate significantly from the real values, as shown by the high MAPE (Mean Absolute Percentage Error) value of 0.212, demonstrating the need for better prediction accuracy. Current cardiac rehabilitation prediction models perform below expectations for several reasons. The size of the dataset, which is frequently small and restricts the model's capacity to generalize successfully, is one of the main problems. Furthermore, dealing with datasets of various dimensions presents difficulties that make the situation tougher. Furthermore, the lack of open access to these datasets restricts reproducibility and the building of models that can run on larger and more complex datasets.

Given the aforementioned gaps and problems with current research, there is an urgent need to tackle the problem of more accurately and comprehensively predicting cardiac rehabilitation outcomes. The goal of this study is to investigate the varied aspects of CR and use machine learning techniques to create predictive models that are capable of predicting a broad spectrum of outcomes. This project aims to advance the development of predictive accuracy in cardiac rehabilitation by putting into practice reliable feature selection methods. The goal is to optimize CR programs, boost patient outcomes, and promote knowledge of this crucial stage in cardiovascular care.

CHAPTER 3: RESEARCH QUESTIONS AND OBJECTIVES

Table 3.1 Research Questions and Research Objective

Research Questions	Research Objectives
What are the significant features that contribute to the cardiac rehabilitation prediction?	To identify significant features that contribute to cardiac rehabilitation prediction.
What is the suitable machine learning techniques in predicting cardiac rehabilitation using the identified features?	To identify suitable machine learning techniques in predicting cardiac rehabilitation using the identified features.
How to evaluate the chosen model or technique using suitable metrics?	To evaluate the chosen model or technique using suitable metrics

The first objective of this study is to identify significant features that are crucial in predicting cardiac rehabilitation results. This is an essential step since it speeds up the predictive modelling procedure and improves our understanding of the features that affect a patient's rehabilitation outcome.

Datasets for cardiac rehabilitation frequently include a wide range of variables, including clinical, physiological, and lifestyle-related characteristics. Not every feature in this variety of information is equally important or influential in determining the outcome of cardiac rehabilitation (Cai et al., 20218)

Multiple logistic regression is a particular machine learning technique that may be used to systematically assess the importance of each feature (Yuan et al.,2022). This

involves finding out the coefficient values linked to each feature, showing how much of an impact they make to the predictive model. This process results in the selection of a subset of features that are thought to be clinically and statistically significant for predicting cardiac rehabilitation outcomes. In the next steps, predictive models will be built using the identified top features as a base.

Once the significant features have been identified, the second objective revolves around the development of predictive models using machine learning techniques. The purpose of these models is to use data to accurately predict a patient's outcome after cardiac rehabilitation program. Predictive models serve as advanced tools that utilize the features that have been chosen to generate accurate projections of the results of rehabilitation. These models enable healthcare providers to adapt therapies and resources more effectively. Predictive models can be built using a variety of machine learning models, including Random Forest, Support Vector Machine, and XGBoost. The type of data and the literature review determine the methodology selection. It's crucial to test various feature sets in order to find the one that gives the best performance.

The final objective is to assess the performance of the developed predictive models using suitable metrics. This step is essential to evaluate the models' ability to generalize to new and unexplored data as well as their usefulness in clinical situations. Making robust and accurate predictions is the main aim of predictive modelling. Model evaluation helps in checking that created models satisfy this requirement and can be relied upon for decision-making. Several performance metrics, such as accuracy, specificity, sensitivity, and ROC AUC (Receiver Operating Characteristic Area Under the Curve), are commonly used to assess model performance. Each metric provides a unique perspective on how well the model is performing. These three main objectives

make up a structured approach for predicting the outcomes of cardiac rehabilitation using machine learning. Researchers can improve cardiac rehabilitation programs and patients' general wellbeing by first finding significant features, creating predictive models, and then thoroughly analyzing their performance. This systematic approach enables the delivery of individualized and successful rehabilitation programs and promotes the use of evidence-based decision-making in healthcare.

Figures, like tables, are printed within the body of the text at the centre of the frame and labelled according to the chapter in which they appear. Thus, for example, figures in Chapter 3 are numbered sequentially: Figure 3.1, Figure 3.2.

Figures, unlike text or tables, contain graphs, illustrations or photographs and their labels are placed at the bottom of the figure rather than at the top (using the same format used for tables). If the figure occupies more than one page, the continued figure on the following page should indicate that it is a continuation: for example: ‘Figure 3.7, continued’. If the figure contains a citation, the source of the reference should be placed at the bottom, after the label.

To insert label below a figure, click “Insert Caption” under the “References” tab and select “Figure” in the dropdown list. Click “Update Table” to update the List of Figures.

CHAPTER 4: LITERATURE REVIEW

Recent research has focused a lot of attention on the use of machine learning to anticipate the results of CR programs. In this review of the literature, we examine the most significant research in this area, emphasizing the models, procedures, and performance indicators that are employed to forecast the outcomes of cardiac rehabilitation. The following table shows analysis of technique on cardiac rehabilitation based on machine learning model performance.

Table 4.1 Table of Analysis of technique on cardiac rehabilitation based on machine learning model performance.

Reference	What is being predicted?	Dataset availabi-lity	Dataset size	Best model	Result
Lofaro et al. (2016)	Rehabilitation program divided in four different programs (A, B, C, D)	No	129	Lasso	Performance: Accuracy 0.935, Error rate 0.063, Precision 0.941, Recall 0.901
Yuan et al. (2022).	Return to work (RTW)	Yes	929	AdaBoost model, with the top 20 features selected using a RFE method	Performance: ROC AUC 0.924, Accuracy 0.864, Sensitivity 0.928, Specificity 0.733
Torres et al. (2023)	Probability of cardiac rehabilitation	No	207 retrospect ive, 20 prospectiv e	XGBoost	Performance: NMSE= 0.030 ± 0.013, R^2= 0.630 ± 0.189, r= 0.760 ± 0.162, MAE= 0.086 ± 0.021, MAPE= 0.212 ± 0.120
Jahandideh et al. (2021)	Individual intention to engage in outpatient cardiac rehabilitation (CR) programs	No	217	RF-CIT	Performance: Accuracy=0.6916
Okada et al. (2023)	Non-home discharge among acute heart failure patients	No	128,068	The model using 1-SE rule of Lasso regression performs closely the	Performance: E:O ratio 0.974, CITL0 .037, and slope 1.036

				same with model that uses all the 26 variables	
Chicco and Jurman (2020)	Survival of patients with heart failure	Yes	299	Random forests	Performance: MCC 0.384, F1 Score 0.547, Accuracy 0.740, PR AUC 0.657, ROC AUC 0.800
Naami et al. (2022)	1-year major adverse cardiovascular events (MACE) using cardiac rehabilitation performance	Yes	516	Recursive partitioning tree	Performance: ROC 0.994, Specificity 0.996, Sensitivity 0.985
Claes et al. (2019)	Discharged individuals' intention to engage in an outpatient CR program	No	50	SVM that supplementing the baseline data with the system use, employing the data from a 4-week familiarization period	Performance: ROC=0.94, Specificity 0.80, Sensitivity 0.955
Wei et al. (2018)	Automatic enjoyment estimation during an exercise: 5-Likert scale with 1 being not enjoyable at all and 5 being very enjoyable	No	46	Random forests with geometric features	Performance: Accuracy=0.49
Gupta et al. (2020)	Readmission in AMI patients at the time of discharge: (1)	Yes	7018	All models showed similar performance and modest discrimination. No model	Nil

	readmission within 30 days, and (2) readmission within 1-year			outperformed other models by significant differences.	
Hadanny et al. (2022)	1-year mortality after ACS (acute coronary syndrome)	No	9270	RSF	Performance: Harrell's C-index=0.924, Time-dependent IPCW=0.928, Brier score=0.006
Lowres et al. (2020)	Classify the SMS text messages into the 12 categorical variables and whether it need staff review or not	Yes	3118	Ensemble method	Performance: 1.43% false negatives and 16.2% false positives, sensitivity of 93.5%, specificity of 81.3%
Wallert et al. (2018)	iCBT adherence in MI-ANXDEP patients.	No	90	Breiman random forest model with RFE. This model used 56% (19/34) of the provided predictors and performed significantly better than a random model	Performance: Accuracy 0.64
Iwamoto et al. (2020)	Activities of daily living (ADL) dependence of stroke patients	No	994	classification and regression tree (CART)	Performance: Accuracy 0.830, Sensitivity 0.770, Specificity 0.820
Galli et al. (2021)	Prediction of CRT response and prognosis	No	193	Random forest	Performance: AUC=0.81
Ahmad et al. (2018)	1-year survival outcome of the	Yes	44,886	Adaptive lasso	Performance: ROC AUC=0.759

	patients.				Adaptive lasso is well calibrated and appropriately calibrated model fits the red diagonal line (out of sample calibration)
Howell et al. (2021)	Short-term (6-month) response to cardiac resynchronization therapy	No	741	Random forest	Performance: Accuracy 0.9719
Ogbuabor et al. (2020)	Physical activity recognition	No	Not stated	Random forest	Performance: Accuracy 0.9719
Chen et al. (2022)	Poor prognosis at 90-day	Yes	10,967	CatBoost	Performance: AUC 0.839, Accuracy 0.942, PPV 0.660, NPV 0.951, F1-score 0.404, Brier score 0.047
Chen et al. (2022)	Poor prognosis at 90-day	Yes	10,967	CatBoost	Performance: AUC=0.839, Accuracy=0.942, PPV=0.660, NPV=0.951, F1-score=0.404, Brier score=0.047
Pervaiz et al. (2018)	Activity detection (running, walking, upstairs, downstairs, idle)	No	1292	SVM	Performance: Accuracy 0.98795
Shen et al. (2022)	Cardiopulmonary Exercise Testing-Related Cardiovascular Events	Yes	2,455	Logistic regression model with lasso regression (nomogram model)	Performance: Hosmer–Lemeshow test: there was no statistically significant difference between the predicted probability of the model and the actual observed probability. AUC 0.830

A benchmark study for this research by Torres et al. (2023) was focused on forecasting the likelihood of cardiac rehabilitation. Their dataset, which included 227 observations, was used as the basis for comparing different predictive algorithms. Notably, XGBoost was the model that performed the best in this research, with performance metrics are normalized mean squared error (NMSE) of 0.030, an R-squared (R2) value of 0.630, a correlation coefficient (r) of 0.760, a mean absolute error (MAE) of 0.086, and a mean absolute percentage error (MAPE) of 0.212.

A noteworthy study by Naami et al. (2022) examined the effectiveness of cardiac rehabilitation in predicting major adverse cardiovascular events (MACE) after one year. They used a recursive partitioning tree model and multivariable linear logistic regression with a forward conditional strategy. With a receiver operating characteristic (ROC) score of 0.994, specificity of 0.996, and sensitivity of 0.985, this study showed remarkable prediction capacities. This degree of prediction accuracy for MACE outcomes demonstrates the potential of machine learning algorithms to improve patient care and risk assessment.

A thorough analysis of the current literature by Torres et al. (2023) demonstrates that much of the earlier research focused mainly on elements linked to the patient's health status, such as blood pressure monitoring and the detection of cardiovascular disease. CR program patient adherence and motivation have been the subject of certain research. However, only a small number of studies have expressly sought to predict the outcome of rehabilitation. This finding highlights the importance of conducting more study in this understudied field, particularly considering the potential advantages it may have for enhancing patient outcomes and maximizing CR programs.

In their research, Naami et al. (2022) provided an in-depth description of outcome, focusing on the incidence rate of major adverse cardiovascular events (MACE) during 1

year of the CR program. MACE covered a variety of serious incidents, including as fatalities, hospitalizations for heart failure, coronary angiographies, and admissions for transient ischemic attacks (TIA) or cerebrovascular accidents (CVA). The researchers used logistic regression to create a CR-score that measures the chance of these events. A multimodal method to predict cardiovascular outcomes in the context of rehabilitation was provided by this score, which took into account exercise time, speed of work, incline percent, and workload.

The literature review shows the increased interest in applying machine learning approaches to forecast the results of cardiac rehabilitation. While instance studies like Naami et al. (2022) have demonstrated the potential of predictive algorithms to precisely predict major cardiovascular events, benchmark studies like Torres et al. (2023) have established a performance baseline.

Across all literature review, it is observed that Random Forest and XGBoost recorded high performance in terms of accuracy, specificity, sensitivity and ROC AUC. We can consider to include those models and exploring some other models such as support vector machine (SVM) in the modelling process to evaluate the performance of those models using the given datasets in this research.

The field is still developing, and there are many opportunities for further study. There is still potential for greater research on predicting rehabilitation outcomes because most of the existing work focuses mostly on patient health status and adherence. Researchers can continue to develop the field of cardiac rehabilitation prediction by improving prediction models, including various datasets, and integrating innovative techniques.

CHAPTER 5: RESEARCH METHODOLOGY

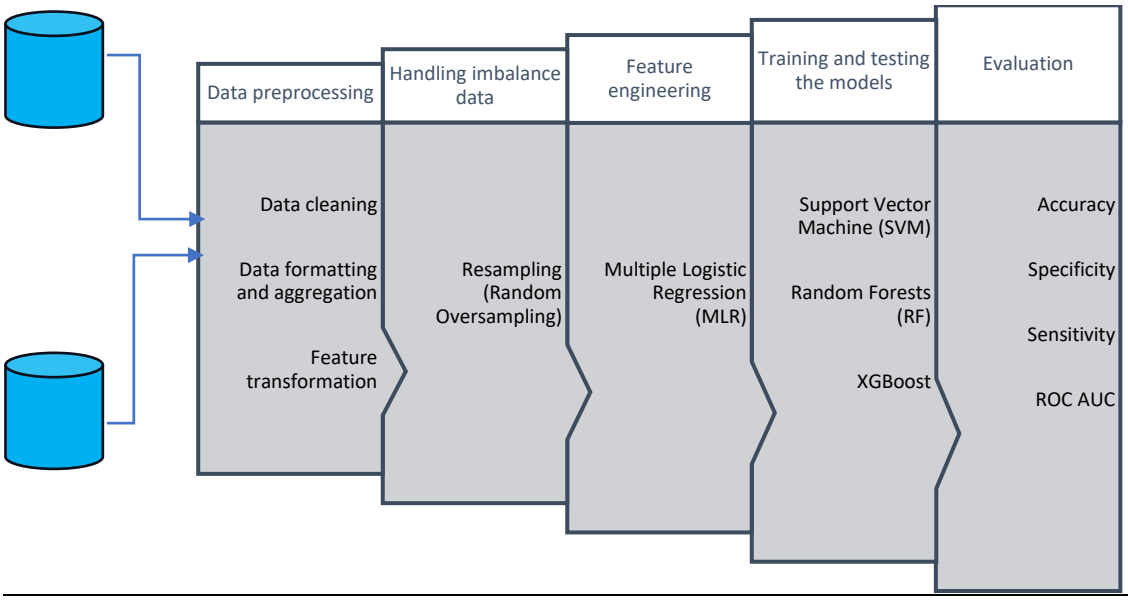


Figure 5-1 CRISP-DM Methodology

For this research, CRISP-DM (cross-industry process for data mining) methodology is used. According to Schröer et al. (2021), CRISP-DM consists mainly of six phases based on the CRISP-DM user guide, which includes business understanding, data understanding, data preparation, modelling, evaluation and lastly, deployment. From Figure 5-1, given the cardiac rehabilitation datasets, business understanding and data understanding process is carried out to determine the research objectives and methodology. The data mining type according to Schröer et al. (2021) was determined as classification. Furthermore, data preprocessing and handling imbalance data is the two steps included in the data preparation process in order to improve the quality of the datasets. Modelling process included machine learning models which are Support Vector Machine, Random Forest and XGBoost. The performance of all models was evaluated in the evaluation process and then the final report is prepared for deployment process.

5.1 Raw datasets

To begin with, this research project is provided with a raw and unprocessed dataset from Malaysia regarding the details and information of patients that are in cardiac rehabilitation process. The exact source of the datasets was not informed. Going deeper into the detail of the dataset, the duration of the datasets was recorded from 2018 to 2020. In total, there were 170 patients in that duration. Some of the patients have no data at all and almost all the remaining patients have missing data in their row. There are 2 different datasets for each year which are **Prestress** and **PT1**. Prestress dataset is mainly about the condition of the patients during the cardiovascular exercise testing together with the preliminary's information of the patients. Excluding **PatientID** column which act like the identifier of the patient, there are 39 other features exist in the dataset. Meanwhile, **PT1** dataset consists of three different sets which are Info which slightly the same with **Prestress** dataset, **Recommendation** which shows the recommendations of each patient under the cardiac rehabilitation program and **Evaluation** which evaluate the patient's condition during and after exercise testing. Under the diagnosis feature of both Prestress and PT1 dataset, it specified the cardiac rehabilitation phase of each patient. Lofaro et al. (2016) explained every single phase involved in cardiac rehabilitation which included 3 main phases: phase I (inpatient cardiac rehabilitation), phase II (outpatient cardiac rehabilitation), and phase III (maintenance). Phase I programmes primarily focus on education on lifestyle changes; an activity training component is not included. Hospital-based outpatient programmes last for two to four months. Phase II cardiac rehabilitation can take many different forms depending on the facility, but it typically consists of individual or group exercise, education, and counselling (Lofaro et al., 2016). Phase 3 of cardiac rehabilitation is a community-based approach with an emphasis on compliance with heart-healthy

behaviour and access to local resources to manage and maintain optimal cardiac health and function (Yuan et al., 2022). The dataset features ranging from various aspects: history of the patients (disease suffered, cardiac attack and complications), the personal background and habits (marital status, family member, lives with, smoking, alcoholic etc.), physical condition (cardiac intervention history, musculoskeletal problem, breathing pattern, electrocardiogram etc.) and exercise testing results and recommendations. Before the research start, National Data Privacy Agreement (NDPA) was signed to protect the privacy of the patients involved in the dataset. The benefit of this research is it can promote the integration of medical field which enriched with daily patients' data and the emerging and advanced machine learning model.

5.2 Data Preprocessing

Since there are several datasets to be worked on, it is better to create a single dataset that unified all the different datasets from different years. Firstly, all datasets from the same year were combined according to the PatientID feature. After that, the datasets from each year; 2018,2019 and 2020 is merged into a single mother dataset. Basically, the mother dataset consists of data of patients from those 3 years with all the features from PT1 and Prestress dataset is collected altogether in a table. All the processes in creating a single reference dataset were mainly done using Microsoft Excel. Furthermore, each feature was filtered and analysed whether they are suitable and compatible to be processed. The next process will be to determine which features to be included and excluded. Almost all numerical features except blood pressure, SPO2 percentage and others were included in the modelling process later. Some of the numerical features were excluded due to the difficulty of the nature of the features and the same features literally have almost the same range of values across all the patients which might have zero impact in the experiment because of there is no significant difference between the patients. In addition, features such as remarks from doctor or physician, plan, past

history, ECG analysis and other features with lengthy text and a lot of abbreviations were excluded from the research due to the nature of the output that were hard to handled and generalized in order to ease the modelling process. Variables with more than 70% missing values also were excluded from the dataset. Next process is feature transformation. For this process, Anaconda Navigator was used which the main language is Python. All the feature processing were done in the IPython notebook in the Anaconda Navigator. Some features such as BMI, Smoking, Alcoholic, Walking and other related can be categorized into several categories to group similar output in a single group by using common keywords in the columns. This process aims to smooth out the modelling process. For example, BMI can be grouped into 3 different group which are ‘normal’, ‘overweight’ and ‘obese’ based on the BMI values. To deal textual output from some of the output, the same process is done where all the similar and slightly the same were generalized to form a new category. For example, PA feature which stands for Personal Activity has tons of different output, but they can be generalized into two different main group which are ‘Active’ and ‘Sedentary’. After feature transformation process, all the categorical features are converted into ranks in form of integers. The following snapshot shows how to carry out the ranking process.

```
smoking_to_integer = {'Non-Smoker': 1, 'Ex-Smoker': 2, 'Smoker': 3}
df2['Smoking'] = df2['Smoking'].map(smoking_to_integer)

breathing_to_integer = {'normal': 1, 'diaphragmatic': 2}
df2['Processed Breathing'] = df2['Processed Breathing'].map(breathing_to_integer)

standing_to_integer = {'good': 1, 'fair': 2}
df2['Balance in Standing'] = df2['Balance in Standing'].map(standing_to_integer)

walking_to_integer = {'normal': 1, 'independent': 2, 'aided': 3}
df2['Walking Group'] = df2['Walking Group'].map(walking_to_integer)

gait_to_integer = {'normal': 1, 'independent': 2, 'slow stepping gait': 3, 'limping':4}
df2['Gait Group'] = df2['Gait Group'].map(gait_to_integer)

dm_to_integer = {'non-diabetic': 1, 'prediabetes': 2, 'diabetes': 3}
df2['DM Category'] = df2['DM Category'].map(dm_to_integer)

activity_to_integer = {'active': 1, 'sedentary': 2}
df2['PA Group'] = df2['PA Group'].map(activity_to_integer)

bmi_to_integer = {'normal': 1, 'overweight': 2, 'obese': 3}
df2['BMI Category'] = df2['BMI Category'].map(bmi_to_integer)

df2
```

Figure 5-2 Ranking Process

Again, ranking process is done basically to aid the modelling process. Next, for numerical features, some of the output of the feature consists of a text embedding the value of the feature. Text filtering was carried out to extract only the value or number that represents the feature and removed all the text embedded to that value. For numerical features with different units of measurement, all the output were converted to the same unit for standardization. For example, exercise duration feature has output with only minutes and minutes and seconds. To standardize, each output with minutes and second is converted to minutes.

After all process of feature transformation has been done, the dataset now is ready to enter next process which is data cleaning. The aim of data cleaning is to deal with missing values. Almost all features have missing values. Therefore, mean and mode imputation were introduced to the dataset to replace those missing values.

```
# Impute missing values
for column in df.columns:
    if df[column].dtype == 'category':  # Categorical column
        mode_value = df[column].mode()[0] # Compute the mode
        df[column].fillna(mode_value, inplace=True)  # Impute missing values with mode
    else:  # Numerical column
        mean_value = df[column].mean()[0]  # Compute the mean
        df[column].fillna(mean_value, inplace=True)  # Impute missing values with mean
```

Now that the dataset is cleaned already, the target variable was identified to be **PA Group.** To clarify, PA Group is a new feature generated after preprocessing of PA features. PA group from my point of view is the best target variable because this research aims to predict the cardiac rehabilitation outcome of each using machine learning and by observing the activity of the patients in their daily life, we can conclude whether the patients are recovering well from the cardiac rehabilitation program or otherwise. As mentioned before, PA Group consists of two category which are ‘active’ and ‘sedentary’. In the preprocessing part, patient who referred as ‘active’ are likely to already RTW (returned to work), doing some physical activities such as walking and running, filling their times with active hobbies such as gardening, cooking and house cleaning and other related activities. On the other hand, ‘sedentary’ referred to the

patients which done almost nothing and being inactive throughout the day. From the cleaned dataset, the frequency of ‘active’ and ‘sedentary’ from a total of 170 patients are 141 and 29 patients respectively. No patients are being excluded from the dataset due to the small size of dataset. Patients with no data in all columns except the PatientID is being replaced using mean and mode imputation that were being implied to the dataset just now. From there, it is concluded that there is a class imbalance in the target variable. Class imbalance might result in misleading interpretation of performance metrics later in the modelling process. Thus, resampling method is carried out. Specific method used for the resampling is random oversampling. Considering that the size of cleaned datasets is small, random oversampling method can increase the dataset size by adding up the number of minority class which is the ‘sedentary’ category.

```
import numpy as np

# Separate the minority and majority classes
minority_class = data[data['PA Group'] == 'sedentary']
majority_class = data[data['PA Group'] == 'active']

# Determine the oversampling ratio (e.g., 2x, 3x, etc.)
oversampling_ratio = 5

# Calculate the number of instances to oversample
num_minority_samples = len(minority_class)
num_samples_to_generate = oversampling_ratio * num_minority_samples

# Randomly oversample the minority class
oversampled_minority = minority_class.sample(num_samples_to_generate, replace=True)

# Concatenate the oversampled minority class with the majority class
df2 = pd.concat([majority_class, oversampled_minority], ignore_index=True)
```

Figure 5-3 Oversampling

5.3 Feature Engineering

Since data preprocessing, data cleaning and resampling method is completed, the new processed dataset are now ready for the feature analysis and modelling process. In this section, every feature in the dataset is analysed to evaluate the coefficient of each feature in predicting the PA Group that indicates the outcome of the cardiac rehabilitation. For the feature selection, multiple logistic regression is used to determine the top predictors due to its popularity among machine learning researchers and have been used in numerous studies too (Yuan et al., 2022). From the process, the top 20

predictors are determined based on the value of the coefficient of each feature. Later, this top 20 features will be utilized in the modelling part.

```
from sklearn.linear_model import LogisticRegression

X = df2.drop(columns=['PA Group'])  # Features
y = df2['PA Group']  # Target variable

model = LogisticRegression()
model.fit(X, y)

coefficients = model.coef_[0]
feature_names = X.columns

coefficients_dict = dict(zip(feature_names, coefficients))
sorted_coefficients = sorted(coefficients_dict.items(), key=lambda x: abs(x[1]), reverse=True)
```

Figure 5-4 Multiple Logistic Regression

5.4 Modelling

Before modelling process, the dataset is first split into train set and test set. The dataset is split into 60% train set and 40% train set. Since the target variable is a categorical variable or to be more specific, a binary feature, classification machine learning model will be fitted into the dataset. Based on literature review of analysis of technique on cardiac rehabilitation based on ML model performance from numerous journal articles and research report, Random Forest, and Extreme Gradient Boosting (XGBoost) will be used in the modelling process because of large number of research that implement both models in their research and results in better performance than any other model. For the exploration purpose, Support Vector Machine (SVM) classifier will used to test the performance of the model and compare it to other two model. After all the model involved in this research is fitted to the cleaned dataset with all the features, the dataset will now be fitted into the same three model as before but using three different set of features:

1. Top 10 features
2. Top 15 features
3. Top 20 features

As mentioned in feature engineering part, this top 10, 15 and 20 features are determined by multiple logistic regression in the feature selection process. The performance of all the set of features will be compared to the previous model that used all the features

involved. The overall performance of each model will be evaluated from all four set of features and the best performance will be highlighted.

5.5 Evaluation Metrics

To evaluate the performance of each model used in this research, four evaluation metrics is used. Since the research is mainly about classification problem, accuracy will be enormously important to identify the percentage of each class of the feature being correctly sorted. Since the class imbalance already been dealt with, accuracy can be calculated without any misleading interpretation. The other three metrics used are specificity, sensitivity, and ROC AUC. These three metrics interrelated to each other. Specificity refers to the rate of which the positive class where in this case indicated by 'Active' output of the target variable PA Group, is correctly identified meanwhile specificity means the opposite from the sensitivity where it calculates the rate of negative class 'Sedentary' being correctly classified. The ROC curve is helpful in choosing the optimal cut-offs for clinical application since it graphically illustrates the trade-off between sensitivity and specificity (Florkowski C. M., 2008). The use of ROC AUC, which considers the performance of both classes by using sensitivity and specificity, will be more suited to assess the models' performance (Yuan et al., 2022).

The formula of each evaluation metrics is provided below:

1. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
2. $Specificity = \frac{TN}{TN+FP}$
3. $Sensitivity = \frac{TP}{TP+FN}$

where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative. For ROC AUC, it is calculated using area under the curve of following curve

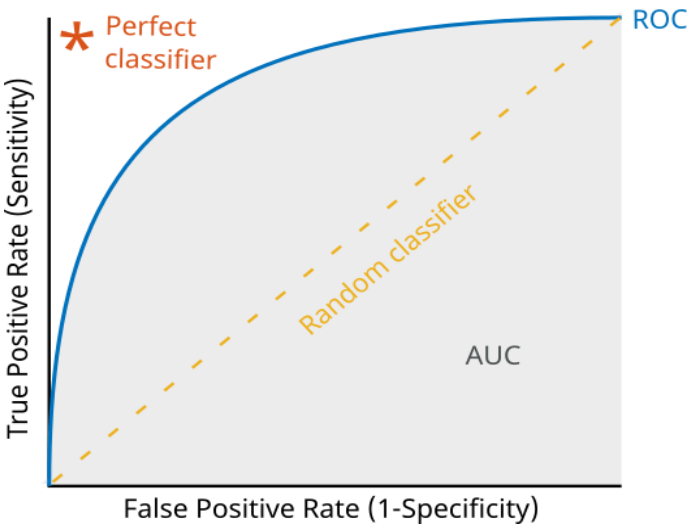


Figure 5-5 ROC AUC

CHAPTER 6: RESULT AND FINDINGS

6.1 Processed and Cleaned Dataset

After all preprocessing, data generalization and data cleaning has been done, the final dataset had 170 patients' data and 27 columns of features including the patient's identifier, PatientID. The full list of 27 features in the dataset are as follows:

```
data.columns
```

```
Index(['PatientID', 'METS', 'Smoking', 'Alcoholic', 'Exercise Frequency',
       'Exercise Duration (minutes)', 'Processed Breathing',
       'UL Muscle Power-Right', 'UL Muscle Power-Left',
       'LL Muscle Power-Right', 'LL Muscle Power-Left', 'Balance in Standing',
       'Walking Group', 'Gait Group', 'Risk Number', 'DM Category', 'PA Group',
       'BMI Category', 'Heart Rate at Rest', 'Peak Heart Rate', 'Processed EF',
       'Processed HR Reserve', 'Heart Rate Recovery', 'Processed MHR',
       'Recumbent Bike Duration in Minutes', 'Max HR during Recumbent Bike',
       'Heart Rate during Evaluation'],
      dtype='object')
```

Figure 6-1 All features in the dataset.

After resampling process is done, the frequency of both class is shown below.

```
class_distribution = df2['PA Group'].value_counts()
print(class_distribution)
```

```
sedentary    145
active       141
Name: PA Group, dtype: int64
```

Figure 6-2 Resampling result.

6.2 Multiple Logistic Regression (MLR)

For the feature selection part, the result of multiple logistic regression is shown in the table below. The top 20 features were determined by evaluating the coefficient of each feature in the logistic regression model.

Table 6.1 Multiple Logistic Regression

Top 20 predictors	Coefficients
BMI Category	0.9750
DM Category	-0.7367
Smoking	0.7366

UL Muscle Power-Right	0.6581
Gait Group	0.5314
LL Muscle Power-Left	-0.5270
Recumbent Bike Duration in Minutes	0.4664
METS	-0.3466
LL Muscle Power-Right	0.3242
UL Muscle Power-Left	0.3097
Processed HR Reserve	-0.2508
Walking Group	-0.2112
Risk Number	-0.1988
Max HR during Recumbent Bike	-0.1842
Balance in Standing	-0.1164
Alcoholic	-0.0634
Processed MHR	0.0584
Processed EF	-0.0510
Heart Rate Recovery	0.0470
Peak Heart Rate	0.0272

From the table above, BMI Category has the highest coefficient value compared to other predictors which indicates that BMI Category has the greatest positive effect to the target variable, PA Group classifier. On the other hand, Peak Heart Rate has the lowest positive effect to the target variable classifier. What can be interpreted from this value of

coefficients is a feature has a positive coefficients value implied that the higher the rank or the lower the value of a feature depending on whether the feature is categorical or numerical respectively will result in higher rank in the target variable. Take note that higher rank does not mean rank with larger number. Rank value 1 is always the highest and in the context of this research, it indicates the best condition of the patients. Take example of BMI category, rank 1 of BMI imply that the patient is in normal BMI and rank 2 and rank 3 indicates patients with overweight and obese BMI. The order of top 20 predictors is also based on the absolute value of the coefficient which means that it neglects whether the coefficient takes positive or negative value.

Meanwhile, negative value of coefficients denotes that the feature is inversely proportional to the target variable. For example, METS is a numerical feature that has negative value of coefficient which imply that higher value of METS that exceed the threshold might result in lower rank of PA Group which is rank 2. Unfortunately, there are some anomaly values of the coefficient in the top 20 predictors for instance the negative coefficient value of Walking Group. Logically speaking, person with poor walking pattern will have a less active lifestyle near to sedentary. But the negative value implied oppositely. This might be explained by the class or category imbalance present in the Walking Group feature that led to such results.

6.3 Modelling

As mentioned before, three machine learning model which are Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) were fitted into the dataset that was split into 60% train set and 40% test set in which the first run of the model is using all the features in the cleaned dataset which consists of 26 features and for the second, third and fourth run of the model, it will involve the top 10, 15 and 20 features in the dataset determined by MLR. The full results of each model are shown in the following tables:

1. Random Forest (RF)

Table 6.2 Random Forest

RF Model	Accuracy	Specificity	Sensitivity	ROC AUC
All features	0.96	0.91	1.00	0.9991
Top 10 features	0.90	0.88	0.93	0.9822
Top 15 features	**0.94**	**0.93**	**1.00**	**0.9994**
Top 20 features	0.92	0.91	0.93	0.9867

From the table, almost all values of evaluation metrics recorded is 0.9 or more. Therefore, the performance of RF is quite excellent across all set of features. The best performance from RF model is using top 15 features suggested by MLR. Although the accuracy of RF with top 15 features is the second best in the table where the highest accuracy score is recorded with all features being included, the third model still performed the best in other three metrics. To highlight, accuracy is one of the best measures to be used in a classification problem, but the other metrics indicate the performance of the model according to the class which referred to how many 'active' and 'sedentary' class is correctly classified.

2. Support Vector Machine (SVM)

Table 6.3 SVM

SVM Model	Accuracy	Specificity	Sensitivity	ROC AUC
All features	0.71	0.81	0.62	0.7601

Top 10 features	0.53	0.32	0.74	0.7018
Top 15 features	0.70	0.77	0.64	0.7532
Top 20 features	**0.72**	**0.74**	**0.71**	**0.7795**

From the table, majority of the performance metrics ranging from 0.60 to 0.70. The highest value recorded was 0.81. It is understandable that the performance of SVM across all set of features in each run is quite fair and above average but not exceptionally good. The best performance from SVM is by using set of top 20 features. This set of features have the highest value of accuracy, sensitivity, and ROC AUC even though specificity is slightly lower than model fitted using all features and top 15 features.

3. Extreme Gradient Boosting (XGBoost)

Table 6.4 XGBoost

XGBoost Model	Accuracy	Specificity	Sensitivity	ROC AUC
All features	**0.90**	**0.84**	**0.95**	**0.9704**
Top 10 features	0.80	0.79	0.81	0.8807
Top 15 features	0.89	0.77	0.93	0.9625
Top 20 features	0.87	0.82	0.91	0.9428

The best performance by XGBoost model is when using the set of all features in the dataset. The performance metrics of model fitted using all features is the highest among other set of features. In summary, the XGBoost model records a very good performance

across all set of features except set of top 10 features where all the performance metrics have the lowest value and the difference between any other set is quite significant.

4. Overall performance of three machine learning models used.

Table 6.5 Overall Performance

Model Involved	Number of Features	Accuracy	Specificity	Sensitivity	ROC AUC
RF	**15**	**0.94**	**0.93**	**1.00**	**0.9994**
SVM	20	0.72	0.74	0.71	0.7795
XGBoost	26	0.90	0.84	0.95	0.9704

By taking the best performance of each model involved and comparing it using performance metrics, it is clear daylight that random forest (RF) has the best performance over Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). Regardless of which set of features involved, all value of performance metrics in the RF table is the highest compared to the SVM and XGBoost table. From the number of features column, RF only needs top 15 features from the dataset determined by multiple logistic regression, which is the least number of features compared to the other two models. The difference of performance metrics of RF and XGBoost is quite close to each other. While not as accurate as Random Forest, Support Vector Machine showed sufficient predictive ability, making it an alternate option. In terms of sensitivity and ROC AUC, XGBoost demonstrated high predictive performance as well, adding to the pool of trustworthy models for cardiovascular rehabilitation prediction. Overall, these findings show how machine learning may be

used to improve cardiac rehabilitation programmes, which will eventually improve patient care and outcomes.

CHAPTER 7: RESEARCH LIMITATIONS AND FUTURE STUDY

In recent years, the medical field has made considerable strides, with machine learning approaches playing a key role in enhancing patient care and results. Predicting the results of cardiac rehabilitation is one such area of application, which can help in adapting rehabilitation programs to specific needs of patients and enhancing the overall efficacy of these interventions. While the current research has provided information on this vital issue, it is crucial to recognize its limitations and suggest potential directions of exploration for future investigations in order to improve the reliability and applicability of prediction models in cardiac rehabilitation.

The size of the datasets used for this study is one of its main constraints. In the context of machine learning, the data used in this study can be considered as quite small, which increases the likelihood of overfitting. A model is said to be overfit when it overlearns the training set of data and underperforms on untrained data. Future research should take into consideration using larger and more diverse datasets to overcome this constraint. In order to train models that are better at generalizing patterns and making accurate predictions, a larger dataset is essential. Additionally, integrating data from several sources or collaborating with other healthcare organizations can result in a dataset which is more insightful.

Mean and mode imputation were used in this study to deal with missing data. Although popular, this technique has disadvantages because it might introduce bias into the research findings. In order to better handle missing data, future studies should investigate advanced imputation techniques, such as multiple imputation or machine learning-based imputation methods. These methods can produce estimates that are more accurate and less biased in order to improve the reliability of prediction models.

This research mainly focused on using coefficients from multiple logistic regression, which is an important step in developing machine learning models. Despite this method having advantages, future research can be gained from an enhanced feature engineering process. By potentially discovering new features or relationships that can enhance model performance, exploratory data analysis (EDA) can be utilized to obtain insightful information from the data. EDA enables researchers to comprehend the data more effectively and facilitates the selection of important features for predictive modelling.

Three machine learning models were taken into account in this research: Random Forest (RF), Support Vector Machine (SVM), and XGBoost. Even if these models are well-known in the industry, there might be others that haven't yet been found or researched that perform better. Future study should focus on investigating and testing a wider variety of machine learning algorithms to find the model that is most effective in forecasting the results of cardiac rehabilitation. In this situation, technologies like neural networks, ensemble methods, or deep learning architectures can be promising.

The next study limitations are the exclusion of textual data, in particular remarks from physicians and each patient's individual rehabilitation plans. Textual information can offer insightful context and additional viewpoints on a patient's health and development. But it was excluded since it was difficult to process lengthy and unorganized text data. The extraction of useful data from textual material using natural language processing (NLP) techniques can be used in future research to build prediction models. Potentially, this inclusion will result in a greater understanding of patient outcomes.

Accuracy, specificity, sensitivity, and ROC AUC were used in this research to assess the model performance. Although these measurements provide insightful information about model performance, future research might expand the list of evaluation metrics to

offer a more comprehensive evaluation. Precision, F1-score, and confusion matrices are the metrics that can provide a broader view of model performance, particularly in healthcare applications.

The development of ensemble models offers exciting potential for future research. Several machine learning algorithms' predictions are combined by ensemble models to produce a single, potentially more reliable predictive model. Ensemble models can reduce the limitations of individual models and boost overall accuracy by leveraging the strengths of various models. In order to develop an ensemble model specifically for predicting the outcomes of cardiac rehabilitation, researchers can investigate a variety of ensemble methods, including bagging, boosting, and stacking.

CHAPTER 8: CONCLUSION

In the entire process of care for patients recovering from cardiac procedures or events, cardiac rehabilitation (CR) is a crucial stage. To improve patients' physical health, lower cardiac risks, and facilitate a successful return to normal life, CR programs must be optimized and personalized to meet the individual needs of each patient. In this study, we set out on an objective to create the predictive model of CR using machine learning model, addressing the obstacles mentioned in the problem statement and evaluating the performance of the models involved.

Our first objective aimed to identify significant features that contribute to cardiac rehabilitation prediction. We used multiple logistic regression to identify the most significant variables because we were aware that CR datasets frequently contain a large number of features, including both relevant and redundant ones. Consequently, we determined the top 20 features that were crucial to our further modelling process. By taking this action, we were able to reduce the issue of feature overload and improve the interpretability of our predictive models (Cai et al., 2018).

Building upon the insights gained from feature selection, our second objective involved the creation of predictive models using machine learning techniques. We used three different models: Random Forest (RF), Support Vector Machine (SVM), and XGBoost. Four different feature sets were used in our strategy: all features, the top 10, the top 15, and the top 20 features. We determined the target variable, "PA Group," and divided observations into the "Active" and "Sedentary" categories. This process was crucial in solving the problem of low model performance and insufficient use of significant features. We aimed to create comprehensive predictive models capable of producing precise and accurate predictions for cardiac rehabilitation outcomes by using various models and varying feature sets.

To determine the practical utility of our predictive models, we understood how important it was to assess their performance. In order to achieve this, we used four performance metrics: ROC AUC, accuracy, specificity, and sensitivity. These measures gave a comprehensive overview of model performance, taking into account aspects such as accuracy, bias, and prediction. Our analysis led us to the conclusion that Random Forest performed the best of the models we studied at. The Random Forest model demonstrated its potential as a useful tool in cardiac rehabilitation prediction with an exceptional ROC AUC score of 99.94% and accuracy, specificity, and sensitivity values of 94%, 93%, and 100%, respectively. These outcomes addressed the issue of inadequate model performance, which had been found, and were in line with the overall objective of improving predicted accuracy in this crucial area of healthcare.

The understanding that the integration of data science and healthcare offers new prospects to revolutionize patient care, particularly in the field of cardiac rehabilitation, served as the foundation for our research path. The problem statement highlighted the difficulties brought on by feature overload, the underuse of critical features, and the pursuit of excellent model performance. These difficulties are a good example of the complicated, high-dimensional, and interpretability-required nature of healthcare data.

In conclusion, considerable advancements have been made in feature discovery, model development, and performance assessment during our work to improve cardiac rehabilitation prediction using machine learning. The field of data-driven healthcare is dynamic and constantly changing, despite the fact that we have made significant progress in addressing the problems identified in the problem statement. To further improve the accuracy and effectiveness of cardiac rehabilitation prediction, future research projects should continue to investigate innovative methods, make use of larger and more diverse datasets, and embrace emerging machine learning techniques.

REFERENCES

Aguirre, A., Pinto, M. H., Cifuentes, C. A., Perdomo, O., Díaz, C. a. R., & Munera, M. (2021). Machine Learning Approach for Fatigue Estimation in Sit-to-Stand Exercise. *Sensors*, *21*(15), 5006. https://doi.org/10.3390/s21155006

Ahmad, T., Lund, L., Rao, P., Ghosh, R., Warier, P., Vaccaro, B. J., Dahlström, U., O'Connor, C. M., Felker, G. M., & Desai, N. R. (2018). Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. *Journal of the American Heart Association*, *7*(8). https://doi.org/10.1161/jaha.117.008081

Bivona, D., Tallavajhala, S., Abdi, M., Oomen, P. J. A., Gao, X., Malhotra, R., Darby, A. E., Monfredi, O., Mangrum, J. M., Mason, P. K., Mazimba, S., Salerno, M., Kramer, C. M., Epstein, F. H., Holmes, J. W., & Bilchick, K. C. (2022). Machine learning for multidimensional response and survival after cardiac resynchronization therapy using features from cardiac magnetic resonance. *Heart Rhythm O2*, *3*(5), 542–552. https://doi.org/10.1016/j.hroo.2022.06.005

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Chen, S., You, J., Yang, X., Gu, H., Huang, X., Liu, H., Feng, J., Jiang, Y., & Wang, Y. (2022). Machine learning is an effective method to predict the 90-day prognosis of patients with transient ischemic attack and minor stroke. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/s12874-022-01672-z

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-1023-5

Claes, J., Filos, D., Cornelissen, V., & Chouvarda, I. (2019). *Prediction of the Adherence to a Home-Based Cardiac Rehabilitation Program*. https://doi.org/10.1109/embc.2019.8857395

Davis, A. T. (2020). Cardiac Rehabilitation. In *Elsevier eBooks* (pp. 678–683). https://doi.org/10.1016/b978-0-323-54947-9.00123-1

Galli, E., Rolle, V. L., Smiseth, O. A., Duchenne, J., Aalen, J. M., Larsen, C. M., Sade, E., Hubert, A., Anilkumar, S., Penicka, M., Linde, C., Leclercq, C., Hernandez, A. C., Voigt, J., & Donal, E. (2021). Importance of Systematic Right Ventricular Assessment in Cardiac Resynchronization Therapy Candidates: A Machine Learning Approach. *Journal of the American Society of Echocardiography*, *34*(5), 494–502. https://doi.org/10.1016/j.echo.2020.12.025

Gupta, S., Ko, D. T., Azizi, P. M., Bouadjenek, M. R., Koh, M., Chong, A., Austin, P. C., & Sanner, S. (2020). Evaluation of Machine Learning Algorithms for Predicting Readmission After Acute Myocardial Infarction Using Routinely Collected Clinical Data. *Canadian Journal of Cardiology*, *36*(6), 878–885. https://doi.org/10.1016/j.cjca.2019.10.023

Hadanny, A., Nagler, A., Wu, J., Gale, C. P., Unger, R., Zahger, D., Gottlieb, S., Matetzky, S., Goldenberg, I., Beigel, R., & Iakobishvili, Z. (2022). Machine learning-based prediction of 1-year mortality for acute coronary syndrome☆. *Journal of Cardiology*, *79*(3), 342–351. https://doi.org/10.1016/j.jjcc.2021.11.006

Hayes, A. (2023). Multiple Linear Regression (MLR) Definition, Formula, and Example. *Investopedia*. https://www.investopedia.com/terms/m/mlr.asp

Howell, S. J., Stivland, T., Stein, K. M., Ellenbogen, K. A., & Tereshchenko, L. G. (2021). Using Machine-Learning for Prediction of the Response to Cardiac Resynchronization Therapy. *JACC: Clinical Electrophysiology*, *7*(12), 1505–1515. https://doi.org/10.1016/j.jacep.2021.06.009

Iwamoto, Y., Imura, T., Tanaka, R., Imada, N., Inagawa, T., Araki, H., & Araki, O. (2020). Development and Validation of Machine Learning-Based Prediction for Dependence in the Activities of Daily Living after Stroke Inpatient Rehabilitation: A Decision-Tree Analysis. *Journal of Stroke and Cerebrovascular Diseases*, *29*(12), 105332. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105332

Jahandideh, S., Jahandideh, M., & Barzegari, E. (2021). Individuals' Intention to Engage in Outpatient Cardiac Rehabilitation Programs: Prediction Based on an Enhanced Model. *Journal of Clinical Psychology in Medical Settings*, *28*(4), 798–807. https://doi.org/10.1007/s10880-021-09771-7

Khan, A. Y. F., Ramli, A. S., Razak, S. A., Kasim, N. a. M., Chua, Y., Ul-Saufie, A. Z., Jalaludin, M. A., & Nawawi, H. M. (2022). The Malaysian HEalth and WellBeing AssessmenT (MYHEBAT) study Protocol: An initiation of a national registry for extended cardiovascular Risk evaluation in the community. *International Journal of Environmental Research and Public Health*, *19*(18), 11789. https://doi.org/10.3390/ijerph191811789

Lofaro, D., Groccia, M. C., Guido, R., Conforti, Caroleo, & Fragomeni. (2016). Machine learning approaches for supporting patient-specific cardiac rehabilitation

programs. *2016 Computing in Cardiology Conference (CinC)*, 149–152. https://ieeexplore.ieee.org/document/7868701/metrics#metrics

Louridi, N., Douzi, S., & Ouahidi, B. E. (2021). Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00524-9

Lowres, N., Duckworth, A. D., Redfern, J., Thiagalingam, A., & Chow, C. K. (2020). Use of a Machine Learning Program to Correctly Triage Incoming Text Messaging Replies From a Cardiovascular Text–Based Secondary Prevention Program: Feasibility Study. *Use of a Machine Learning Program to Correctly Triage Incoming Text Messaging Replies From a Cardiovascular Text–Based Secondary Prevention Program: Feasibility Study*, *8*(6), e19200. https://doi.org/10.2196/19200

Naami, R., Naami, E., Omari, T., Lowi, S. G., Natanzon, S. S., Patel, V., Lerner, A., Rozner, E., Turgeman, Y., & Koren, O. (2022). Cardiac rehabilitation performance predicts 1-year major adverse cardiovascular events. *Clinical Cardiology*, *45*(10), 1036–1043. https://doi.org/10.1002/clc.23890

Ogbuabor, G. O., Augusto, J. C., Moseley, R., & Van Wyk, A. (2020). Context-aware Approach for Cardiac Rehabilitation Monitoring. *Context-Aware Approach for Cardiac Rehabilitation Monitoring*. https://doi.org/10.3233/aise200039

Okada, A., Kaneko, H., Konishi, M., Kamiya, K., Sugimoto, T., Matsuoka, S., Yokota, I., Suzuki, Y., Yamaguchi, S., Itoh, H., Fujiu, K., Michihata, N., Jo, T., Matsui, H., Fushimi, K., Takeda, N., Morita, H., Yasunaga, H., & Komuro, I. (2023). A machine-learning-based prediction of non-home discharge among acute heart failure patients. *Clinical Research in Cardiology*. https://doi.org/10.1007/s00392-023-02209-0

Pervaiz, U., Khawaldeh, S., Aleef, T. A., Minh, V. T., & Hagos, Y. B. (2018). Activity monitoring and meal tracking for cardiac rehabilitation patients. *International Journal of Medical Engineering and Informatics*, *10*(3), 252. https://doi.org/10.1504/ijmei.2018.093365

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q. V., Litsch, K., . . . Dean, J. M. (2018). Scalable and accurate deep learning with electronic health records. *Npj Digital Medicine*, *1*(1). https://doi.org/10.1038/s41746-018-0029-1

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shen, T., Liu, D., Lin, Z., Ren, C., Zhao, W., & Gao, W. (2022). A Machine Learning Model to Predict Cardiovascular Events during Exercise Evaluation in Patients with Coronary Heart Disease. *Journal of Clinical Medicine*, *11*(20), 6061. https://doi.org/10.3390/jcm11206061

Shen, T., Ren, C., Zhao, W., Tao, L., Xu, S., Zhang, C., & Gao, W. (2022). Development and Validation of a Prediction Model for Cardiovascular Events in Exercise Assessment of Coronary Heart Disease Patients After Percutaneous Coronary Intervention. *Frontiers in Cardiovascular Medicine*, *9*. https://doi.org/10.3389/fcvm.2022.798446

Sherman, E., Alejo, D., Wood-Doughty, Z., Sussman, M. S., Schena, S., Ong, C. S., Etchill, E., DiNatale, J., Ahmidi, N., Shpitser, I., & Whitman, G. J. (2021). Leveraging Machine Learning to Predict 30-Day Hospital Readmission After Cardiac Surgery. *The*

Annals of Thoracic Surgery, *114*(6), 2173–2179. https://doi.org/10.1016/j.athoracsur.2021.11.011

Torres, R., Zurita, C., Mellado, D., Nicolis, O., Saavedra, C., Tuesta, M., Salinas, M., Bertini, A., Pedemonte, O., Querales, M., & Salas, R. (2023). Predicting Cardiovascular Rehabilitation of Patients with Coronary Artery Disease Using Transfer Feature Learning. *Diagnostics*, *13*(3), 508. https://doi.org/10.3390/diagnostics13030508

Tschuggnall, M., Grote, V., Pirchl, M., Holzner, B., Rumpold, G., & Fischer, M. (2021). Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. *Informatics in Medicine Unlocked*, *24*, 100598. https://doi.org/10.1016/j.imu.2021.100598

Wallert, J., Gustafson, E., Held, C., Madison, G., Norlund, F., Von Essen, L., & Olsson, E. J. (2018). Predicting Adherence to Internet-Delivered Psychotherapy for Symptoms of Depression and Anxiety After Myocardial Infarction: Machine Learning Insights From the U-CARE Heart Randomized Controlled Trial. *Journal of Medical Internet Research*, *20*(10), e10754. https://doi.org/10.2196/10754

Wei, H., Moran, K., & O'Connor, N. E. (2018). *Automatic Estimation of Enjoyment Levels during Cardiac Rehabilitation Exercise*. https://doi.org/10.1145/3264996.3265003

Yuan, C. J., Varathan, K. D., Suhaimi, A., & Ling, L. W. (2023). Predicting Return to Work after Cardiac Rehabilitation using Machine Learning Models. *Journal of Rehabilitation Medicine*, *55*, jrm00348. https://doi.org/10.2340/jrm.v54.2432