Politécnico
de Viseu

Escola Superior
de Tecnologia
e Gestão de Viseu

Machine Learning Models for Detecting Chronic Disease Progression Based on VO2max: A Study Applied to Cardiovascular Diseases and Type 2 Diabetes

Eduardo Manuel Abreu Pina

Trabalho de Projeto

Mestrado em Engenharia Informática - Sistemas de Informação

Trabalho efetuado sob a orientação de
Professor Doutor Rui Pedro Duarte

Dezembro de 2025

Politécnico
de Viseu

Escola Superior
de Tecnologia
e Gestão de Viseu

Machine Learning Models for Detecting Chronic Disease Progression Based on VO2max: A Study Applied to Cardiovascular Diseases and Type 2 Diabetes

Eduardo Manuel Abreu Pina

Trabalho de Projeto

Mestrado em Engenharia Informática - Sistemas de Informação

Trabalho efetuado sob a orientação de

Professor Doutor Rui Pedro Duarte

Dezembro de 2025

Success is going from failure to failure without losing your enthusiasm.

Winston S. Churchill

Acknowledgements

The path to complete this thesis has been fought with both professional and personal challenges. I am deeply grateful to the Escola Superior de Tecnologia e Gestão de Viseu for providing me with a smartwatch, which allowed me to develop this project. I would also like to express my appreciation to my advisor, Professor Rui Duarte, for his belief in my capabilities and academic potential to undertake this project with confidence and commitment. His guidance and support enriched my research experience, making this journey not only possible but also enjoyable.

The author acknowledges the use of the Open Source Python Garmin Connect API, licensed under the MIT license, without limitations on software restrictions.

I am also grateful to my friends, who provided me with distractions when needed and encouragement when it seemed impossible to continue. A special thanks to Afonso and Henrique for the discussions and help provided throughout this project.

Lastly, i would like to thank my family for their continuous support. Their belief in me kept my spirit and motivation high during this process. I would also like to thank my dog for all the entertainment and emotional support he gave me during this time.

Abstract

Chronic diseases, namely cardiovascular diseases and type 2 diabetes, remain the leading causes of mortality and morbidity around the world. The early detection and prevention of these conditions are crucial to improve health outcomes and reduce the burden on healthcare systems. Among existing health biomarkers, VO_2max is a key feature traditionally used to evaluate individual performance during high intensity activities. Moreover, there have been associations between VO_2max levels and the presence of CVD and T2D. The appearance of wearable devices, such as smartwatches, enables the measurement of VO_2max and other biomarkers, in a non-invasive way, providing real time data for a reliable and validated set of information. The data obtained allows the implementation of machine learning models capable of predicting patterns and chronic disease progression. To this end, the project acquisition of data using a Garmin Venu 2 Plus smartwatch, led to the creation of four datasets, with the help of OpenAI ChatGPT, due to the limited real world data collected for CVD and T2D assessment. These datasets comprised of data derived from the smartwatch and fully synthetic data to be benchmarked under different risk distributions. The feature selection technique using Recursive Feature Elimination identified VO_2max, heart rate variability, sleep time, body mass index, respiration metrics, oxygen saturation and activity features as the most relevant features. The application of two supervised machine learning classification algorithms, SVM and LR, are evaluated, with SVM outperforming LR across both diseases. Notably, although models trained on fully synthetic data showed a strong training performance, datasets derived from smartwatch data demonstrated a better generalisation during testing. Overall, the results confirm the validity and effectiveness of combining non-invasive features with the application of machine learning models for chronic disease risk assessment, highlighting the importance of the application of realistic physiological data in predictive health monitoring systems.

Keywords: *Cardiovascular Diseases* (CVD), *Type 2 Diabetes* (T2D), Biomarkers, *Maximum Oxygen Consumption* (VO_2max), *Machine Learning* (ML), Smartwatches

Resumo

As doenças crónicas, nomeadamente, doenças cardiovasculares e diabetes tipo 2, continuam a ser as principais causas de mortalidade e morbilidade no mundo. A deteção precoce e a prevenção destas condições são cruciais para melhorar o estado de saúde da população e reduzir a carga sobre os sistemas de saúde. Dos biomarcadores de saúde existentes, o VO_2max é uma métrica chave utilizada tradicionalmente na avaliação do desempenho individual durante atividades de alta intensidade. Além disso, cientistas têm encontrado associações entre os níveis de VO_2max e a presença de doenças cardiovasculares e diabetes tipo 2. O aparecimento de dispositivos vestíveis, como os relógios inteligentes, permitem a medição de VO_2max e outros biomarcadores de forma não invasiva, proporcionando dados em tempo real para um conjunto de dados fiáveis e validados. A obtenção destes dados permite a implementação de modelos de *machine learning*, capazes de encontrar padrões e prever a progressão de doenças crónicas. Neste sentido, a acquisição de dados do projeto, através do uso de um relógio inteligente da Garmin Venu 2 Plus, levou à criação de quatro *datasets*, com a ajuda da ferramenta ChatGPT OpenAI, devido à limitação de dados recolhidos no mundo real, para avaliar DCV e DT2. A utilização da técnica de seleção de características, Seleção Recursiva de Características, identificou VO_2max, variabilidade da frequência cardíaca, índice de massa corporal, métricas respiratórias, saturação de oxigénio e características de atividade como os biomarcadores mais relevantes. A aplicação de dois algoritmos de classificação supervisionado, SVM e LR, é avaliada, com o SVM a superar o LR em ambas as doenças. Notavelmente, embora os modelos treinados com *datasets* totalmente sintéticos apresentam um forte desempenho de treino, os *datasets* derivados dos dados do relógio inteligente demonstram uma melhor generalização durante a fase de testes. No geral, os resultados confirmam a validade e eficácia da combinação de características não invasivas, com a aplicação de modelos de *machine learning* para avaliar o risco de doenças crónicas, destacando a importância da aplicação de dados fisiológicos reais em sistemas de previsão de monitorização de saúde.

Palavras-Chave: Doenças Cardiovasculares, Diabetes Tipo 2, Biomarcadores, VO_2max, Machine Learning, Relógios Inteligentes

Contents

List of Tables

List of Figures

Listings

List of Acronyms

Oz	*Ounces*
ACS	*Acute Coronary Syndrome*
AF	*Atrial Fibrillation*
AI	*Artificial Intelligence*
ANOVA	*Analysis of Variance*
API	*Application Programming Interface*
AUC	*Area Under the Curve*
BLG	*Blood Glucose Levels*
BMI	*Body Mass Index*
BMR	*Basal Metabolic Rate*
BP	*Blood Pressure*
BPM	*Beats Per Minute*
CGM	Continuous Glucose Monitoring
CIDR	*Classless Inter-Domain Routing*
CRF	*Cardiorespiratory Fitness*
CV	*Cross Validation*
CVD	*Cardiovascular Diseases*
DT	*Decision Trees*
ECG	*Electrocardiogram*
EDA	*Electrodermal Activity*
FCMIM	*Fast Conditional Mutual Information Maximization*
GCS	*Glasgow Coma Scale*

GPS	*Global Positioning System*
GridSearchCV	Grid Search Cross Validation
GSR	*Galvanic Skin Response*
HF	*Heat Flux*
HOMA-IR	Homeostatic Model Assessment Insulin Resistance
HR	*Heart Rate*
HRV	*Heart Rate Variability*
HTN	*Hypertension*
ICD	*Implantable Cardioverter Defibrillator*
IHD	*Ischemic Heart Disease*
IP	*Internet Protocol*
iPOP	*Integrated Personal Omics Profiling*
JSON	*JavaScript Object Notation*
Kcal	*Calories*
KG	*Kilograms*
KNHANES	*Korea National Health and Nutrition Examination*
KNN	*K-Nearest Neighbors*
L	*Litres*
LASSO	*Least Absolute Shrinkage and Selection Operator*
LLBFS	*Local Learning Based Feature Selection*
LOOCV	*Leave One Out Cross Validation*
LR	*Logistic Regression*
LVF	*Left Ventricle Function*
LVH	*Left Ventricular Hypertrophy*
MI	*Mutual Information*
ML	*Machine Learning*

mL	*Millilitres*
mRMR	*Minimal Redundancy Maximal Relevance*
ms	*Milliseconds*
MYI	*Myocardial Infarction*
NB	*Naive Bayes*
NCD	*Noncommunicable Diseases*
NN	*Neural Networks*
OVR	One Versus Rest
PHI	*Prehospital Index*
PPG	*Photoplethysmography*
RandomizedSearchCV	Randomised Search Cross Validation
RF	*Random Forest*
RFE	*Recursive Feature Elimination*
RHR	*Resting Heart Rate*
RPM	*Respiration Per Minute*
RR	*Respiratory Rate*
Sag	Stochastic Average Gradient
SFS	*Sequential Forward Selection*
SHAP	*Shapley Additive Explanations*
SPO_2	*Blood Oxygen Saturation*
SQL	*Structured Query Language*
SVC	Support Vector Classification
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
SVT	*Supraventricular Tachycardia*
T1D	*Type 1 Diabetes*

T2D	*Type 2 Diabetes*
TimeSeriesSplitCV	Time Series Split Cross Validation
VO_2max	*Maximum Oxygen Consumption*
VAS	*Visual Analogue Scale*
VM	*Virtual Machine*
WHO	*World Health Organization*
WiFi	*Wireless Field*
WoS	*Web of Science*

Chapter 1

Introduction

This chapter highlights the importance of early detection in the management of chronic diseases, providing a comprehensive contextualisation of *Cardiovascular Diseases* (CVD) and *Type 2 Diabetes* (T2D) with an emphasis on their global impact and challenges associated with common diagnostic approaches. Furthermore, this chapter presents key objectives related to data acquisition, feature selection and the development of *Machine Learning* (ML) models for disease predictions in their early stages. The expected results of this project emphasise how the integration of ML algorithms with smartwatch data and biomarkers, like *Maximum Oxygen Consumption* (VO_2max), can improve early diagnosis in healthcare. The work plan provides an overview of this project development to achieve the mentioned objectives. Lastly, the Thesis Structure details the organisation of this project.

1.1 Motivation

Global health and international governments face significant challenges in the fight against the prevalence of chronic diseases. Within the spectrum of the *Non-communicable Diseases* (NCD), CVD, and T2D stand out as the leading causes of death, which places a significant burden on healthcare systems. The creation of innovative solutions has been in demand to prevent such diseases from further spreading and to prevent an even higher rate of mortality. Despite various global initiatives being launched to prevent NCD and mitigate its effects, the results have

been insufficient, in order to achieve their goals. A pressing concern lies in scheduling sporadic appointments by patients. Many healthcare professionals lack access to the most recent clinical data from patients, which, ultimately, limits their accurate assessment of conditions, delays the detection of early diseases and prevents the development of effective treatment plans. Nevertheless, progress in the fields of medicine and engineering represents a significant step forward in the prevention of these diseases, as well as in relieving pressure from healthcare systems.

The use of ML to predict the risks of data collected from wearables, namely smartwatches, allows the early intervention for individuals at risk before the development of the disease. Conventional approaches often fail to fully grasp the data and understand everything it has to offer, especially when faced with vast volumes of data collected by modern devices, including patient health history, diagnostic tests, and biomarker measurements. Under these circumstances, biological markers play a crucial role in prevention and treatment, as indicators, to guide the early detection of chronic diseases. The VO_2max marker, the gold standard metric for assessing *Cardiorespiratory Fitness* (CRF), represents the maximum oxygen uptake during intense physical activity. Its strong association with CVD and insulin resistance highlights its role in understanding and addressing the early progression of T2D. The predictive capacity of this feature in recognising early risk profiles and following disease evolution makes it an important key aspect in research.

Through this approach, ML algorithms offer powerful tools for processing large health datasets, uncovering patterns, and to enhance predictive analysis. With the conversion of raw information into actionable insights, ML supports more accurate diagnoses and effective disease management. Building on this potential, the present project applies ML and advanced analytical techniques to integrate VO_2max with other biomarkers linked to CVD and T2D. The goal is to identify the most informative indicators for early detection and use them to evaluate these conditions in their initial stages.

1.2 Contextualization

As the leading causes of death, worldwide, CVD and T2D have a major impact on both developed and underdeveloped nations [Bernfort et al., 2020, Hossain et al., 2020, Mosenzon et al., 2021, Ehlers et al., 2021]. Cardiovascular disease is responsible for one third of all global deaths [Laranjo et al., 2024]. Since 1990, the number of CVD cases has increased steadily [Roth et al., 2020] with ischemic heart disease and stroke accounting for 65% of CVD cases linked to smoking and high *Body Mass Index* (BMI). Similar results from a study show nearly 20 million deaths reported in 2021, with projections estimating a continued increase in CVD numbers in the coming years [Timmis et al., 2022].

A global initiative, in 2012, aimed to accelerate the progress in the prevention and control of NCD. This initiative aimed to get the collaboration of various member states to adopt policies and programs to improve NCD outcomes by 2025 [Organization, 2023]. Recent updates in the same report indicate that the initiative has not yet met its goals, leading to an extension of its timeline to 2030. Despite promising advancements in the medical field over the past 50 years, at least 80% of CVD cases continue to occur in regions with limited access to healthcare resources, and preventive services pose significant challenges in controlling the diseases [Di Cesare et al., 2023].

Diabetes mellitus, particularly T2D, represents a major health concern due to complications that alter the balance human body. To tackle this disease, collaboration is necessary across various fields, through research, prevention, and care [Abdul Basith Khan et al., 2020], with associations linked to lifestyle and diet characteristics [Abdul Basith Khan et al., 2020]. Data from the *World Health Organization* (WHO) [Organization et al., 2019] report that this health concern is more prevalent among adults, with cases ranging between 90% to 95% detected in low and medium economic countries. Various reports further indicate an increasing number of children and adolescents affected below the age of 25 [Zheng et al., 2018, Group, 2021]. This metabolic disorder occurs when the body can no longer regulate blood sugar levels, because of insufficient insulin production, or the cell's inability to respond accurately to insulin that spreads through other organs [Galicia-Garcia et al., 2020]. A number of reasons that lead to the appearance of T2D, are the intake of highly processed foods, sweetened sugar beverages, elevated BMI, physical inactivity, and sedentary habits [Rutters et al., 2024]. To prevent high levels of blood sugar, insulin treatments help regulate blood glucose and restore glycemic control [Yu et al., 2021, James et al., 2021].

The progression of CVD and T2D links through shared biomarkers that reflect rudimentary metabolic, and physiological imbalances [Balducci et al., 2023]. Among these, high performance athletes and the general public widely use VO_2max as a measure in aerobic exercises. Researchers consider this biomarker the gold standard for assessing CRF, as it commonly serves as a predictive factor for cardiovascular health and morbidity [Harber et al., 2024, Gomez and Giang, 2024]. Additional studies demonstrate a strong association between reduced VO_2max levels and an elevated risk of chronic diseases, particularly CVD, and T2D [Kaminsky et al., 2019, Myers et al., 2021, Petersen et al., 2022, Abbas et al., 2024]. In addition, low VO_2max levels face significantly higher risks of fatal and non-fatal cardiac events, as well as its use as a reliable predictor of functional capacity in patients with heart failure [Ducharme and Gibson, 2021, Mohajan and Mohajan, 2023]. In T2D, lower VO_2max is associated with greater insulin resistance and faster disease progression, whereas regular aerobic exercise that improves VO_2max can enhance insulin sensitivity and

help maintain healthy blood glucose levels [Almutairi et al., 2024].

To evaluate CRF, VO_2max can be measured via two methods, either by direct or indirect approaches [Harber et al., 2024]. A direct approach involves a graded test conducted in a controlled laboratory environment, which measures a person's maximum volume of oxygen by reaching the maximum level of physical activity. Most common tests are performed on treadmills or stationary bicycles to achieve a plateau where the person cannot maintain the current intensity, reaching the peak volume of oxygen. In contrast, the indirect approach does not require the measurement of oxygen consumption during exercise. Alternatively, it uses the heart rate assessments resulting from movement and exercise intensity to estimate the calculation of VO_2max using mathematical formulas in wearable devices or software applications [Gudhmundsson, 2024]. Although direct methods for VO_2max calculation are accurate, they are also more expensive, time consuming to set up, require trained specialists to operate and are typically limited to a small personalized group of individuals. However, indirect methods are more accessible to larger populations and less time consuming. Its accuracy may be lower as it relies on interpolation to estimate VO_2max rather than directly measuring it [Neshitov et al., 2023].

Wearable technology is continuously growing in popularity and predominance in both recreational and professional athleticism [Ometov et al., 2021, Carrier et al., 2023]. According to the report in [Hacker, 2024], the estimated cost of healthcare investments is projected to reach 47 trillion dollars by 2030, reflecting the growing financial demands on global healthcare systems. In the same context, as wearable devices become sophisticated, their interest expands around the world [Lee and Lee, 2020], showing great promise in translating the medical use of clinical settings into wearable devices [Babu et al., 2024].

The concept and design of wearables focus on the attachment to the human body to acquire real time data and continuously monitor the individual's health status in a non-invasive way [Yang et al., 2024]. Several appliances of these devices can vary in sensor placement, such as wrists or the chest. Among these devices, smartwatches can serve various areas, including education, medicine, industry and fitness [Cheng et al., 2021, Moshawrab et al., 2023].

Several brands of smartwatches, such as Apple Watch, Garmin, Samsung Watch and Pixel watches, allow monitoring heart rate, physical activity, among other health tracking tools with integrated sensors. These watches use proprietary algorithms to analyse the data obtained from several sensors, which can be used to make decisions to prevent the early and long term chronic diseases [Hughes et al., 2023]. Some of the variables measured or estimated by current smartwatches include biomarkers such as *Heart Rate* (HR), *Heart Rate Variability* (HRV), VO_2max, *Blood Oxygen Saturation* (SPO_2), BMI and sleep quality [Kumar et al., 2023]. With the emergence of new smartwatch models and the development of scientific studies in this area, these

devices show promising precision results in measuring VO_2max with the assistance of ML algorithms [Lee et al., 2022, Javaid et al., 2022, Molina-Garcia et al., 2022, Saad et al., 2024, Caserman et al., 2024]. Nevertheless, despite their increasing popularity, further research is still needed, since a significant number of smartwatch devices lack rigorous and independent testing to determine their accuracy, reliability, and validity [Miller et al., 2022, Migliaccio et al., 2024, O'Grady et al., 2024].

1.3 Problem Statement

The rapid progress in clinical research presents major challenges in integrating wearable technologies into routine medical practice. Present challenges, such as device accuracy, clinical validity, lack of standardised policies and regulations, with concerns over patient privacy, continue to delay the widespread adoption of wearable devices in healthcare [Bayoumy et al., 2021].

Healthcare individuals continue to rely on sporadic appointments to assess their health status [Amin et al., 2021]. This conventional approach often struggles for medical professionals to detect the early stages of diseases, such as CVD or T2D, without the support of continuous monitoring systems, which in the end makes healthcare professionals overwhelmed.

Advances in *Artificial Intelligence* (AI), particularly ML, have opened new opportunities to improve diagnostic accuracy, and enhance patient assessment. These algorithms process extensive health datasets to identify patterns and make predictions, providing new ways of disease prevention in healthcare [Mathur et al., 2020]. In parallel, the proliferation of wearable devices, such as smartwatches, enables continuous non-invasive monitoring of critical biomarkers like VO_2max. The combination of ML techniques enables smartwatches to contribute to predictive models capable of detecting early signs of chronic disease progression. Integrating smartwatches and ML predictive algorithms shows a path towards early health interventions that can potentially improve patient healthcare.

To this end, this project combines the potential of ML algorithms with VO_2max and other physiological features associated with CVD and T2D collected from smartwatches. Through this work, the project aims to have a significant impact on public health evaluation and contribute to scientific research in the early detection of chronic diseases.

Given the stated problem, we raise several questions:

1. (*RQ1*) – How can we ensure patient data privacy during collection and processing in real time from smartwatches? If so, what mechanisms or techniques can we implement to protect sensitive health information?

2. (*RQ2*) – How can the precision of VO_2max and other physiological indicators measured by smartwatches be validated to ensure reliable forecasts?

3. (*RQ3*) – What is the success rate of ML algorithms applied to predict the early risks of CVD and T2D using observable metrics from smartwatches?

4. (*RQ4*) – How can we train the models to minimise false positives or negatives in early disease progression?

Therefore, it is necessary to focus on five domains: the target population, key health indicators, including VO_2max, feature processing algorithms, predictive algorithms and statistical analysis.

1.4 Objectives

Following the problem specified in the previous section, the main objective of this work is to develop ML models to predict the early progression of CVD and T2D based on data from VO_2max and other biological markers acquired by smartwatches. This main objective can be decomposed into four specific objectives as follows:

1. (*O1*) – Explore the correlation between VO_2max and additional health metrics with the early progression of CVD and T2D risks by identifying predictive indicators.

2. (*O2*) – Familiarise with the technical requirements of wearable healthcare devices and retrieve real time data to apply ML algorithms.

3. (*O3*) – Apply feature selection techniques and ML algorithms to analyse the impact of VO_2max and additional health metrics on the early progression of CVD and T2D.

4. (*O4*) – Evaluate predictive models using performance metrics relevant to CVD and T2D risk assessment.

These objectives will establish a detailed structure for the development and evaluation of predictive models.

1.5 Expected Results

The expected results of this project align with the objectives identified in the previous section. The following outcomes are:

1. (*R1*) – A comprehensive analysis of VO_2max and other health metrics concerning CVD and T2D risks. This analysis will include visualisations and

statistical evidence identifying the most relevant indicators of disease progression, serving as a foundation for feature selection and model building (related to *O1* and *O3*).

2. (*R2*) – A technical framework that outlines the requirements for wearable healthcare devices, including data gathering, real time processing and secure storage from smartwatches. This architecture should support scalability, data integrity, and real time functionality (related to *O2* and *O3*).

3. (*R3*) – A ML pipeline incorporating feature selection and multiple predictive models optimised for chronic disease risk analysis. This pipeline should be tested on smartwatch device data, demonstrating precision and computational efficiency. (related to *O3*, *O4*, and *O5*).

4. (*R4*) – A performance evaluation on the predictive models using metrics. This will validate the model's effectiveness and computational performance (related to *O4*).

1.6 Work Plan

To achieve the objectives mentioned in the previous sections with a focus on the expected results, the development of this work relies on the following steps:

1. **Literature Review:** A detailed review of the current literature regarding VO_2max as a cardiovascular health indicator is performed. The relation between VO_2max levels and the incidence of CVD and T2D is explored, with references to studies highlighting the importance of cardiorespiratory conditioning. In addition, it should focus on ML applications in health, including the prediction of the progression of chronic diseases. Methods such as *Neural Networks* (NN), *Logistic Regression* (LR), *Support Vector Machine* (SVM) and *Random Forest* (RF) are examined, as they will be used to build predictive models based on health data. Finally, the use of wearable devices to continuously monitor VO_2max and other biomarkers will be analysed. Studies validating the effectiveness of wearable devices in health management and real time data collection will be discussed;

2. **Data Collection and Processing:** A smartwatch will continuously monitor VO_2max and other physiological biomarkers such as HR, SPO_2, and sleep quality. Data will be extracted from the Garmin Health *Application Programming Interface* (API) and integrated into a structured database. Complementary health data, such as clinical history and chronic disease diagnoses, may also be provided, with full compliance with privacy and data protection standards;

3. **Predictive Modelling:** The collected data will be used to train ML models. Algorithms such as LR, SVM, NN, and RF may be implemented. Metrics for the models will include VO_2max, *Resting Heart Rate* (RHR), sleep patterns, and other biomarkers from the smartwatch device. Feature selection techniques will be applied to identify the most relevant metrics for predicting disease progression;

4. **Validation and Testing:** A training and a test dataset will be used to evaluate model performance. Evaluation metrics may include accuracy, sensitivity, specificity, and the AUC-ROC curve to assess the model's ability to identify high-risk patients accurately;

5. **Suggested Tools and Technologies:** Python with ML libraries such as scikit learn, TensorFlow, and Keras. The Garmin Health API will be used for data collection. Cloud computing platforms like Google Cloud or AWS are recommended for data processing.

1.7 Thesis/Proposal Structure

The structure of this project includes seven chapters. Chapter 2 describes the bibliographic search process, including the platforms and keywords used to identify relevant articles and journals. Moreover, situates itself within the broader academic context by exploring its relevance, contemporary significance, and the background definition of the concepts mentioned in this study. It also presents the research gaps detected by analysing research studies, while highlighting key areas that may guide this project. Chapter 3 shows the conceptual architecture implemented in this project, outlining the functional and non-functional requirements alongside the implementation of the system, from data acquisition, feature selection, machine learning, and performance metrics. Chapter 4 discusses the results obtained from this study by evaluating the implementation of pre-processing techniques and machine learning models. Chapter 5 synthesises the overall process of implementation and aggregation of healthcare data acquired from smartwatches with the application of supervised machine learning algorithms for classification problems. Lastly, Chapter 6 presents the limitations of the study and future research directions.

Chapter 2

Literature Review

This chapter examines bibliographic research using various search engines to filter relevant conference articles and journals for related work, including definitions, with a focus on ML approaches to predict chronic disease progression and models that account for physiological markers such as VO_2max in conjunction with other biomarkers for the early stages of disease progression. Lastly, it highlights the strengths and gaps of current studies and the essential steps that this project aims to address.

2.1 Bibliographic Research

2.1.1 Methodology used

The investigation process started with the search for articles and journals using multiple scientific research platforms, including Google Scholar, Scopus, Conference Ranks, and *Web of Science* (WoS). We evaluated each selected article and journal by assessing the journal's percentile or associated conference rank. We use Scopus to ensure that the corresponding journal is equal to or above the 75% percentile for journal writing. For conference papers, we verify the associated conference rank ranges within the performance classes A1 (highest), A2, or B1.

2.1.2 Topics and Areas of Research

The use of two search engines requires careful consideration of potentially duplicated articles or journals. To address this, we systematically identify and exclude duplicates. The selection of keywords, for the search, focused on terms directly related to the scope of the work, including "early detection", "smartwatches", "VO_2max", "health features", "machine learning algorithms", "CVD", "T2D", "supervised machine learning", "feature selection techniques", "glucose and insulin prediction" and "continuous glucose monitoring". We consciously exclude works mentioning "long term" and "type 1 diabetes", as this study focuses on the early detection of CVD and T2D, leading to 24 prominent results for full text examination.

2.1.3 Search Methods

The research relied on two scientific search engines, Google Scholar and WoS, illustrated in Figure 2.1. As an initial exploratory search engine, Google Scholar, a well known search engine, provides a wide range of academic resources, including articles, journals and books, among other documents. In addition, it provides extensive indexing, which led to the first option. To complement and reinforce the use of this preliminary bibliographic search, we use WoS. This engine offers a more robust search process that uses keyword queries and filter options to narrow the results of high quality peer reviewed articles and journals.

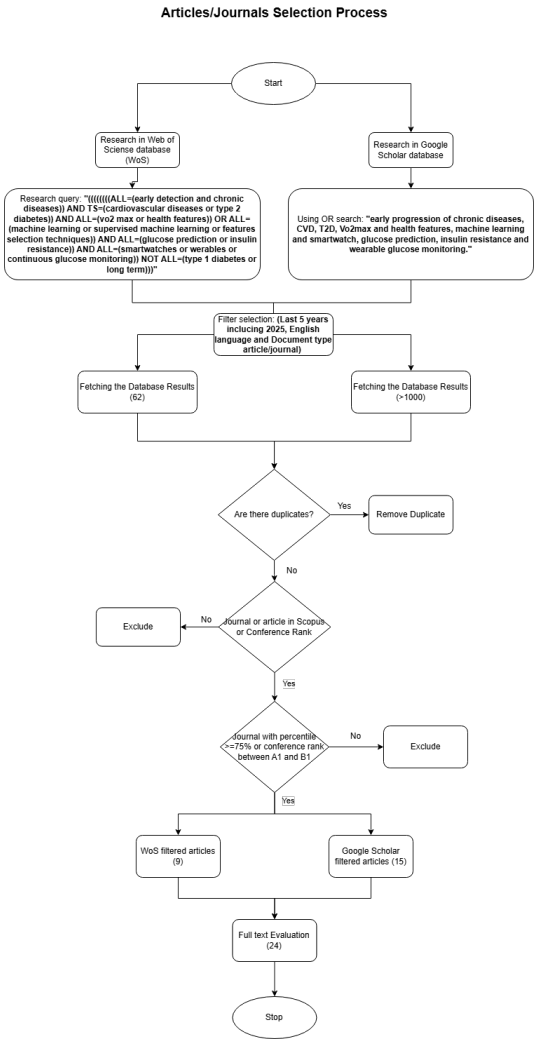


Figure 2.1: Investigation Workflow using two search engines, Google Scholar and Web of Science

This search strategy allowed the selection of high quality sources, forming a solid foundation for the development of the subsequent phases of the system implementation and analysis.

2.2 Related Work

The following subsection presents articles that use data from smartwatches to predict or monitor the progression of CVD and T2D. The focus is on the applied algorithms ML and the strategies adopted to explore physiological data from wearable devices.

2.2.1 Machine Learning Algorithms on CVD and T2D progression using Smartwatch Data

Himi, S. et al. [Himi et al., 2023] present a MedAI system that can predict up to twelve diseases, including CVD and diabetes. This system employs a smartwatch named "Sense O'Clock" with eleven sensors that collect twenty physiological attributes such as HR, SPO_2 levels, body temperature, *Blood Pressure* (BP), and other vital biomarkers. Using this data and ML algorithms embedded in an Android application, the system provides real-time preventive notifications without requiring manual input. The study uses eight algorithms, including SVM [Cortes, 1995], *K-Nearest Neighbors* (KNN) [Fix, 1985, Cover and Hart, 1967] and RF [Ho, 1995, Breiman, 2001]. RF demonstrated an accuracy of 99.4%, which is the highest among the implemented methods. Although this prototype offers a comprehensive range of sensors, high accuracy, and affordability due to its low cost and real-time predictions, which enhance its suitability for daily use, it presents some limitations. The researchers only used 260 records to train the models, which may lead to less reliable predictions and higher error rates. This limitation affects the validity of the metrics, which increases the risk of overfitting, as the models might capture noise rather than generalizable patterns. Lastly, this framework faces substantial data volume and hardware challenges, since predicting twelve diseases requires large volumes of data. This requirement places a significant strain on computational resources and increases hardware complexity.

Similarly, [Kim, 2021] proposes an ML model to predict CVD and diabetes, explicitly targeting smartwatch users. In this study, the authors select nine attributes from the *Korea National Health and Nutrition Examination* (KNHANES) [Control and Agency, 1998] corresponding to metrics measurable by a smartwatch, such as BP, HR, *Blood Glucose Levels* (BLG), BMI, stress levels and demographic variables as age and gender. The model uses three ML algorithms: LR [Berkson, 1953], NN [Rosenblatt, 1958] and SVM. Among these, SVM delivered the best performance in accuracy, precision, recall, and F1-score: 83.04%, 91.96%, 84.81%, and 88.24%, respectively, despite requiring longer computational time due to the complexity of its parameters. The NN and LR also show promising results above 80%, resulting in hopeful ML algorithms for CVD research studies. In the SVM algorithm, the authors used a feature selection technique to identify which features are essential with the help of the Classification and Regression Training (CARET) package [Kuhn, 2008] in R language. This package contains a function called "varlmp", which calculates the feature's importance based on the performance of the proposed SVM algorithm model. The main features selected were age, systolic BP, BLG, and BMI. The least essential feature identified in this study was gender. Although the authors reference Samsung's smartwatch as an example of a device capable of health monitoring, they

did not use it directly in the study. The authors only demonstrated the practical applicability of data gathering. Additionally, using a smartwatch to measure health parameters directly, the authors could use real-time data, which might improve the model's validation and verification. Applying additional parameters available on smartwatches could also enrich the model, making it more robust and ensuring its validity. Another limitation identified in the study is the measurement of stress levels. This variable compromised the model's evaluation, given its subjective nature. However, using these parameter readings from a smartwatch could lead to a better understanding of the value measurements. Finally, during the data preparation phase, the authors reduced the number of participants from 8110 to 6170 due to eliminating missing or abnormal data, resulting in a significant data loss of 24% of its original dataset.

The authors [Dunn et al., 2021] investigate the effectiveness of data collected from an Intel Basis smartwatch in predicting clinical laboratory results using ML models. This study aims to enhance non-invasive health monitoring and diagnostics with particular interest in features like HR, skin temperature, and hydration. The *Integrated Personal Omics Profiling* (iPOP) cohort dataset included 54 participants monitored over an average of 343 days, with various clinical laboratory measurements collected during routine visits. The study employs several ML models, including Lasso Regression and RF, and evaluates them using *Leave One Out Cross Validation* (LOOCV) [Kohavi, 1995]. The findings revealed that smartwatch derived features outperformed clinical data in predicting laboratory results. HR data from smartwatches provide a more reliable and consistent representation of baseline physiology than the higher variability observed in clinic HR measurements. Clinic oral temperature was more stable than smartwatch skin temperature since it varies by the sensor's location in the pulse, environmental conditions, or physiological factors. Furthermore, this work presents several advantages and limitations. A considerable strength of this study lies in the continuous and comprehensive nature of smartwatch monitoring. The ability to capture physiological data over extended periods reduces variability and provides a more precise understanding of an individual's health. Integrating smartwatch data into ML models, mainly RF, reached a higher accuracy. The study highlighted the potential of smartwatch devices for continuous health monitoring and identifying signs of chronic health conditions. However, the study presented several limitations. The dataset included a small sample size, which limits generalizability and may influence overfitting. The models relied on features restricting their ability to adapt to new health indicators. In addition, specific metrics such as skin temperature were sensitive to environmental factors, reducing reliability. Lastly, LOOCV requires a high computational cost since it is the most extreme version of k-fold cross validation.

In [Qureshi et al., 2020], the authors implement a model for an innovative Health

mobile system using ML to predict and classify CVD accurately. In this research, the authors use a dataset combining clinical health records and physiological signals collected from sensors. The smartwatch includes features like HRV, RHR, temperature variability, stress, and cholesterol indicators. The study implemented two ML algorithms, SVM and *Decision Trees* (DT), for classification and risk prediction. The model's performance used accuracy and sensitivity to identify patients with high risk (True Positives) and specificity to identify patients not at risk (true negatives). This study presents advantages and limitations. One advantage is the continuous monitoring of physiological data from smartwatches, enabling early detection of anomalies. The use of real-time data validation and reliability regarding the patient's health conditions using smartwatches and non-invasive smartwatches eliminates the need for invasive tests. Regarding limitations, signals like skin temperature and *Galvanic Skin Response* (GSR) are susceptible to environmental factors, reducing reliability. Signal quality can vary depending on the smartwatch's proper placement. Lastly, one of the common limitations in health is that using small datasets can lead to less generalizable predictive models.

In [Bogue-Jimenez et al., 2022], the authors evaluate the selection of non-invasive features in wrist based wearables, specifically smartwatches, to predict BLG concentrations. The investigated non-invasive features include HR, skin temperature, *Heat Flux* (HF), electrodermal activity, GSR, SPO_2, BP, ambient temperature, and ambient humidity. These features show a great performance in predicting BLG or have a moderate correlation with blood sugar, and their respective sensors can be implemented in a smartwatch like wearable device, enabling continuous daily measurements and rapid commercial adoption. However, this work could be improved in several ways. One is that the dataset was limited in size, with only two participants, which impacted the generalizability of the results. Another strong aspect is the correlation between BP and HR features, suggesting redundancies in implementing real-time devices. Another aspect is the focus on nondiabetic individuals, limiting the application to patients with early risk of T2D, and the lack of hyper and hypoglycemic attributes, which are critical for evaluating the performance of glucose scenarios.

The study [Lehmann et al., 2023] explores the use of smartwatches to develop a non-invasive ML approach to detect hypoglycemia, a common and potential complication in individuals with diabetes. A smartwatch, Garmin vivoactive 4S, is used to collect real-time data. The dataset has key physiological parameters, such as HR, HRV, *Electrodermal Activity* (EDA), and time. It includes 22 adults with diabetes, with only three having T2D. The authors applied a DT ML algorithm to evaluate the model performance. Additionally, the integration of *Shapley Additive Explanations* (SHAP) [Lundberg and Lee, 2017] allowed the authors to highlight that increased HR, decreased HRV, and higher EDA values were strongly associated

with hypoglycemia. Despite showing interesting evaluation possibilities, the study shows some limitations. Regarding the dataset, the individuals included both forms of diabetes, *Type 1 Diabetes* (T1D) and T2D, with unbalanced sizes, given that T2D only had three people. However, the findings have implications for T2D individuals. Furthermore, applying more ML algorithms could improve and compare the study model's reliability with other models.

The following study [Tatli et al., 2024] investigated the validity of detecting the early assessment of T2D using wearable devices. The study combined smartwatch sensors with a Continuous Glucose Monitoring (CGM) system to obtain both behavioural and metabolic data. The authors used two derived feature datasets with related glucose characteristics and accelerometer features focused on activity levels. These multimodal features were aggregated and used as features inputs to a SVM classifier to determine normal glycemic or prediabetic individuals. Although the study presents encouraging classification results above 80%, there are limitations. The dataset presents a small sample size, which limits the diversity and generalisation of the findings. The study, despite using a smartwatch to offer non-invasive measurements, used a CGM glucose device which prevents the system from being fully non-invasive. The feature aggregation altered the temporal acquisition of the data, which changes the interpretation of glucose fluctuations. Furthermore, the collection of data without wear time and missing data further limits the validation of the model. Finally, the study lacks the use of further data, such as HR and *Respiratory Rate* (RR) biomarkers that could add another layer of physiological data to improve and interpret the correlation of features and performance evaluation.

The subsequent subsection, examines relevant studies contributions that addresses CVD progression through ML models that incorporate the domains of feature selection techniques and the hyperparameter tuning.

2.2.2 Machine learning Models on CVD progression using Feature Selection

In [Ghosh et al., 2021], the authors propose a model that predicts CVD with multiple datasets and two feature selection techniques. This approach merges five datasets to establish a diverse training dataset, overcoming common data limitations in medical predictive studies. The authors apply two feature selection techniques: Relief [Kira and Rendell, 1992a, Kira and Rendell, 1992b] and *Least Absolute Shrinkage and Selection Operator* (LASSO) [Tibshirani, 1996]. These two methods identify and prioritize features based on rank values drawn from medical references, effectively minimizing the risks of overfitting and underfitting, which are prevalent challenges in ML models when dealing with clinical data. Relief and LASSO methods demonstrate a distinct advantage in identifying the most essential features instead of

not using them. A highly correlated and optimized feature set enhances the model's interpretability and predictive accuracy. Despite these strengths, the authors acknowledge certain limitations. The model's reliance on the Relief feature selection technique could hinder its generalization when adapted to different datasets or other feature selection methods. Lastly, although the authors use imputation to manage missing data, the model's performance may be affected when faced with datasets containing substantial amounts of missing information.

In [Li et al., 2020], the authors introduce a *Fast Conditional Mutual Information Maximization* (FCMIM) feature selection algorithm [Fleuret, 2004], integrating it with multiple classification algorithms to enhance heart disease diagnosis within electronic healthcare systems. They aim to improve prediction, accuracy, and processing efficiency by implementing various classification methods, including SVM, NN, LR, KNN, *Naive Bayes* (NB) [Bayes, 1968] and DT [Belson, 1959]. The Cleveland Heart Disease dataset [Janosi, 1989] is used as the primary data source. After preprocessing to remove missing values, the dataset comprises 297 samples and 13 selected attributes. These features include vital indicators such as type of chest pain, maximum heart rate, exercise induced angina, resting BP, and thallium status, which the study identifies as critical predictors of heart disease. During the preprocessing step, the authors apply four feature algorithms Relief, *Minimal Redundancy Maximal Relevance* (mRMR) [Ding and Peng, 2005], Least Absolute Shrinkage and Selection Operator, and *Local Learning Based Feature Selection* (LLBFS) [Sun et al., 2009], to identify which features give good results during the application of classification algorithms, in terms of accuracy and computational costs. SVM and LR performed best among the classification algorithms using four feature selection techniques. However, compared to the proposed method, FCMIM-SVM achieved an accuracy of 92.35%. The FCMIM feature selection algorithm reduced model complexity while focusing on the most relevant attributes, which enhanced classification accuracy and reduced computational cost. However, given the study's promising results, there are two main limitations. The dataset size limits the model's generalization, with 297 samples. Lastly, although feature selection optimizes relevance, some selected features may add minimal diagnostic value, potentially introducing noise.

The study [El-Sofany, 2024] focuses on leveraging ML techniques to improve heart disease prediction. A private dataset contains several features like age, gender, BP, cholesterol levels, *Electrocardiogram* (ECG), and chest pain. Researchers employ a combination of feature selection methods and multiple ML classifiers to identify an optimal model for the prediction of early stage heart disease. Three feature selection methods, chi-square [Pearson, 1900], *Analysis of Variance* (ANOVA) [Fisher, 1970] and *Mutual Information* (MI) [Shannon, 1948], were used. Additionally, the study applies several ML classifiers, including NB, SVM, DT, KNN, RF, and LR, to determine the most accurate model. Among the feature selection algorithms,

the one that showed promising performance was ANOVA. This study highlights limitations, including the reliance on private datasets, which limits the validation and reliability of data quality and pose challenges in applying the model across diverse populations. Searching for other data sources and evaluating algorithm choices are recommended.

The study [Liu et al., 2023] explores the application of ML algorithms to predict VO_2max, an important marker for cardiovascular health and chronic disease risk, using non-exercise data. This approach uses publicly available data from NHANES to develop models that can estimate CRF. Notable features in this dataset are age, ethnicity, gender, BMI, height, weight, resting pulse rate, total cholesterol, and BP. The authors employ the SHAP method to identify key contributors to the predictions, ensuring the model selects the most relevant variables. Subsequently, they applied several ML algorithms, including KNN, RF, and *Support Vector Regression* (SVR). This study identifies limitations on the dataset constrained by the age, which ranges from (16-49) years, and excludes the older range of chronic disease risk population, which may restrict generalizability. A cross-sectional design applied to the ML models assumes that the relationships between VO_2max and the selected predictors have remained stable over the past two decades, which may not be valid. Lastly, the absence of real-time data integration and the reliance on indirect measurements of VO_2max reduce the accuracy of the model prediction.

The study [Li et al., 2019] focuses on developing ML models for early prevention of chronic CVD, including stroke and heart failure. The dataset contains demographic and clinical features, including age, gender, smoking history, BMI, HR, diabetes status, and BP. The authors apply a chi-square and SVM-*Recursive Feature Elimination* (RFE) [Weston and Guyon, 2012] algorithm to identify the most relevant features. Subsequently, they employ ML algorithms, including SVM, NB, and LR machine learning algorithms, to predict the outcome. The performance of these models uses metrics such as accuracy, sensitivity, specificity, F1-score, and *Area Under the Curve* (AUC). Among the applied ML algorithms, SVM demonstrated the highest sensitivity and specificity in heart failure predictions, while LR showed promising results, particularly in the detection of strokes. This research presents limitations in the dataset imbalance, primarily due to the small clinical sample size, which narrows the focus on hypertensive patients and restricts the model's generalizability. This research study did not explore other CVD, such as stroke and heart disease, that were analysed. Lastly, even though the subjects of this study are hypertensive patients, the dataset could include features regarding diabetes to enrich the size and diversity of the dataset, as well as the model applicability and analysis to other populations.

The study [Jabeen et al., 2019] introduces a system capable of diagnosing and

managing CVD. This system is capable of classifying eight types of CVD: *Myocardial Infarction* (MYI), *Acute Coronary Syndrome* (ACS), *Atrial Fibrillation* (AF), *Hypertension* (HTN), *Ischemic Heart Disease* (IHD), *Left Ventricular Hypertrophy* (LVH), *Left Ventricle Function* (LVF) and *Supraventricular Tachycardia* (SVT). The dataset comprises one hundred cardiac patients collected under the supervision of cardiologists. Key features include age, gender, BP, HR, chest pain type, ECG, smoking habits, and diabetic history. Then, the authors implement a preprocessing pipeline involving noise reduction, missing values, and feature selection to ensure optimal model performance. A feature selection algorithm called *Sequential Forward Selection* (SFS) [Devijver and Kittler, 1982] where the authors focused on continuously adding best performance features until the accuracy of the results keeps increasing, while stopping when the recognition rate starts to decrease. Then, the features are mapped into training classifiers to identify each disease class. Several supervised ML algorithms, including SVM, RF, and NB, enable classifying patients conditions into eight CVD subtypes. According to the cardiologist's recommendations, the system evaluates performance using precision, recall, and mean absolute error metrics. RF performs better among these ML algorithms with a 98% accuracy. This study presents several advantages and limitations. One considerable strength is the classification of eight types of CVD. Another notable strength is the definition of a recommended lifestyle answer after identifying the disease. Regarding limitations, the proposed system is limited by the data acquisition by a single hospital, without external validity, which limits reproducibility and independent benchmarking. Another aspect is that the dataset is private, which in turn brings security and privacy considerations for sensitive health data, which are not mentioned in this work. With only 100 records, it would be interesting to apply a stronger validation, rather than using 2 fold cross validations, which is not sufficient to be sure that the model actually performed well within different folds. Another aspect is the limited feature space, which could have additional predictors to support the identification of the mentioned diseases. Overall, this work provides an idea to develop a diagnostic tool with the ability to support real-time data collection and scalability to identify and provide insights to improve the lifestyle, given the possibility to identify 8 types of CVD.

The research paper [Bouqentar et al., 2024] addresses the need for early prediction of CVD by developing an ML system optimized for accuracy and reliability. This study identifies and applies several criteria essential for early predictions. The authors use the Cleveland dataset, an internationally known dataset for its balanced and comprehensive health data. Furthermore, the authors apply preprocessing techniques, which include handling missing values and removing duplicates and data types of attributes to improve the model performance. In the development phase, the authors compare six ML algorithms: DT, RF, SVM, LR, and KNN. Performance

assessment uses evaluation metrics such as precision, accuracy, recall, F1-score, and AUC. Afterwards, the authors conduct a feature selection, or hyperparameter tunning process using Grid Search and Random Search [Bergstra and Bengio, 2012]. This procedure ensures a higher accuracy and reduced error during model training and testing. The results highlight SVM and RF as the most effective models, achieving the highest balance between sensitivity and precision, with more than 95% precision in detecting healthy patients and 85% with sick patients. On the other hand, the study presents some limitations. For instance, the dataset only comprises clinical features. Although clinical features are accurate and validated, CVD is approached in various ways and could be included in this dataset to be more robust and diverse. Another example, is that the dataset consists of 303 samples, while the preprocessing method reduces the sample population to 287. Another limitation of this dataset is the lack of validation of the age feature due to the population size. Lastly, it would be interesting to include an ethnic demographic feature to describe each patient for a more precise application of the model to different populations.

In the study [Qadri et al., 2023], the authors investigate how to assess heart diseases with the help of machine learning algorithms while using feature selection algorithms to select the most predominant features. The author's chosen dataset comes from Kaggle [Goldbloom, 2010], a publicly available website containing several datasets for data analysis. This dataset includes 1.025 patient samples with a distribution of 526 samples with heart diseases, making a nearly balanced dataset. Eight of the fourteen features are selected using a novel feature selection technique proposed by the authors, which uses a linear transformation mechanism. These features include age, sex, chest pain, resting BP, cholesterol, fasting blood sugar, maximum HR, and exercise induced angina. This study compared several classification algorithms, including LR, SVM, DT, NB, KNN and RF, regarding machine learning algorithms. Among these ML algorithms, RF and DT present 100% accuracy, while SVM and LR also showed promising results. This study used accuracy, precision, recall, F1-score, and runtime computation for performance metrics. On the other hand, this study presents some limitations. Firstly, the authors do not use a real-time dataset. Using a dataset from Kaggle limits the generalization and the validated results. Another aspect is the lack of explanation of the feature selection technique used, which further limits the validation of the selected features. Lastly, despite obtaining 100% accuracy with DT and RF, given the points stated above raises concerns about the model's robustness and applicability to real-world scenarios.

The subsection below presents relevant approaches and considerations for the application of ML models, including the data types to predict and monitor T2D assessment.

2.2.3 Machine Learning Models on T2D progression

The work [Stolfi et al., 2020] presents an ML model approach to predict T2D risk using synthetic data [Castiglione, 2013]. This dataset comprises 46.170 virtual subjects with the following parameters: gender, age, weight, height, food intake (carbohydrates, proteins, and fats), and physical activity (number of sessions, duration, and intensity). One major disadvantage of this work is that synthetic data does not provide a reliable and validated dataset for other studies, despite its size, given that the data is artificial.

The work [Butt et al., 2021] explores the implementation of ML algorithms for diabetes prediction, emphasizing its role in enhancing healthcare applications. This work used a dataset from the Kaggle repository, the PIMA Indian Diabetes Dataset [Schnor, 2023], which includes attributes such as BLG, BP, BMI, age, insulin, and a binary target attribute to indicate the presence of diabetes. The authors use several ML algorithms, including NB, RF, LR, and SVM, to classify the 768 samples. Among the several performance metrics that exist, this study used precision, recall, and accuracy metrics to evaluate the model's performance. However, this study presents some limitations. This study relies on a dataset that is not collected in real-time through smartwatches, limiting the applicability of the results in a real-world scenario. Another limitation is the missing opportunity to explore wearable devices like smartwatches, given the growing role in diabetes monitoring.

The authors conducted a study focused on early warning of diabetes mellitus, including characteristics and risk factors associated with T2D [Wang et al., 2023]. The dataset consisted of a survey that enrolled 4.886 participants, excluding 95 individuals due to incomplete data, obtaining a final sample of 4.791 participants. Among this cohort, 700 individuals were diagnosed with diabetes. The study identified several significant predictors of diabetes, including age, BMI, hypertension, HR, and obesity. Additionally, family history, hyperlipidemia, and obesity emerged as relevant predictors for T2D.

Recent research has explored non-invasive and wearable approaches that allows the early prediction of insulin resistance to prevent the development of T2D. Traditional methods, such as fasting insulin and glucose tests that are used to calculate Homeostatic Model Assessment Insulin Resistance (HOMA-IR), require blood collection and laboratory analysis, which limits their utility for large scale approaches or continuous monitoring. This study [Metwally et al., 2025] introduces a framework for predicting insulin resistance using non-invasive and wearable devices that offers an alternative to traditional biomedical assessments. The authors recruited 1.100 participants to collect multiple derived physiological signals, focusing on HR, HRV, activity levels, sleep duration and demographic features. The authors state that derived wearable signals provide a strong predictive value even in models with limited clinical input. This highlights the potential of wearable devices for continuous,

non-invasive monitoring to serve as an early warning system for insulin resistance. However, this study presents several limitations. The study focused on HOMA-IR data, which limits the ability to capture temporal causality or progression from insulin sensitivity to resistance. Despite being an indirect measurement of insulin resistance, it can not reflect dynamic insulin glucose relationships as accurately as the gold standard method, such as the glucose clamp technique, which introduces a bias in model towards training and evaluation. Finally, the study lacked additional measurements, such as medication and stress level data, which are known to significantly influence insulin resistance sensitivity.

The following review [Alhaddad et al., 2022] focuses on glucose monitoring research towards non-invasive and wearable technologies capable of detecting and predicting blood glucose using physiological signals, rather than direct invasive sampling. Traditional glucose measurements, like finger tests and CGM, although effective, remain invasive and costly. The review identifies several wearable modalities, coupled with ML algorithms for signal interpretation and glucose classification. Wristbands or smartwatches often comes integrated with *Photoplethysmography* (PPG), to extract cardiovascular and blood readings, such as HRV and blood volume pulse. These features allow to measure the effect of heart rate variation and blood glucose levels as relevant markers for T2D. Similarly, ECG sensors that capture cardiac electrical pulses have shown sensitivity towards glycemic fluctuations. Bioimpedance sensors, typically configured as skin contact electrodes or rings, estimate glucose tissue electrical properties, while electromagnetic sensors and optical systems use reflection and absorption responses at specific wavelengths to indirectly infer glucose concentration. Lastly, accelerometer motion sensors also contribute useful information to detect behavioural patterns that accompany hypoglycemia. Collectively, the information acquired by these derived signal features allow the creation of an input space of biomarkers for ML assessment of glucose classification models. To map these signal features into glucose states, several ML have been explored. Tree based algorithms, like DT and RF are among the most frequently used due to their interpretability and robustness against noisy wearable data, with detection accuracies above 80%. Support vector machines have also classified glucose levels using optical and derived saliva data, with accuracies ranging between 80% and 85%. Although ML and wearable technologies show potential for non-invasive glucose prediction, there are limitations that hinder their use in clinical applications. Physical features extracted from PPG, ECG, or bioimpedance sensors can be influenced by motion, skin temperature, hydration, stress and light factors that can obscure glucose signal changes, which limits the reliability of predictive models. Lastly, reliance on invasive devices for correct labelling of glucose measurements further means that current wearable devices are not yet fully validated for glucose estimation. The authors emphasise the need for multimodal approaches, along with

standard protocols for better data collection, validation and performance evaluation.

The study [Nurmi and Lohan, 2023] presents a machine learning framework for predicting insulin resistance using data from wearable devices combined with routine blood biomarkers, to predict insulin resistance, a key indicator of T2D. The study goal was to explore whether non-invasive or minimal invasive data sources collected through consumer wearables could improve early metabolic risk detection over invasive clinical testing. The dataset consisted of 1.165 participants who provided data, including RHR, HRV, breathing data, step count, sleep duration, BMI, activity levels and clinical blood biomarkers. This study presents several limitations that constrains the model application in the real world. Regarding the dataset, only a quarter of the participants had complete wearable and biomarker data, which creates possible selection bias. Another limitation lies in the reliability of invasive insulin resistance methods, such as the glucose clamp technique, which is expensive and invasive. Furthermore, the study design limits the insight into how insulin resistance evolves over time with short term signal variability introducing noise into the model. Finally, an imbalance of demographic population may reduce generalizability to diverse populations.

The authors [Huang et al., 2025] investigated the potential of non-invasive wearable sensors combined with ML to predict glucose levels on healthy individuals. The study used two wearable devices that combined physiological and behavioural signals, alongside a CGM to provide glucose values. Some of the features include HR, temperature, blood volume pulse, BMI, RR, caloric levels and demographic values. This study presents various limitations. The sample population consisted solely of healthy individuals, without analysing the conditions for those with diabetes. Furthermore, the model requires external validation across a higher demographic diversity or real world uncontrolled context to determine if the model has a good performance evaluation.

Lastly, the following subsection reviews the use of biometrics relevant to the medical field over the clinical assessment of CVD and T2D progression. Furthermore, includes useful performance metrics, capable of assessing the predictive strength and reliability of the models applied to the mentioned chronic disease conditions.

2.2.4 Relevant Metrics in Medical Machine Learning

The authors [Jiang et al., 2021] analyse a dataset comprising 17.661 emergency department visits from the Second Affiliated Hospital of Guangzhou Medical University, covering the period from August 2015 to December 2018. The dataset includes patients aged fourteen and older presenting with suspected CVD. Key data features consisted of age, gender, vital signs (temperature, pulse rate, systolic and diastolic BP, RR, SPO_2, BLG, the *Visual Analogue Scale* (VAS) pain score, the *Glasgow*

Coma Scale (GCS) score and the *Prehospital Index* (PHI). The authors effectively trained ML models, including Multinomial LR and RF, to classify patients into different triage levels. The study encountered several limitations. The exclusion of cases with more than half the data missing could have introduced selection inclination, though comparisons indicated that the analytic and non-analytic cohorts were generally similar. The research execution at a single institution without external validation constrains the generalisation of the findings to other settings. Lastly, the study focused solely on data available at triage, omitting comprehensive clinical and socioeconomic data that could have enhanced the model's accuracy.

Another research [Khan and Algarni, 2020] focused on identifying key characteristics of heart disease prediction using a proposed ML algorithm. The datasets used for this prediction are composed of several attributes: Age (medium, old, very old), gender, type of chest pain (typical angina, atypical angina, non-anginal pain, asymptomatic), resting BP (Low, medium, high, very high), serum cholesterol (low, medium, high, very high), BLG (higher than 120 mg/dl - yes or no), ECG results (normal, having ST-T wave abnormality, left ventricular hypertrophy), maximum HR (low, medium, high, very high) and thallium scan (normal, fixed defect, reversible defect). Another aspect of this work is the use of metrics for performance evaluation, such as precision, accuracy, sensitivity, and specificity.

Numerous studies that investigated the potential of ML algorithms in assessing the early progression of chronic diseases, such as CVD and T2D, reveal key insights and gaps that can guide the trajectory of this project.

2.3 Research Gaps

The studies reviewed above show critical gaps that the current project can tackle to further enhance the investigation process for the development of ML models in the early detection CVD and T2D.

A critical point in the development of this project is improving the dataset comprehensiveness and quality, as studies [Qureshi et al., 2020, Stolfi et al., 2020, Kim, 2021, Dunn et al., 2021, Bogue-Jimenez et al., 2022, Himi et al., 2023] highlight limitations posed by small and selective datasets. This project could incorporate a diverse real-time dataset that includes VO_2max metrics and additional biomarkers from smartwatch devices. This approach could improve the understanding of chronic disease progression, where implementing real-time data would enable reliable and validated predictive capabilities, leading to Objectives *O1* and *O2*.

Feature selection is another vital area for optimising model performance, where studies [Jabeen et al., 2019, Li et al., 2020] and [Ghosh et al., 2021] demonstrate that selection techniques such as Relief, Lasso, SFS, and RFE [Guyon et al., 2002] can

reduce model complexity and improve model performance. Using feature selection to prioritise metrics directly associated with CVD and T2D risks, with VO_2max and related biomarkers in mind, this project could enhance the model interpretability and reduce unnecessary computational load, leading to Objective *O3*. Furthermore, the implementation of SVM and LR algorithms is essential, as these algorithms achieve high performance in chronic disease predictions across multiple studies [Li et al., 2019, Qadri et al., 2023, Bouqentar et al., 2024], and could be suitable for identifying patterns with VO_2max data. By deploying both algorithms in this project, a baseline for precision and reliability in disease progression could be established, reinforcing Objective *O3*.

Finally, studies [Khan and Algarni, 2020] and [Jiang et al., 2021] emphasise the importance of accuracy, recall, precision, F1-score and AUC to assess model efficiency. This project, with the use of these performance metrics, can provide a strong assessment of chronic disease prevention that aligns with real world health monitoring requirements, leading to objective *O4*.

2.4 Definitions

This section presents CVD and T2D concepts using mathematical formulas to understand their role in today's healthcare, alongside performance metrics used in the implementation of machine learning models.

Diabetes mellitus type 2 is a chronic disease in which the body presents high sugar levels due to insulin resistance. This disease occurs when the body can no longer act to the reaction of this hormone, increasing the difficulty of absorbing the sugar in the blood and transforming it into energy. A common approach to quantify insulin resistance in a non-invasive way is the use of the HOMA-IR formula [Matthews et al., 1985]. The mathematical calculation of this formula is as follows:

$$HOMA-IR = \frac{FastingInsulin(\mu U/mL) \times FastingGlucose(mmol/L)}{22.5} \quad (2.1)$$

This formula calculates the product of fasting insulin and fasting glucose and normalises it using a constant. This value originates from typical fasting levels in healthy individuals, around 5µU/mL for insulin and 4.5 mmol/L for glucose, resulting in the 22.5 value. This normalisation helps standardise the HOMA-IR insulin resistance values among different individuals.

The VO_2max consists of the oxygen capacity that our body manages to capture, transport, and use to supply the muscles during high intensity aerobic exercises, to achieve a plateau where the human body can no longer maintain the current intensity, reaching the maximum volume of oxygen. One possible approach to calculate

VO_2max in a non-invasive way is a formula focused solely on heart rate measurements [Uth et al., 2004].

$$VO_2\ max = 15 \times (HeartRate\ max / HeartRate\ rest) \tag{2.2}$$

This formula calculates VO_2max by estimating the maximum heart rate divided by the resting heart rate. This value represents the heart rate reserve, where a high ratio is associated with better cardiovascular fitness. The constant represents the best fit to scale the heart rate ratio to match the actual VO_2max values in ml/kg/min. Given this formula, the higher the VO_2max levels, the greater the individual aerobic capacity and overall cardiovascular fitness, while lower levels show a correlation with the progression of chronic diseases such as CVD and T2D.

The F1 score measures the harmonic mean of precision and recall. This metric balances the importance of precision and recall and is preferable to accuracy when dealing with imbalanced datasets. It also provides a balanced trade-off between false positives and false negatives. The mathematical formulas is given by:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

When both precision and recall have perfect scores of 1.0, this score will also have the same perfect score of 1.0. If precision and recall are close in value, F1 score will be close to their value. When precision and recall are far apart, the F1 score will be similar to which metric presents the worst value.

Precision measures the proportion of all the model's positive classifications that are correctly identified as positive. Its formula is defined as:

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2.4}$$

A higher precision indicates that when the model predicts a sample, belonging to a given class, it is usually correct.

Recall, or Sensitivity, represents the proportion of all actual positives that were classified correctly as positives. Its formula is described as:

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.5}$$

It assesses the ability to detect true positives and is especially important when missing positives cases carries a high cost. A higher recall indicates that the model is effective in capturing one or more relevant classes.

The balanced accuracy accounts for the presence of imbalanced classes on the dataset, where it computes the average recall across all classes. Each class contributes equally to the final metric, regardless of how many samples it contains. The formula is defined as:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} + \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}\right) \tag{2.6}$$

This formula is a reliable indicator of model performance in scenarios where some classes are under represented.

Given the relationships between VO_2max, CVD and T2D progression, the following section presents studies focused on predictive analysis to assess disease risks and early stages of cardiovascular and insulin resistance symptoms based on VO_2max and relevant biomarkers.

Chapter 3

Implementation of Machine Learning Models for Early Detection of CVD and T2D

This chapter presents the methodology which introduces the project implementation process, along with specific functional and non-functional requirements and the implementation process of the proposed system. The first step involves the acquisition of real data from a smartwatch. Then we analyse the data and apply the necessary transformations so that the data is adequate for each CVD and T2D profiles, alongside the implementation of a risk stratification strategy. After data preparation, we apply a feature selection technique to select the most relevant features to build predictive classification models. Lastly, we apply two supervised machine learning algorithms to benchmark the performance of two datasets for both types of diseases and determine which model(s) are most appropriate to evaluate their testing predictions.

3.1 Implementation Process

The conceptual architecture of the practical implementation of this project, as illustrated in Figure 3.1, outlines the workflow and interactions involved in data

acquisition, data transformation, machine learning model implementation, and evaluation.

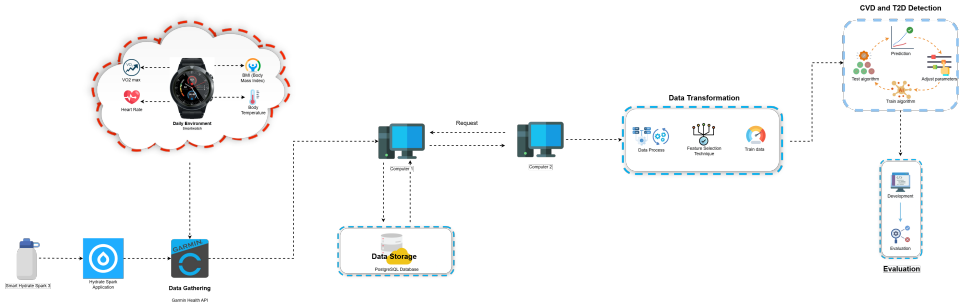


Figure 3.1: Conceptual Architecture

The implementation of this architecture begins with the data collection process. First, the data is acquired using the Venu 2 Plus Garmin smartwatch provided by the institute. The smartwatch is used on a daily basis to record health values in real time and measures several metrics such as VO_2max, HR, sleep time, SPO_2, among other features. To access the data from the smartwatch, we use the Garmin Connect v5.11 application for Android, which stores and monitors the collected data from the smartwatch via Bluetooth and *Wireless Field* (WiFi) network connection. To connect to the application, we create an account on the Garmin Connect Portal website by providing an email and a password. In addition to the data obtained from the Garmin application, the study integrates hydration intake using the HidrateSpark 3 smart water bottle, a legacy device from 2019 with a capacity of 20 *Ounces* (*Oz*) / 592 *Millilitres* (mL). The bottle remains compatible with the Hidrate Spark v4.1.3 Android application, despite its release date, and calculates the water consumption based on the difference in weight between a full bottle and an empty bottle. This option offers a more accurate measurement compared to the estimated manual input. As the Garmin Connect platform does not support automated integration with Hidrate Spark, the recorded hydration values are manually saved into the Garmin Connect to ensure consistency and completeness of the dataset. We use two computers, one for hosting a local PostgreSQL version 17 relational database to store all the information collected from the smartwatch using the Garmin Connect open source project [Klinkien, 2020] and a second computer with a *Virtual Machine* (VM) to pre-process the data and apply supervised ML algorithms to determine the presence of early CVD and T2D. The computers characteristics are the following:

Computer 1

- **CPU:** AMD Ryzen 3 2300U with Radeon Vega Mobile Gfx 2.00 GHz
- **RAM:** 8.00GB
- **Disk:** 120GB SSD SanDisk

Computer 2 with Oracle Virtual Box

- **CPU:** Intel(R) Core(TM) i7 6700 3.40 GHz
- **RAM:** 16.00GB
- **Disk:** 1 TB HDD WDC

  Virtual Machine:

  - **Operative System:** Linux Mint 22.1 xfce 64bit
  - **Cores Allocated:** 5 Cores
  - **Acceleration:** Nested Paging, Hyper-V Paravirtualization
  - **RAM:** 8.00GB
  - **Disk Space Allocated:** 50GB

The connection between the computer that hosts the database and the computer with the VM follows three steps.

On the first step, we access the database `pg\_hba.conf` file, which controls the client authentication, inside the PostgreSQL folder, we add a new non-local connection as shown in Figure 3.2.

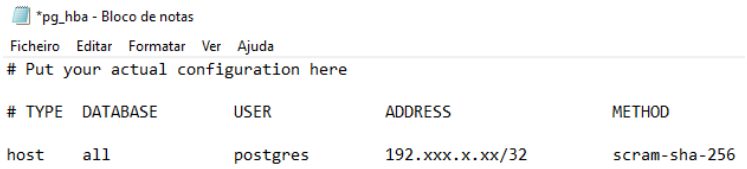


Figure 3.2: PostgreSQL Authentication Configuration File

In this connection string, we set a new host following the directives of the configuration file. For this, we need to set five values:

- TYPE: Indicates the type of connection between two machines. We use "host" parameter since it is a non-local connection;
- DATABASE: Indicates the database name to which the connection accesses. In this case, we set the parameter to "all", since there is only one database created, even though, for better security, we could set it to the actual name of the database;
- USER: Indicates the user role to which we have access. We only have one user, so we set the "postgres" parameter, since it is the only created user role;

- ADDRESS: Indicates the *Internet Protocol* (IP) address of the machine to which we are allowing the database to connect. Alongside the IP address, we indicate a *Classless Inter-Domain Routing* (CIDR) mask that represents an integer between 0 and 32 for IPv4 or 128 for IPv6. We use the 32 integer value since it is the most commonly used for small host routes and specifies a direct route to a single host.

- METHOD: Indicates the encryption methods used to send encrypted passwords. The most viable method presented by the file is scram-sha-256. This method is the most secure option to prevent password sniffing.

The second step is to configure the laptop's firewall, which hosts the database, to open the default port 5432 and allow access to the PostgreSQL database from the VM. Since it is on a local network, only devices within the same network can communicate with the PostgreSQL instance. To achieve this, we create a new input rule within the firewall by going to the advanced settings and setting the port with all domains activated, including the port number.

Lastly, we set the environment variables for the connection to the database, including, the host address, database name, database user and password, and the port to be used in the connection string.

3.2 Functional and Non-Functional Requirements

3.2.1 Functional Requirements

To ensure the system fulfils the intended purpose, it must have a group of clear defined capabilities. The functional requirements for the system are:

- The system must collect real-time data from smartwatches, using the Garmin Connect API.

- The system must preprocess the acquired data by handling missing values, normalise features, and structure data for analysis.

- The system must implement filtering techniques to identify the most relevant biomarkers to predict CVD and T2D.

- The system must be capable of training machine learning algorithms, using labelled health data for supervised classification problems. It must also support the use of *Cross Validation* (CV) techniques and hyperparameter tunning.

- The system must output a risk score or a classification report using performance metrics to assess the validity and reliability of the acquired model(s) over the early prediction of CVD and T2D.

Following these requirements, the system will be able to effectively acquire, process, and analyse health data to deliver accurate and interpretable health risk assessments.

3.2.2 Non-Functional Requirements

In addition to the core functionalities, the system must adhere to a set of non-functional requirements that ensure performance, reliability, and interpretability. The following non-functional requirements are:

- The system must process smartwatch data in a fast manner. Feature selection and machine learning techniques must be optimised to handle real-time or close to real-time analysis.
- The architecture must support a horizontal scale to support small to medium sized data volumes.
- The system must ensure a successful completion of a single time data request, with appropriate error handling and output logging in case of API failure, timeout, or authentication issues.
- The system must be modular, to allow easy updates to data pipelines, feature techniques, or machine learning algorithms.
- The system must provide clear visual feedback on data analysis and prediction outcomes.
- The system must achieve a minimum of 80% performance for the predictions of CVD and T2D.

These non-functional requirements support the system's overall functionality, in order to provide a reliable, secure, and efficient environment for the early prediction of CVD and T2D.

3.3 Data Acquisition

The implementation process begins with the configuration of the smartwatch, in which we define specific user parameters, such as age, height, weight and daily water intake, as well as the placement of the device on the wrist. For this study, the watch is placed on the left wrist, the non-dominant arm, to minimise interference with routine activities.

Garmin offers multiple APIs, including Activity, Courses, Health, Training and Women's Health APIs. This project uses the Garmin Health API, which enables access to biometric traits such as step count, VO_2max, HR beats per minute, sleep

time, RR per minute, SPO_2, BMI, among other health indicators. Access to the API requires Consumer and Secret Key credentials, which the institute provides since it partners with the Garmin Developer Program. Initially, we used the Garmin Health API to retrieve data in *JavaScript Object Notation* (JSON) format. However, during data extraction, several endpoints yielded limited information, and some failed to function correctly when queried using the Epoch/Unix time format, which Garmin Connect uses to search for data.

As an alternative to the Garmin Health API, we found an open source Python project called Garmin Connect API [Klinkien, 2020]. This tool uses web scraping to extract data from web interfaces and retrieve information directly from the Garmin Connect web portal. Unlike the limitations of the Garmin Health API, this approach enabled access to a broader range of detailed biometrics and activities recorded by the smartwatch. These include comprehensive activity data, defined user goals, challenges and smartwatch configuration settings. Figure A.1 demonstrates the available options that scrape the data from the Garmin Connect portal, providing a rich and structured data for later in depth analysis. This project uses a Python module, Garth, developed explicitly for interacting with the Garmin Connect portal. Garth uses web portal login credentials, email and password, and the API credential keys mentioned above to authenticate and retrieve data aligned with the project's analytical requirements. Each function of the program enables the extraction of unique data for a specific day. For this study, we collect data from six menu options, selected according to the relevance of their associated metrics in the related scientific literature in Section 2, which contains eighteen important metrics as follows in Table 3.1.

Menu Option	**Variables**
Stats and Body Composition	Steps, Minimum Heart Rate, Maximum Heart Rate, Resting Heart Rate, Average Respiration Rate, Average Waking Respiration Rate, Lowest Respiration Rate, Weight
Hydration	Daily Water Intake
Sleep	Sleep Duration, Sleep Average Respiration, Sleep Resting Heart Rate
SPO_2	Average Sleep SPO_2, Lowest Sleep SPO_2, Average Daytime SPO_2
Max Metric	VO_2max
Fitness Age	Body Mass Index, Fitness Age

Table 3.1: Garmin Connect API Menu Options and Tracked Variables

The Garmin Venu 2 Plus smartwatch collects and calculates specific health metrics on a daily basis. To that end, we contextualise and briefly explain how each metric works and is monitored. The device uses optical sensors and user input data to obtain physiological indicators throughout the day and night.

The VO_2max score is estimated during outdoor running activities, above the ten minute mark. According to Garmin, the calculation requires sustained physical effort that elevates the user's heart rate to at least 70% of the estimated maximum registered, otherwise it will not be able to calculate the score. The device uses heart rate, speed and distance data, with the help of an accelerometer, to estimate cardiorespiratory fitness with its proprietary algorithms. This study focused on 10 to 15 minute daily runs to calculate the score.

Regarding the SPO_2, it uses a pulse oximeter sensor integrated into the watch's optical heart rate sensor to monitor its frequency. The smartwatch measures the oxygen level in the blood by shining light on the skin and measures how much light is absorbed. The device records SPO_2 continuously during sleep and periodically throughout the day, presenting the readings in percentage. Similarly, measuring the RR using heart rate and body movement variations reports the results during daytime activities and nighttime sleep.

Heart rate data, including minimum, maximum and resting values, uses Garmin's optical sensor to monitor the frequency on a daily basis.

Sleep tracking uses a combination of heart rate monitoring and body movement data to categorise different stages of sleep (light, deep, REM, awake) and determine total sleep time. The smartwatch also tracks sleep RR and SPO_2, adding a more comprehensive view of sleep.

In this study, we manually enter the weight into the Garmin Connect application. The application uses this value and the user's height to calculate the BMI. Then, Garmin's proprietary smartwatch algorithms estimate fitness age based on BMI, VO_2max and activity data.

The step count uses a built-in accelerometer that detects repetitive motion patterns, whether by walking or running, with the assistance of *Global Positioning System* (GPS) tracking.

Lastly, as mentioned in Section 3, the device does not automatically track water intake. However, the user manually inputs the values into the Garmin Connect application with the help of the Hidrate Spark application.

On the other hand, we modified the original Garmin Python program to save the extracted information in JSON files to facilitate structured data storage. We append the new data to the corresponding files by adjusting the selected dates. To upload the data to a local PostgreSQL database, we create a Python script that inserts each health metric's information stored in JSON format. The script, for each record, attempts to parse the complete JSON structure into a variable named `details`, which maintains the hierarchical organisation of the original file. Before the insertion of each record, the script checks whether a record, for a specified date, exists on the target table. The script skips that entry to prevent data duplication if a matching date is found. Since the data is in JSON format, the database exhibits

a semi-structured design. Although the data shows some degree of organisation, the data does not fully conform to a fixed relational schema. With this in mind, we create five tables corresponding to distinct categories of different health metrics groups to accommodate the following format:

- **User Metrics**: Stores VO_2max daily tracking information;
- **Daily Activity Data**: Stores information regarding respiration rate, heart rate, daily steps and weight;
- **Fitness Data**: Stores information regarding the estimated fitness age and body mass index;
- **Hydration Data**: Stores daily hydration intake values;
- **Pulse Oximeter Data**: Stores blood oxygen saturation levels;
- **Sleep Data**: Stores sleep time and metrics measured during sleep;

These tables store raw JSON data directly. PostgreSQL provides two options for handling JSON data. Plain text format retains the original structure and formatting, while the binary format known as JSONB, stores data in a decomposed binary representation. Since we only extract a specific subset of values from the JSON structure, we opt for the binary format. This configuration offers better performance for querying, given its compressed binary structure and is optimised to retrieve key value pairs from JSON objects efficiently.

3.4 Pre-Processing

The preprocessing step begins by fetching data from all five tables created previously and merging only the necessary subset of features into a new table. We perform this step within the database to reduce the network traffic when requesting the processed data from this new table, instead of requesting all tables to perform the data handling in the Python script. As illustrated in Figure 3.3, the query extracts and filters only the most essential health/fitness metrics for downstream use.

```sql
INSERT INTO person_record_ml (
    calendarDate, Vo2MaxPreciseValue,
    steps, minHeartRate, maxHeartRate, restingHeartRate, highestRespirationValue, lowestRespirationValue, avgWakingRespirationValue, weight,
    fitnessAge, bmi,
    hydrationValueInML,
    averageSpO2, lowestSpO2, avgSleepSpO2,
    sleepTimeSeconds, sleepAverageRespirationValue, sleepRestingHeartRate
)
SELECT
    (u.details->'generic'->>'calendarDate')::VARCHAR AS calendarDate,
    (u.details->'generic'->>'vo2MaxPreciseValue')::REAL AS Vo2MaxPreciseValue,
    (sb.details->>'totalSteps')::INTEGER AS steps,
    (sb.details->>'minHeartRate')::INTEGER AS minHeartRate,
    (sb.details->>'maxHeartRate')::INTEGER AS maxHeartRate,
    (sb.details->>'restingHeartRate')::INTEGER AS restingHeartRate,
    (sb.details->>'latestRespirationValue')::REAL AS highestRespirationValue,
    (sb.details->>'lowestRespirationValue')::REAL AS lowestRespirationValue,
    (sb.details->>'avgWakingRespirationValue')::REAL AS avgWakingRespirationValue,
    (sb.details->>'weight')::REAL AS weight,
    (fa.details->>'fitnessAge')::DOUBLE PRECISION AS fitnessAge,
    (fa.details->'components'->'bmi'->>'value')::REAL AS bmi,
    (h.details->>'valueInML')::REAL AS hydrationValueinML,
    (po.details->>'averageSpO2')::REAL AS averageSpO2,
    (po.details->>'lowestSpO2')::REAL AS lowestSpO2,
    (po.details->>'avgSleepSpO2')::REAL AS avgSleepSpO2,
    (s.details->'dailySleepDTO'->>'sleepTimeSeconds')::INTEGER AS sleepTimeSeconds,
    (s.details->'dailySleepDTO'->>'averageRespirationValue')::REAL AS sleepAverageRespirationValue,
    (s.details->>'restingHeartRate')::INTEGER AS sleepRestingHeartRate
FROM user_metrics u
LEFT JOIN daily_activity_data sb ON u.details->'generic'->>'calendarDate' = sb.details->>'calendarDate'
LEFT JOIN fitness_data fa ON u.details->'generic'->>'calendarDate' = split_part(fa.details->>'lastUpdated', 'T', 1)
LEFT JOIN hydration_data h ON u.details->'generic'->>'calendarDate' = h.details->>'calendarDate'
LEFT JOIN pulse_ox_data po ON u.details->'generic'->>'calendarDate' = po.details->>'calendarDate'
LEFT JOIN sleep_data s ON u.details->'generic'->>'calendarDate' = s.details->'dailySleepDTO'->>'calendarDate'
ON CONFLICT (calendarDate) DO NOTHING;
```

Figure 3.3: Overview of a SQL Query Data Filtration

The *Structured Query Language* (SQL) query starts by inserting daily health records into a table named 'person_record_ml'. The information comes from five different source tables, each storing information in a `details` column formatted in JSONB. The query retrieves the necessary fields, using formatted SQL queries for the key value JSON format. In this step, the query executes LEFT JOIN operations on the remaining tables to align the records by `calendarDate` to maintain chronological consistency. Afterwards, it retrieves and converts multiple data points into their corresponding data types, conforming to the target schema. The query uses the `user_metrics` table, which serves as the anchor table due to its inclusion of the key metric VO_2max that helps shape the structure of the `person_record_ml` table. To avoid inserting duplicate entries, we added a safeguard, using the ON CONFLICT (calendarDate) DO NOTHING clause. This clause ensures that each record is unique, given the date stored.

After the SQL operation, we created another Python script to retrieve data from the database using a SQLAlchemy connection string. SQLAlchemy is an efficient and high performance Python database toolkit module that provides SQL features for database access. Regarding the security purposes of the PostgreSQL database, we use connection credentials, including username, database name, password, host and port, retrieved from environment variables. After establishing a connection between a virtual machine, and a local database hosted on another computer in the same network, we perform further preprocessing operations in preparation for the analytical stage. Once we retrieve the data, we standardise the column names to adhere to specific naming conventions, to show consistency and readability throughout the dataset, as seen in Figure 3.4.

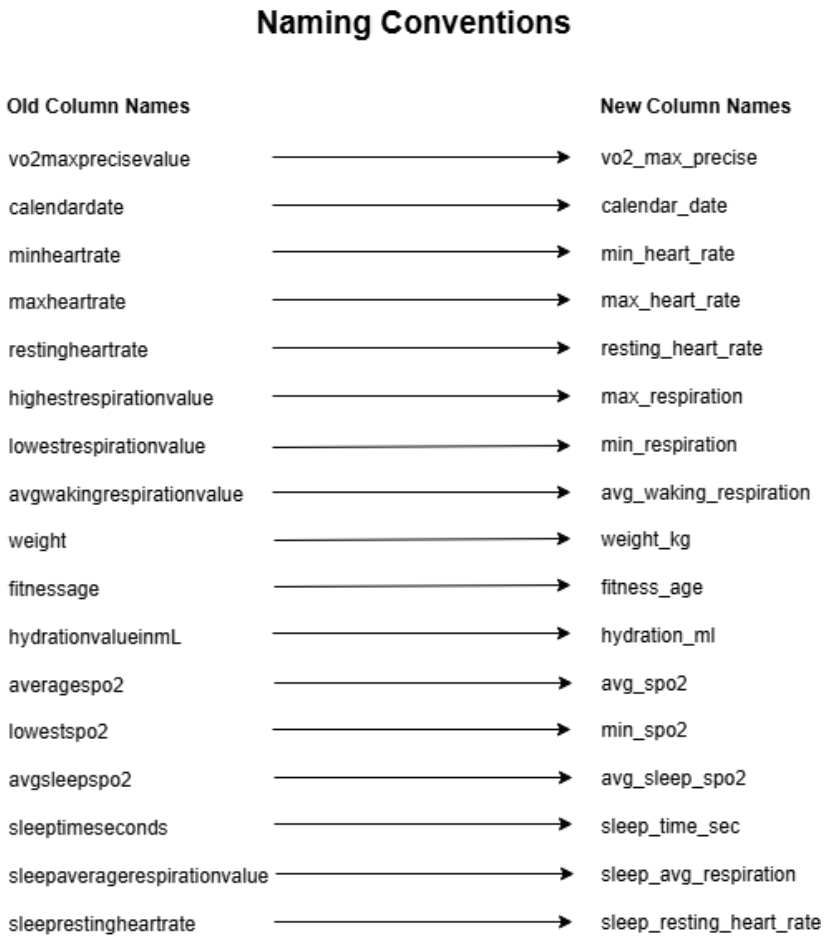


Figure 3.4: Naming Conventions of Column Names Extracted with the Garmin Connect API

Specific values such as weight and fitness age are transformed and rounded to one decimal place. The weight values exhibited scaling anomalies (e.g. 67548 *Kilograms* (KG) instead of 67.6kg), so we divide each value by a factor of one thousand. Finally, we also examine the dataset for missing or invalid values, expressed in the form of *NULL* or *NaN* values.

Given the absence of a predefined target column for assessing different stages of CVD risk and enable supervised learning, we analyse each extracted feature and search for possible combinations of health metrics associated with an increased impact on CVD risk. To this end, the following approach presents the correlations of physiological indicators, supported by the existing literature, as well as independent variables, as relevant markers for CVD:

VO_2max and BMI: The VO_2max and BMI complement each other with their roles in assessing cardiovascular health risks. The VO_2max is the body's ability to use and transport oxygen during intense physical activity, while BMI is a widely used marker that classifies individuals based on the weight relative to the height. Elevated BMI levels have long been linked to elevated risk of cardiovascular diseases,

as confirmed in a cohort study that found a strong association between obesity categories and cardiovascular events [Ghazizadeh et al., 2020]. A recent systematic review and study analysis demonstrate that individuals with high cardiovascular fitness have significantly lower mortality risks. However, compared to individuals classified as unfit, they show two to three times the increased risk of CVD [Weeldreyer et al., 2025]. Therefore, the combination of these two metrics provides a more holistic view for the risk health assessment. To improve the robustness of the risk prediction, we designate individuals below <41.7 VO_2max Garmin score assessment, for people with poor cardiorespiratory fitness. At the same time, we defined a BMI healthy range between 18.5 kg/m^2 to 24.9 kg/m^2 [Banack et al., 2025] based on threshold guidelines. Measurements falling outside this range are considered to be at increased risk of CVD.

VO_2max and Sleep Duration: Following the earlier recognition of VO_2max as a key indicator of cardiorespiratory fitness, we combine this feature with sleep duration to provide another view on cardiovascular health outcomes. While VO_2max reflects the body's aerobic efficiency, sleep duration captures the body's ability to recover, repair and maintain essential biological processes. Research by [Wang et al., 2020] that analyses the sleep durations of several individuals shows that sleep under six or more than eight hours is correlated with a higher risk of developing CVD. To counter the effects of sleep apnea or oversleeping, physical activity can help attenuate or even reverse those effects [Liang et al., 2023]. Together, these two variables show that high cardiovascular fitness can ease the harmful effects of inadequate sleep, giving performance and recovery factors to assess early CVD.

BMI and Resting Heart Rate: The integration of BMI and RHR in cardiovascular risk assessment is supported by evidence highlighting their individual and combined values in cardiovascular outcomes. With BMI as a structural element and indicator of cardiometabolic risk assessment, RHR is a non-invasive, accessible and reliable physiological marker in cardiovascular risk assessment. Elevated RHR above the 60 beats per minute indicator has shown to be a strong predictive indicator for cardiovascular events [Choi et al., 2024], particularly when elevated alongside increased BMI. Recent evidence from studies demonstrated that both BMI and exercise significantly influence RHR over time, suggesting a relationship between physical condition and cardiac regulation [Ehrenwald et al., 2019]. Similarly, another study found that higher RHR and elevated BMI predicted greater coronary atherosclerosis severity, as measured by the SYNTAX score, in patients with stable angina pectoris [Zhang et al., 2018]. These findings show that the pathological interaction between both parameters justifies their combination for a physiologically meaningful evaluation of CVD risk.

Resting Heart Rate and Average Waking Respiration: Having RHR, previously discussed as a key marker of cardiovascular risk, we connected it with the

average respiratory rate, which reflects the efficiency of pulmonary exchange between oxygen and carbon dioxide. Healthy adults RR falls between twelve to twenty times per minute [Pleil et al., 2021, Kumar et al., 2022], where values outside this range may reveal indications of CVD presence. Researchers recognise the importance of these biometrics in the study of heart and lung function. However, there has been no thorough exploration of the advances in these features. A recent study [Zhang et al., 2024] observed that disruptions in these parameters rarely happen in isolation. The researchers emphasise that combining both features may offer a different view in the assessment of acute illness, rather than measuring them separately. In this context, integrating RHR and average waking respiration rate, we can capture subtle physiological dysregulations that could be overlooked when each parameter is analysed separately.

Fitness Age and Resting Heart Rate: Following the discussion on RHR, we explore the combined effect of fitness age as an integrated marker of cardiovascular health. Fitness age consists of an individual's physiological condition compared to chronological age. When the fitness age exceeds the individual's, it indicates a decline in physical fitness and an increased cardiovascular vulnerability. Since both metric brings critical health insights, a recent study [Pittaras et al., 2024] reported that fitness age is consistently not accounted for when using predictive models to assess mortality risk. In parallel, another study demonstrated a strong inverse relationship between RHR and fitness age at the population level, reinforcing RHR role as a non-invasive indicator of physiological ageing [Gonzales et al., 2023]. From this perspective, combining both metrics brings an additional view on cardiovascular health with tracking early warning signs of CVD instead of analysing them individually.

Average Nighttime Blood Oxygen Saturation and Sleep Duration: We consider the combined role of nighttime SPO_2 and sleep duration to advance the cardiovascular risk assessment beyond traditional sleep metrics. While each parameter relates to cardiovascular health, its combination provides a more physiologically holistic view of sleep quality and blood oxygen efficiency. Impaired nighttime oxygenation, reflected in a low average SPO_2 index outside the range of 95-100%, has been strongly associated with poorer cardiovascular health metrics, independent of total sleep time. A study provides evidence, demonstrating that reduced oxygen saturation during sleep significantly correlates with unfavourable behavioural and biological cardiovascular health profiles [Häusler et al., 2019]. When evaluated together with sleep duration, SPO_2 measurements help differentiate simple variations in sleep length and sleep patterns involving underlying sleep dysfunction.

Daily Step Count: Daily step count is a simple metric of physical activity that plays a significant role in CVD prevention. Evidence from an extensive pooled analysis shows a clear inverse relationship where people who take more steps each

day tend to experience rarer CVD events [Paluch et al., 2023]. Researchers analysed the number of steps necessary to reduce CVD incidence and highlighted that an average of 4.500 daily steps can lower the risk of CVD, especially among sedentary populations [Ahmadi et al., 2024]. In summary, we incorporated 4.500 steps as a validated value in the assessment of CVD.

Daily Water Intake: Water is essential to the human body. Healthy individuals, on average, drink between 2.0 to 2.5 *Litres* (L) of water, among which vary from fluids from beverages to moisture present in food [Iversen and Fogelholm, 2023]. A research study indicates that roughly 20% to 30% of this total water intake comes from food, leaving approximately 1.5L as a reasonable target value for direct fluid consumption analysis. This value accounts for the remaining 70% to 80% of total hydration needs, which supports the physiological balance of the human body water percentage. To explore the impact of high serum sodium levels, a group of researchers [Dmitrieva et al., 2022] analysed a population of middle aged individuals, where most individuals were likely to experience faster ageing and be vulnerable to chronic diseases. The study highlights that even without a proper connection to CVD, the authors highlight the need to maintain a proper hydration to regulate the body's fluid balance while safeguarding long term outcomes. In this project, we will use the target value of 1.5L as validated by recent scientific literature in early assessment.

Mean Blood Oxygen Saturation: Blood Oxygen Saturation quantifies the percentage of oxygen in the bloodstream and is considered an important measure of the body's function. Smartwatch based oximeters have been shown to provide a non-invasive way of precise measurements supporting SPO_2 application in technology and potential role in health monitoring. A retrospective study reported that maintaining an SPO_2 target below 90% significantly increased the incidence of hypoxemia [O'Driscoll et al., 2024], whereas healthy adults typically sustain SPO_2 levels between 95% and 100%. Building on this, recent research showed that advanced derived oximetry features, such as the frequency and depth of desaturations, serve as strong independent predictors of CVD [Cade, 2022]. Given that reduced oxygen delivery can aggravate myocardial stress and increase the risk of arrhythmias, ischemia and heart failure, consistently low mean SPO_2 levels can reflect underlying respiratory or circulatory inefficiencies. In this study, we will adhere to the guidelines of 95% to 100% for SPO_2 percentages in assessing early CVD.

Mean Respiration Rate: The RR consists on the measurement of the number of breaths per minute. This metric allows the measurement of the respiratory frequency without subjecting the user to invasive procedures. Respiration rate in adults predominantly ranges from twelve to twenty breaths per minute [Rückert-Eheberg et al., 2025], where abnormal readings often show signs of rapid or shallow

breathing that may lead to conditions like tachypnea or bradypnea. The role of respiration rate in CVD is supported by studies that analysed several individuals after having an acute MYI, where RR variations were linked to poor outcomes [Barthel et al., 2013]. These readings show that RR brings another weight to the assessment of CVD, rather than being a mere measure of breathing. As a result, these findings support the introduction of the respiratory rate in the early CVD risk assessment from which health professionals may detect changes before severe issues develop.

Sleep Resting Heart Rate: During sleep, the RHR tends to be slower, because of the individual's relaxed state during nighttime. Various studies identify a healthy RHR during sleep, which falls between forty and sixty beats per minute, provides a baseline reference for cardiovascular fitness and welfare [Speed et al., 2023, Gonzales et al., 2023]. A cohort study examined patients with *Implantable Cardioverter Defibrillator* (ICD) and found that higher nighttime RHR leads to a higher incidence of cardiovascular mortality, which reinforces its significance in early CVD risk detection [Jiang et al., 2023]. Therefore, monitoring sleep resting heart rate offers another approach to identify individuals with a high risk of early cardiovascular disease, as an independent marker.

Sleep Average Respiration Rate: Respiratory rate, during sleep, enters a more relaxed state, with slower and more regular breathing patterns than daytime RR. Evidence from a cohort study identified a mean nocturnal respiratory rate between fourteen and sixteen breaths per minute as being associated with the lowest risk of cardiovascular risk, and mortality. This suggests that this range is a reference point for evaluating nocturnal respiratory health [Baumert et al., 2019]. Respiratory rates may fall below this threshold in individuals with superior cardiorespiratory fitness, such as athletes, without manifesting related signs. Elevated respiratory rates during sleep may indicate instability or cardiopulmonary stress, which can lead to an increased cardiovascular risk. To support this, researchers examined nighttime heart rate and found that nocturnal physiological markers, such as elevated respiration rate, are independently associated with increased cardiovascular mortality, even in patients with controlled cardiac conditions, including those with ICD [Jiang et al., 2023]. These findings demonstrate the importance of including sleep respiratory rate in cardiovascular assessments.

Afterwards, we assign normalised weight scores to each combination or individual metrics, giving higher weights based on observed correlations and significance, as documented in prior studies. The resulting composite risk score is computed as a weighted sum of the selected metrics, reflecting the classification of low, medium or high CVD multiclass risk levels, as shown in Table 3.2.

This risk stratification criteria will serve to determine what is the risk level of the user developing or having already any symptoms of chronic diseases and how we should proceed based on those risk assessments.

Metric Combination	Weights	Description
VO_2max_bmi	0.13	Relationship between Maximum Oxygen Uptake and Body Mass Index
bmi_resting_hr	0.12	Link between Body Mass Index and Resting Heart Rate
fitness_age_resting_hr	0.11	Combination of physiological age with Resting Heart Rate
avg_waking_respiration_resting_hr	0.10	Association between Average Respiratory efficiency, during wakefulness, with Resting Heart Rate
VO_2max_sleep_time	0.09	Captures the impact of sleep duration and Maximum Oxygen Uptake.
sleep_resting_hr	0.08	Associates Nighttime Resting Heart Rate with cardiovascular diseases
sleep_avg_SPO_2_sleep_time	0.08	Connects Sleep Oxygen Saturation and total sleep time
sleep_avg_respiration	0.07	Represents the Respiratory Rate consistency during sleep as an early CVD risk indicator
avg_SPO_2	0.07	Measures Average Oxygenation as a marker of cardiopulmonary health
steps	0.06	Reflects physical activity, linked to reduced cardiovascular risk
avg_waking_respiration	0.05	Represents the number of breaths per minute monitored during daytime
hydration_ml	0.04	Assesses fluid intake, relevant to cardiovascular health
Total	**1.00**	**Sum of all normalised metric weights**

Table 3.2: Risk Criteria Stratification with Association of Weights to Independent and Correlated Features Combinations

The implementation of this risk criteria using these weights to compute a cumulative risk score is used to classify the CVD risk for all records. The risk stratification criteria is as follows:

- If the score is **less than or equal to 0.3**, then the cardiovascular risk is classified as **Low** risk.
- If the score is **greater than 0.3 and less than 0.5**, then the cardiovascular risk is classified as **Medium** risk.
- If the score is **equal or greater than 0.5**, then the cardiovascular risk is classified as **High** risk.

After iterating through all records and assigning the appropriate risk scores, we analysed the data and found that each record received a low risk score. This indicates that the user is in good health, as defined by the ranges followed in the scientific literature. However, since monitoring data over a certain period of time resulted in only low risk score data, applying supervised ML algorithms might face challenges, since it requires variations so that the model can interpret patterns and manage to predict the risks. To this end, with the help of AI, we created scenarios with monitored data for different CVD risk profiles. In this way, we have distinct profiles with low, medium and high risk output data, so that we can benchmark the profiles using the user data as a starting point for one dataset, while the other uses the actual user data to generate data. Alongside, we provide the AI with an explanation of metrics, including the user's lifestyle, habits and age, to create the datasets with following scenarios:

- The user is healthy and does exercise regularly;
- The user has a sedentary lifestyle;

- The user already has some cardiovascular symptoms.

Following these scenarios, we requested the creation of synthetic data for each profile:

- The AI generates data not related to the user data;
- The AI generates data based on the user data.

This yielded six different scenarios, where each dataset will have three risk scores. Initially, for each scenario, we asked to create some records to analyse the quality and see any dispersion of data related to features. As a result, some of the values generated were not consistent with the expected behaviour, so we applied the following rules to generate data realistically aligned to each profile:

- BMI calculation must use the height provided by the user;
- Fitness age minimum value is 18;
- There must be a correlation between BMI and fitness age, meaning the lower the value of fitness age, the lower the value of BMI.
- There must be a correlation between fitness age and VO_2max, meaning the higher the values of VO_2max the lower the fitness age values;
- There must be a correlation between RHR and VO_2max, meaning the higher the value of VO_2max, the lower the RHR;
- There must be a correlation between weight and BMI since the weight values are used in the BMI calculation;
- The minimum SPO_2 value must be correlated to the average sleep SPO_2 value.
- The average waking respiration must be correlated with minimum and maximum respiration rates;
- For the scenario with the presence of CVD or high risk CVD, generate sleep duration outside the range of six and eight hours;
- For the scenario with High risk of CVD, simulate during sleep an abnormal respiration rate, either indicating sleep apnea or hyperventilation;

Under these rules, the AI provided realistic and dispersed data while maintaining the correlation of variables. As a result, we synthesised 2.000 daily records for both datasets.

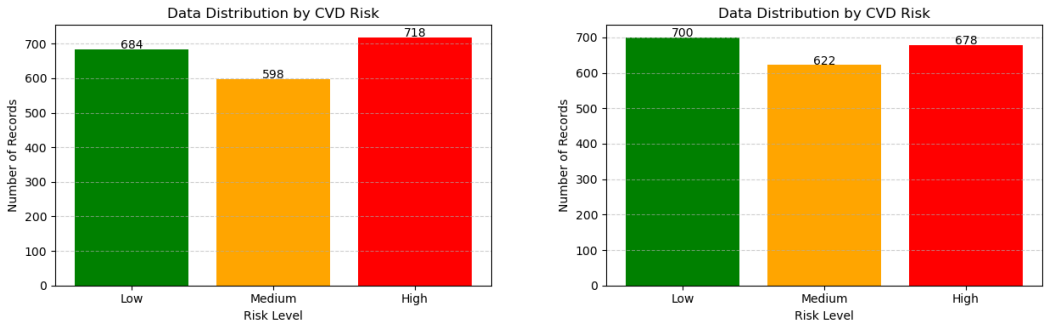
```
profile_types = ['healthy', 'sedentary', 'cvd']
profile_probs = [0.2, 0.4, 0.4] # simulates cvd risk scenarios

...

# --- Main Loop ---
for i in range(records):
    date = start_date + datetime.timedelta(days=i)
    profile = np.random.choice(profile_types, p=profile_probs)
        # choose profile randomly

    if profile=='healthy':
        ...
    elif profile == 'sedentary':
        ...
    else: #cvd
        ...

    ...
```

Listing 3.1: Simulation of Daily Health Profiles for both CVD Datasets

Before iterating each record, the following script randomly selects a risk class based on a reasonably balanced probability distribution, as seen in Listing 3.1. At the same time, it introduces a controlled overlap between risk profiles to reflect the natural variability observed in real world cases. This approach ensures a relatively balanced distribution across CVD risk categories for the practical application of machine learning algorithms.

To illustrate these design choices, Table A.1 presents a detailed overview of the expected synthetic datasets for both applications on CVD and T2D risk profiles, detailing their origin, feature composition, and intended purpose within this study.

To assess the representation and balance of synthetic datasets, we examined the distribution of risk classes across both CVD collections. Figure 3.5a displays the result of the CVD dataset that generated data not using the user data, where it displays 684 records for low risk, 598 for medium risk and 718 for high risk. Despite not being a perfect balanced dataset it still offers a reasonably proportional distribution across the three risk categories.



(a) Distribution of CVD risks excluding user data

(b) Distribution of CVD risks based on user data

Figure 3.5: Comparison of CVD Dataset Volumes

On the other hand, Figure 3.5b shows the 2000 records synthesised that used the actual user data, where 700 corresponds to low risk, 622 medium risk and 678 to high risk of CVD. In this case, it shows a better distribution among the three risks, indicating a more uniform representation across cardiovascular risk categories.

Regarding the T2D dataset, we selected fourteen variables mentioned in Table A.1, derived from the CVD datasets that were related to type 2 diseases. Among those variables, we added four additional metrics to complement this dataset:

- **HRV_day_ms** - HRV estimated during daytime, measured in milliseconds.
- **HRV_night_ms** - HRV estimated during nighttime, measured in milliseconds.
- **total_calories** - Total calories burned throughout the day, including exercise.
- **skin_temp_proxy** - Estimated skin surface temperature, measured in Celcius.

These values were calculated based on obtainable data from the CVD datasets profiles. To simulate realistic physiological responses and derive proxy T2D metrics, a series of transformations are necessary to apply to this group of features. First, to calculate the HRV, two proxy metrics were constructed, one for daytime and another for nighttime. The daytime proxy incorporated daytime RHR, VO_2max and step count. Nighttime HRV proxy included sleep RHR, sleep duration and VO_2max. Both proxies were normalised to *Milliseconds* (ms) ranges, anchored to a reference age between 25 and 30 years. Secondly, a randomised running duration was generated between 10 and 15 minutes for the user. From this, active minutes were estimated by assuming that 85% of the total duration of the runs occurred in the heart rate zones ≥ 3, which according to Garmin represents a moderate to threshold activity range. The caloric expenditure was calculated using two components. The

Basal Metabolic Rate (BMR) via the Mifflin-St. Jeor Equation [Mifflin et al., 1990] to reflect daily energy expenditure influenced by cardiorespiratory fitness.

$$\text{BMR} = (10 \times \text{Weight}) + (6.25 \times \text{Height}) - (5 \times \text{Age}) + 5 \tag{3.1}$$

The second component included exercise induced calories determined by the activity factor based on the user VO_2max value. These activity values determined by the formula were adapted to each range of VO_2max.

- **Poor:** VO_2max $\leq 41.7 \rightarrow$ Activity factor = 1.2
- **Fair:** $41.8 \leq$ VO_2max $\leq 45.4 \rightarrow$ Activity factor = 1.375
- **Good:** $45.4 \leq$ VO_2max $\leq 51.1 \rightarrow$ Activity factor = 1.55
- **Excellent:** $51.1 \leq$ VO_2max $\leq 55.4 \rightarrow$ Activity factor = 1.725
- **Superior:** VO_2max $> 55.4 \rightarrow$ Activity factor = 1.9

The total caloric output combined BMR, scaled by an activity factor linked to individual VO_2max values, and calories burned during exercise.

Finally, a proxy for the skin temperature metric was estimated by integrating RR (daytime and nighttime), SPO_2, water intake, and step count. This composite formula shows the thermoregulatory influenced by respiration, oxygenation and hydration adjusted for physical activity. These engineered features enrich the dataset with interpretable marks relevant to the detection of T2D. In addition, to ensure the same coherence between both types of diseases, we adapt the same risk stratification target column established in the CVD dataset, alongside the same number of records for each dataset profile. This approach will allow us to facilitate the model comparison for both diseases.

The following section, outlines the data profiles corresponding to each scenario within both types of datasets, providing a detailed account of the feature distributions for each multiclass risk assessment.

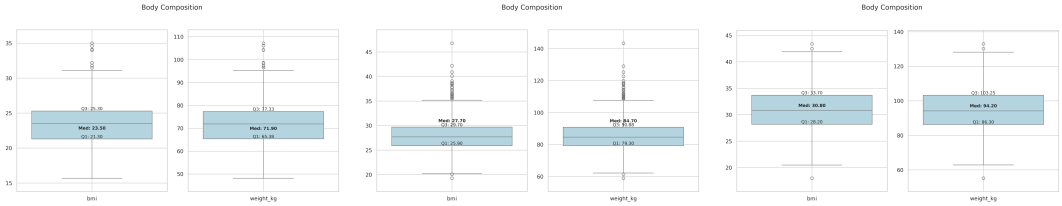
3.4.1 CVD Data profiles

Dataset Not Related on User Data

In this dataset, the data is fully simulated, without connections to the smartwatch user data, alongside a heatmap to show the correlation of metrics.

The following Figure 3.6a displays the observations of BMI and weight metrics. For the various categories of BMI, the present values range between 15.4kg/m^2 and 30.6kg/m^2. The majority of the values within the first and third quartiles are within an optimal range for a healthy profile, which varies between 18.5kg/m^2 and

24.9kg/m^2. However, several records show values outside of this range, as indicated by readings below the optimal range, influenced by the presence of weight values below 54kg and above 73kg. Furthermore, weight values range from 49.6kg to 95.7kg, indicates periods of weight fluctuation over time, where a higher weight does not necessarily indicate poor health, since BMI does not differentiate between muscle and fat.



(a) Healthy User (b) Sedentary User (c) User with CVD

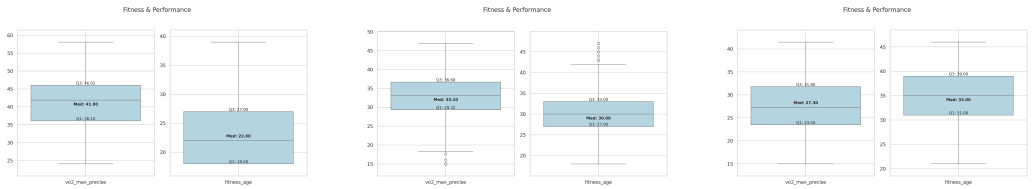
Figure 3.6: Body composition

Regarding Figure 3.6b displays values oriented to a sedentary lifestyle. The observed values highlight indications of body composition worsening that limit the management of BMI. The BMI values range between 20.2kg/m^2 and 35.1kg/m^2 show an individual within the healthy and obese categories, according to the standard BMI classification. Weight measurements ranging from 62.3kg to 108.6kg presents various outliers, with most going above 109kg, influence the calculation of BMI, and places the individual in the obese zone with values going over 35kg/m^2.

Lastly, Figure 3.6c highlights a user with CVD symptoms. The present values shows indications of metabolic concern with BMI values falling within the 20.2kg/m^2 and 42.7kg/m^2 range, and a median of 30.8kg/m^2. This interval range of values, places the user within the normal an obese categories of BMI. The weight measurements used in the BMI calculation range from 83.1kg to 86.5kg. Two outliers exceed this range, which are influenced by two weight values outside of the optimal range, leads to a further increase of the BMI assessment into the obese category. Moreover, a weight outlier, with a value of 55kg places the user at the lower end of the normal category, with a BMI of 18.8kg/m^2. These values capture the essence of the variability between the scenarios, and to that end, they should be recognised.

In terms of performance metrics, an analysis of Figure 3.7a shows the VO_2max scores, of a healthy peron with ranges from 24.7 to 58.8, and a median of 41.9. Garmin considers these results to be between Very Poor and Superior outcomes, which reflects periods of limited to high cardiovascular efficiency. The elevated VO_2max values are inversely proportional to fitness age. A high VO_2max corresponds to a lower fitness age, where an individual is used to regular exercise habits. In contrast, a low VO_2max, leads to a higher fitness age, which indicates that an individual is not suited for cardiovascular fitness exercises. Furthermore, fitness age

values span between 18 and 39, with a median of 22, suggests periods of fitness fluctuations, with most values falling to the optimal alignment of high aerobic capacity of 18 years.



(a) Healthy User (b) Sedentary User (c) User with CVD

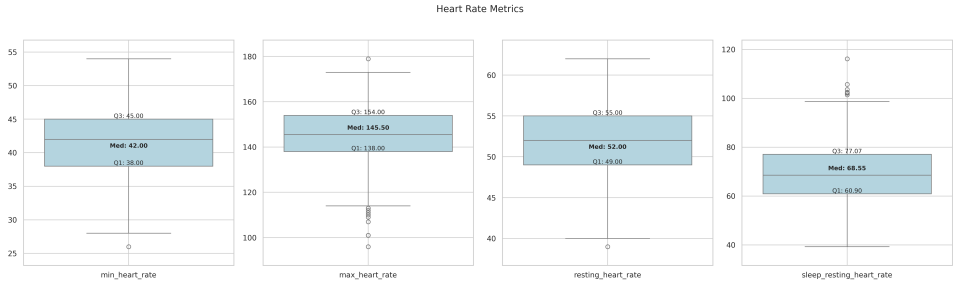
Figure 3.7: Fitness and Performance

Following a sedentary lifestyle, Figure 3.7b presents the VO_2max scores on a range from 18 to 46.7, and a median of 33, places the user within the Very Poor and Fair categories, according to the Garmin score. The presence of variability in aerobic performance, influences the fitness age to yield both good and poor results, with most values exceeding the user's chronological age. The elevated BMI contributes to a reduced aerobic efficiency, which leads to the observed VO_2max score and high fitness age.

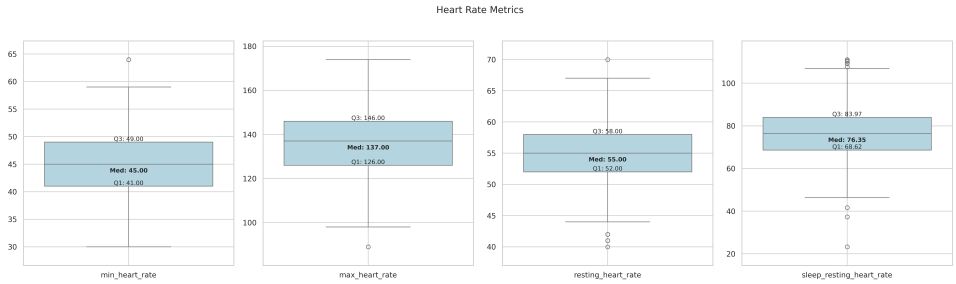
Finally, Figure 3.7c illustrates the performance metrics of a user with CVD. The VO_2max scores fall within a range of 15 to 41.9, classified by Garmin between Very Poor and Fair score. In terms of fitness age values, ranging from 21 to 45.3, with a median of 35, the majority of the values are above the user's chronological age. This trend is influenced by the high BMI values and the low score of VO_2max, reinforces its impact on cardiofitness performance and the presence of CVD.

The assessment of the heart rate metrics, in Figure 3.8a, shows numerous indications of how healthy is the individual cardiovascular system. Regarding RHR, it presents good values below the 60 *Beats Per Minute* (BPM), with most ranging between 35 and 59 BPM. The present values indicate that the user presents a good cardiovascular state for someone within the healthy profile. Minimum heart rate measurements concentrate between 19 and 51 BPM, while maximum heart rate fall between 123 and 198 BPM. This wide maximum margin above the 180 BPM indicates that the user tends to overexerts themselves. Garmin defines this level as a dangerous threshold, especially for those that are not accustomed to intense aerobic exercises. Conversely, values below 170 BPM during activities periods suggests minimal applied efforts. Finally, sleep RHR present variations in readings between 35 and 75 BPM, with several outliers reaching nearly 100 BPM. The usual sleep RHR tends to vary between 40 and 60 BPM, where lower readings are common among athletes. However, readings exceeding 70 BPM can indicate problems related to stress, sleep disorders, or illness, which could have occurred suddenly, at a given time.

(a) Healthy User



(b) Sedentary User



(c) User with CVD

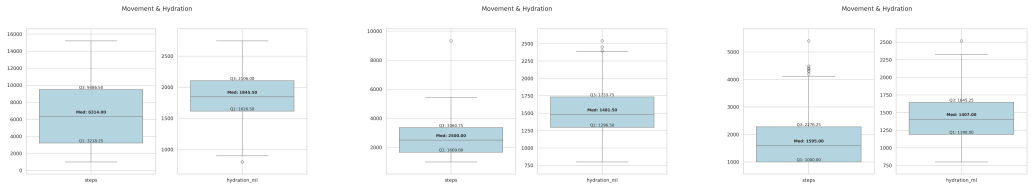
Figure 3.8: Heart Rate Metrics

The following Figure 3.8b displays daily heart rate metrics for a sedentary user. The minimum heart rate values range from 27 to 54 BPM, while the maximum rate ranges between 116 and 174 BPM. With maximum heart rate values above 170 BPM indicates periods of higher physical effort. Outliers below 116 bpm suggest periods of lower daily movement and limited cardiovascular stimulation, which are typical in individuals who follow a sedentary lifestyle. Resting heart rate and sleep resting heart rate values, range from 40 to 63 BPM and between 40 to 99 BPM, respectively. Given the nature of this scenario, values above the 60 BPM range suggest an indication of CVD risk.

Lastly, an individual with prolonged CVD symptoms, are shown in Figure 3.8c. Resting Heart Rate, during the day, ranges between 44.7 and 66.3 bpm. These values indicate a good cardiovascular beat rate, despite the user having CVD symptoms. Sleep RHR values range from 46 to 108 BPM, with most values above 60 BPM,

showing instances of tachycardia, with peaks above 100 BPM. Maximum and minimum heart rate values suggest that, despite having CVD, the user engages in light to moderate intensity exercise, with observed efforts aligning with Garmin endurance exercise zones between Fair and Threshold zones.

The water consumption and daily movement, in Figure 3.9a shows the patterns of a person within a healthy profile group. The observed step count ranges between 1.500 and 15.323 steps, with most values above the daily average for a normal individual. With this level of activity, the user requires a suitable water intake above 1.5L, although there were days when water consumption was inadequate.



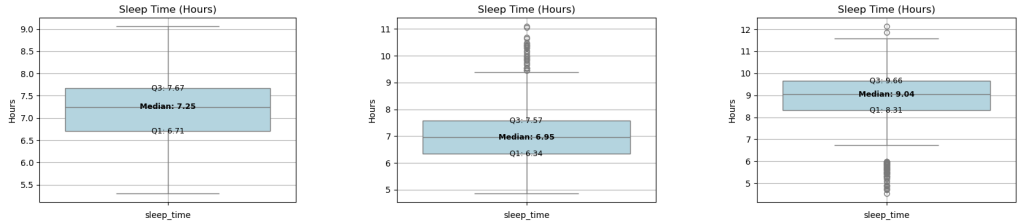
(a) Healthy User (b) Sedentary User (c) User with CVD

Figure 3.9: Movement and Water Intake

Within the sedentary lifestyle scenario Figure 3.9b displays the daily water intake and step count. The number of steps ranging from 1.000 to 5.643, and a median of 2.500 clearly indicate periods of activity. There were also periods with with less movement, below the commonly recommended threshold for active adults of 4.500 steps, clearly characterised by spending much time inactive on the same position. On the other hand, the user attempts to maintain a water consumption above the recommended 1.5L, as indicated by the median of 1.48L. However, the user still shows lapses in hydration that may impact his well being, in particular, individuals engaging in low to moderate levels of physical activity.

Finally, in Figure 3.9c, given the nature of the user's profile, the number of steps ranges between 1.000 and 4.100, with a median of 1.595 steps. This distribution suggests a generally low activity profile although, on occasional days, the user tries to reach the 4.500 step threshold, with one occasion where he surpasses the 5.000 steps. Furthermore, hydration levels indicate insufficient hydration intake, given the median user fluid intake of 1.4L, which indicates that the user manages to surpass the recommended water intake, with values reaching 2.3L.

The sleep duration observed in Figure 3.10a illustrate the sleeping schedule of a healthy person. The average sleeping time ranges from 5 to 9 hours, with periods of sleep below the recommended guidelines. The sleep regulatory mechanism forces the human body to recover lost time by making the body feel sleepier, often leading to prolonged sleep, as evidenced by sleep durations exceeding 8 hours. The majority of sleep duration meets the recommended guidelines for optimal rest and recovery consistent with the healthy CVD scenario.



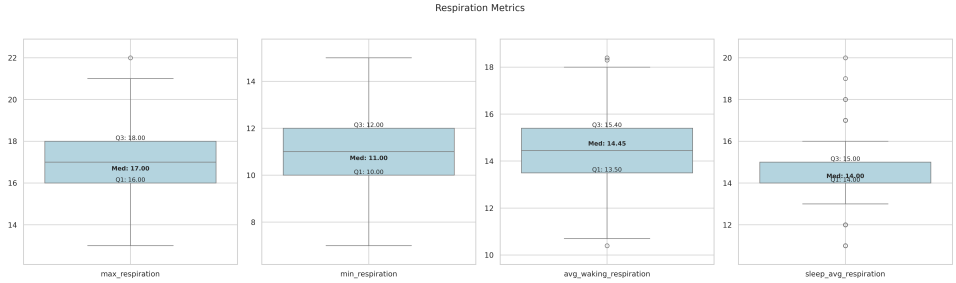
(a) Healthy User (b) Sedentary User (c) User with CVD

Figure 3.10: Sleep Durations

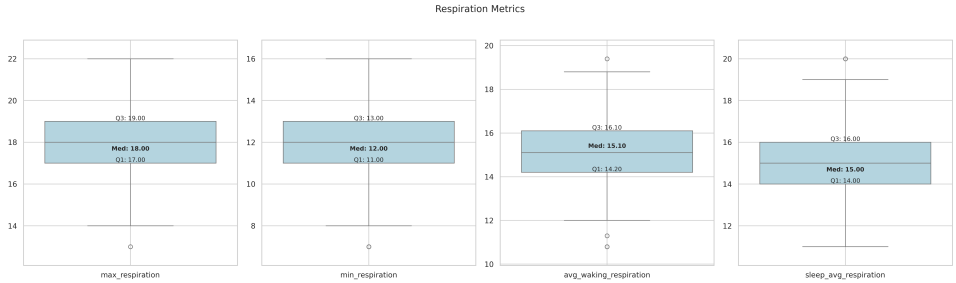
Figure 3.10b shows the sleeping schedule of a sedentary user with with most nights of sleep ranging between 5 and 9.4 hours. The observed sleep falls below the required range of 6 to 8 hours for adults. Instances of sleeping above 9 hours may indicate periods of sleep recovery, or prolonged sleep durations that are associated with an increase risk of CVD.

Finally, sleep time of an individual with prolonged CVD symptoms, as seen in Figure 3.10c shows that the user has difficulties in managing his sleeping schedule. The observations, shows that given the range of 6.8 to 11.5 hours with a median of 9 hours, the user sleep schedule stays outside the recommended range of 6 and 8 hours. Short sleep episodes below 6 hours indicate occurrences of sleep deprivation and may signal an increased risk of cardiovascular disease, metabolic dysfunction and cognitive decline. On the contrary, higher sleep times may compensate for previous nights of insufficient rest.

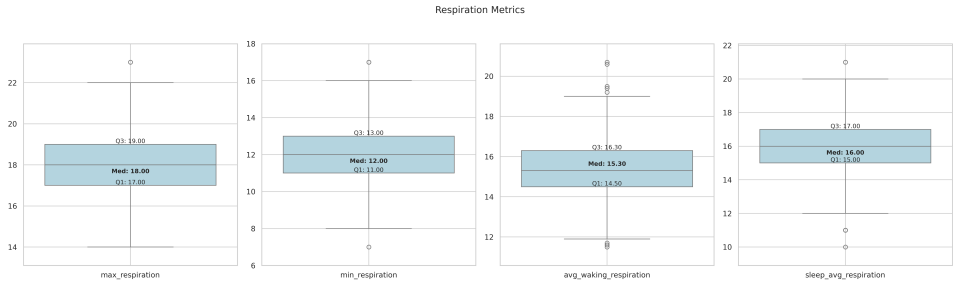
Examination of the results in Figure 3.11a captures the fluctuations of RR measurements over time. The observed sleep average *Respiration Per Minute* (RPM) ranges between 13 and 16 RPM. Several outliers above 16 RPM may be related to previous sleep difficulties, although still within the optimal range of 12 to 20 RPM. Periodic values lower than 12 RPM can indicate potential respiratory issues, such as bradyapnea. During the day, the respiration rate presents a rather stable pattern with values ranging from 11 to 18 RPM. Several readings below 12 RPM show episodes of a low respiratory rate, which indicates the starting symptoms of bradyapnea, but given the fact that the individual still does some exercise, this disruption can be attributed to exercises that require control of the breathing rate. Regarding the minimum RR, the values are slightly lower than usual, ranging from 6 to 15 RPM. These are likely reflecting deep rest or normal variations seen in athletes. The maximum breathing rate shows a range from 13 to 21 RPM, with one outlier of 22 RPM, where values tend to be high mostly during exercise.



(a) Healthy User



(b) Sedentary User



(c) User with CVD

Figure 3.11: Respiration Metrics

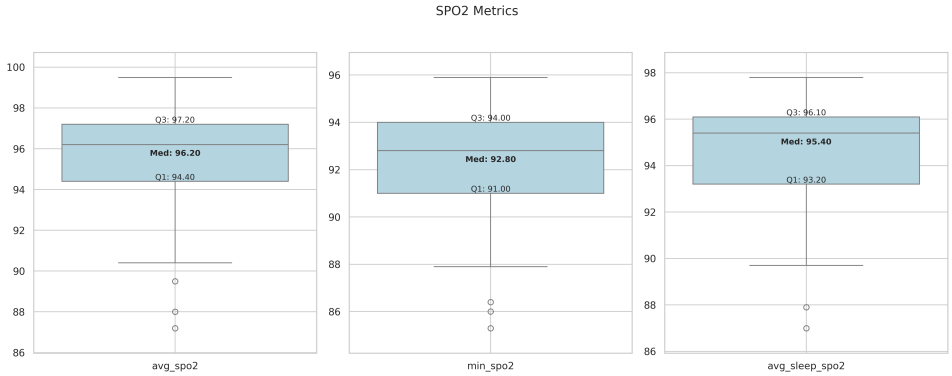
The following Figure 3.11b shows the profile of a person who engages in little activity, with respiration values monitored during day and night. During daytime, the values range between 12 and 19 RPM, which aligns with the optimal breathing range. As for maximum and minimal RR, the values exhibit a relatively stable interval throughout the day. At night, we observe several indications of restless nights with the presence of values below and above the optimal range of 14 to 16 RPM.

Lastly, Figure 3.11c presents various RR metrics for a user with CVD. The maximum respiration rates range from 14 to 22 RPM, with an outlier of 23 RPM. The minimum respiration rates fall between 8 and 16 RPM. Lower values suggest episodes of insufficient respiration, typical of bradyapnea. The average waking respiration rates range between 12 to 19 RPM, with some outliers outside the optimal range. Only two outliers are above the recommended respiration rate, whose values

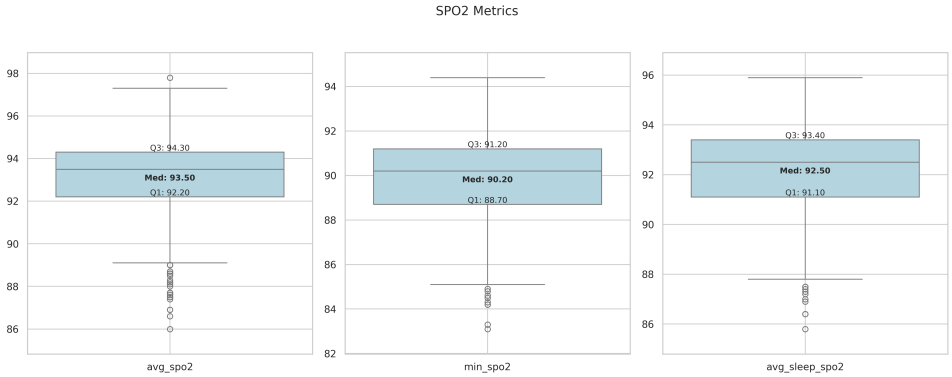
are 20.8 and 21 RPM. During sleep, the RR falls within expected ranges, with occasional values below 12 RPM that might reinforce bradypnea tendencies. An upper value above 20 RPM could indicate a brief episode of tachypnea, possibly linked to disturbed sleep or respiratory issues.

The observations of SPO_2, in Figure 3.12a, show oxygen levels between 90.4% and 99.8%, with most readings around 96.2%. These results show that within the interquartile range, there are no indications of hypoxic events. However, three outliers are present below the 90% range, indicating sporadic periods of hypoxemia events. Nighttime averages yield similar results to daytime with the presence of two episodes of hypoxemia. Although values higher than 90% are still within a reasonable range, the minimum SPO_2 values fall below the 90% may result from awkward or constricting sleeping positions that constrains the blood circulation.

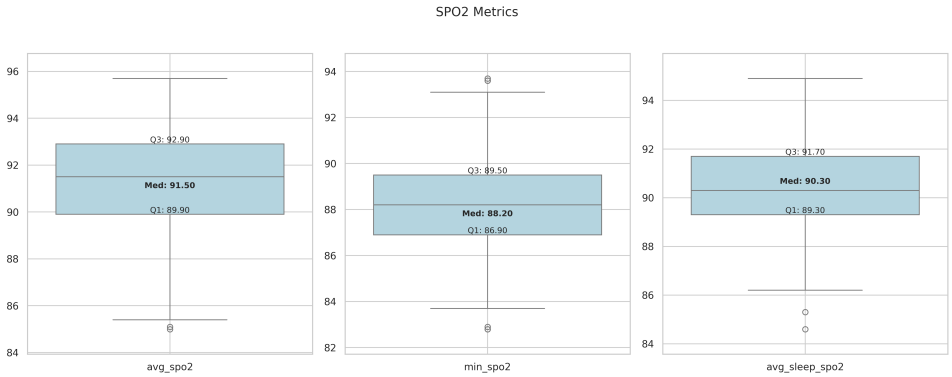
Figure 3.12b presents SPO_2 values during daytime and nighttime for a sedentary user. Daytime SPO_2 readings show values between 89% and 97.6%, indicating a generally adequate variation of SPO_2 levels with some values below the optimal threshold leading to episodes of hypoxemia. Nighttime SPO_2 values similarly suggest intermittent hypoxemia with values ranging between 87.9% and 95.9%, despite most indications inclining to signs of stable levels. To complement the night SPO_2 readings, the minimum SPO_2 further validates the measurements with most values below 90%.



(a) Healthy User



(b) Sedentary User



(c) User with CVD

Figure 3.12: Blood Oxygen Saturation Readings

Lastly, the scenario of CVD presents SPO_2 measurements throughout the day and night, as shown in Figure 3.12c. During daytime, average SPO2 ranges from 85.3% to 95.7%. The presence of outliers around 85.1% and 84.9%, alongside low values below the 90% mark suggests days with insufficient oxygen circulation. Average sleep SPO_2 ranges between 86.2% and 94.9%, with two notable outliers around 84.6% and 85.3%, also indicate episodic hypoxemia with minimum SPO_2 values reaching a low of 83.1%.

The following heatmap shown in Figure 3.13 illustrates the correlations among the eighteen features in the CVD dataset, which is based on the user data. According to the previously established rules, the heatmap reveals several notable relationships between physiological and biometric variables. Notably, a strong inverse correlation exists between resting heart rate and fitness age with VO_2max precise values. The inverse relationship indicates that individuals with higher VO_2max scores exhibit lower resting heart rates and fitness ages, which indicates a high level of CRF. Alternatively, higher VO_2max levels lead to a reduction of BMI and higher maximum heart rate.

A higher number of steps on a daily basis and a high water intake further confirms these correlations, since individuals with elevated VO_2max typically demonstrate a greater physical activity and which leads to a higher concentration of water intake levels. Similar trends are observed on respiration metrics, where a better CRF corresponds to a more efficient and stable respiratory system. Moreover, the analysis confirms that a higher step count is correlated with increased maximum heart rate, consistent with a higher physical effort.

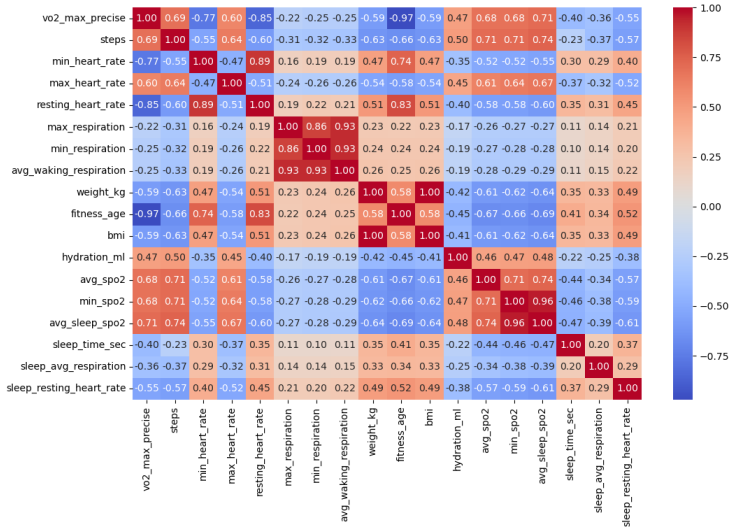


Figure 3.13: CVD Heatmap Correlation with Values Based on User Data

A correlation between BMI and weight is also evident, since weight is a direct measure in the BMI calculation formula. Fitness age and resting heart rate also show a positive correlation, where individuals with an active lifestyle tend to have lower fitness age, and resting heart rates. Furthermore, a positive correlation with BMI further shows that a healthier fitness age leads to a better cardiometabolic system.

The heatmap also reveals a positive relationship between minimum SPO_2, and sleep average SPO_2, as both metrics are intrinsically related. Lastly, a correlation

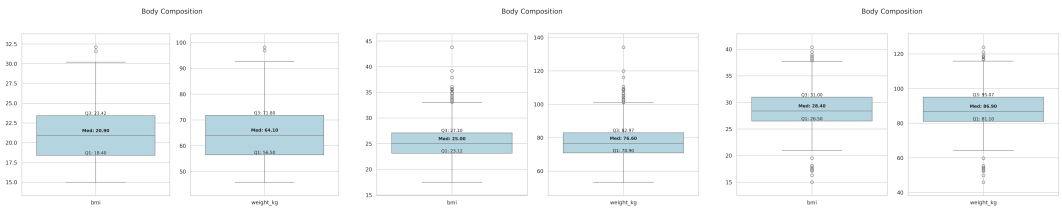
between maximum and minimum respiratory rates with average waking respiration also shows that respiration metrics with variability are intricately linked to breathing patterns measured throughout the day and night.

This heatmap shows that according to the rules implemented, the metrics capture coherent physiological relationships, as intended, according to real world cardiovascular and metabolic health indicators.

Dataset Based on User Data

In this dataset, we generate simulated data based on the user data as a starting point for the scenarios, alongside the correlation of metrics.

The following Figure 3.14a presents the distribution of body composition metrics on a healthy user scenario. The BMI values ranging from 15kg/m^2 to 30.1 kg/m^2 show a diverse range of categories within this profile. As shown in this figure, we observed that most interquartile range values fall within the Healthy classification, despite several data points lying outside the optimal range. To support the BMI present values, the weight exhibits similar dispersion, with values between 45kg and 92kg, alongside two outliers nearing 100kg. These high variations match the changes seen in the BMI values, since the weight variable is a direct measurement to calculate BMI alongside height.



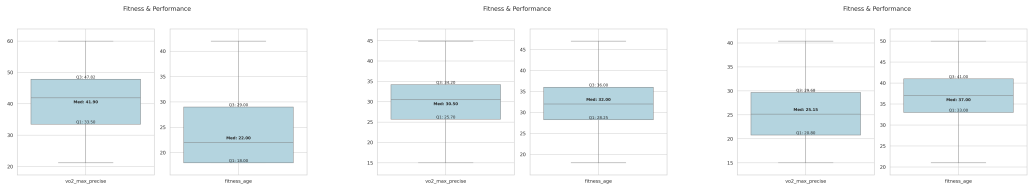
(a) Healthy User (b) Sedentary User (c) User with CVD

Figure 3.14: Body composition

In Figure 3.14b displays the BMI and weight values for the scenario, where the user exhibits signs of sedentary behaviour. The BMI values display signs of optimal range, from 17.4kg/m^2 to 33.3 kg/m^2, with some slight deviations towards higher readings. Several outliers surpass the optimal metabolic range, into the Overweight zone, including days exceeding 30kg/m^2, which marks an entry into the severe Obese zone. The weight values support these elevated BMI values, whose values range from 55kg to 101kg, with multiple outliers exceeding the 100kg range. This firmly positions the individual within the Overweight and Obese zones, which shows a transition from a healthy lifestyle profile to a starting stage of cardiometabolic implications.

Finnaly, in Figure 3.14c presents a scenario where the user has chronic diseases with an aggravated physique. The weights range from 60.8kg to 116kg, which correlate to the corresponding BMI values from 21.3kg/m^2 to 37.2kg/m^2. The presence of outliers shows indications of being between Underweight and Obese. These zones exhibit divergent health trajectories, with underweight values potentially resulting from low muscle mass, while values exceeding 30kg/m^2 suggest increased body fat, and a heightened risk for cardiometabolic complications.

Figure 3.15a illustrates the VO_2max and fitness age metrics. VO_2max reveals a mixture of scores, between Poor and Superior zones, whose values range from 21 to 60. In terms of fitness age, the values range between 18 and 46 years with a median of 22, indicating data variability and a comprehensive cardiovascular age scenario. Once again, this healthy profile, with overlapping data from other profiles, maintains a realistic data spread, ensuring that simulated conditions remain diverse while reflecting plausible human variability.



(a) Healthy User (b) Sedentary User (c) User with CVD

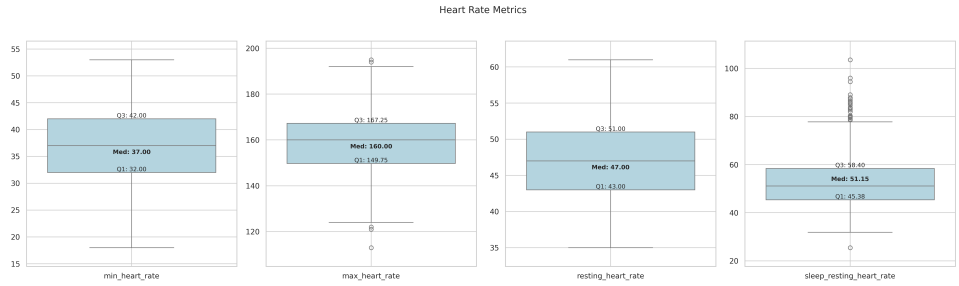
Figure 3.15: Fitness and Performance

Figure 3.15b presents the VO_2max and fitness age metrics for a sedentary user. In this scenario, we observe a worsening of the performance with values ranging from 15 to 45, which are classified in the range of Poor to Good zones, according to the Garmin score. The fitness age shows values between 18 and 47, with a median of 32, where it is visible that the user physiological age surpasses the chronological age within the 25-30 year range. This decline of CRF shows the start of a user sedentary lifestyle.

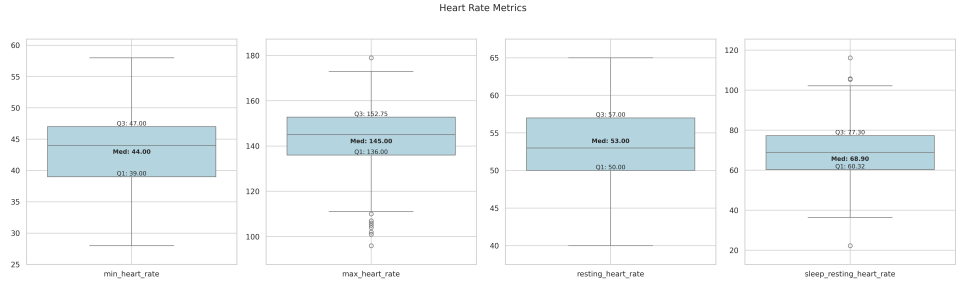
Finally, Figure 3.15c illustrates the VO_2max and fitness age metrics for a user scenario with CVD. The performance values, compared to previous scenarios, display even worse performance, with values ranging from 15 to 40. These values, according to Garmin, fall within the Poor and Fair zones, which shows a worsening effect of the CRF due to CVD. On the other hand, fitness age values ranged from 21 to 50, which indicates a weak cardiovascular system, supported by the aggravated values of VO_2max, despite some indications of good fitness age.

The following Figure 3.16a, presents heart metrics during rest, as well as during daytime. The maximum heart rate values range from 124 to 192bpm, which reflects elevated levels of physical activity, with days of fatigue surpassing 180bpm. In contrast, minimum heart rate values span between 18 and 53 bpm, indicate an excellent

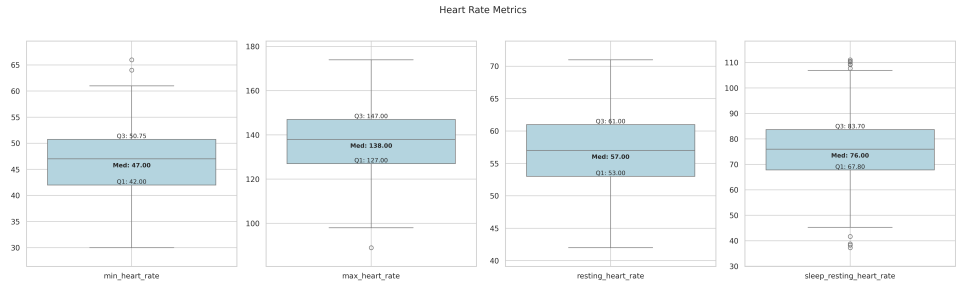
level of cardiovascular conditioning, which is seen in individuals with high aerobic fitness. Resting Heart Rate observations show a range from 35 to 62bpm. Most values are below the 60 bpm threshold, which suggests a strong cardiopulmonary efficiency, despite erratic values surpassing the 60bpm mark. The sleep resting heart rate frequency values range within the optimal range of 34 to 78 bpm, with outliers surpassing 80bpms. This shows that despite the user having a somewhat good sign of heart rate beats, presented by the third quartile, higher values show patterns of poor cardiovascular health during rest. This shows that the user had sporadic events of tachycardia, which shows the variability in data among the scenarios.



(a) Healthy User



(b) Sedentary User



(c) User with CVD

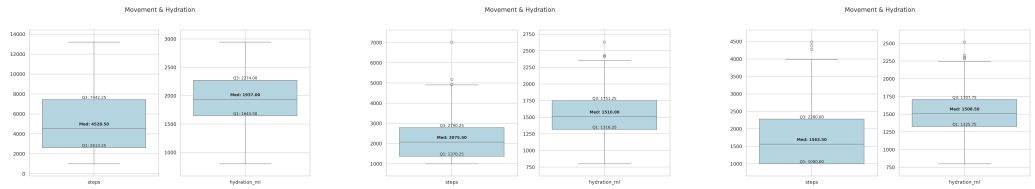
Figure 3.16: Heart Rate Metrics

Figure 3.16b illustrates the HR metrics of a user with poor exercise practice. The maximum HR values range from 111 to 174bpm, indicates that the user, despite doing some physical activity, the intensity remains relatively moderate. In this scenario, some of the values show signs of bradycardia, given the low cardiovascular

response, especially in cases where the user's maximum heart rate outliers fall below 110bpm. The minimal HR values span from 28 to 58bpm, with a median of 47bpm, indicate a moderate conditioned cardiovascular system. The RHR range between 40 and 65bpm is slightly above the ideal threshold with signs of elevated cardiovascular strain during rest. Similarly, most sleep RHR values exceed the recommended range of 40 to 60 bpm, which indicates a suboptimal nocturnal recovery that can be associated with increased CVD risk.

Lastly, Figure 3.16c displays various observations for heart rate metrics on a user with CVD. The RHR, during the day, shows a range of heart rate between 42 and 71 bpm, with some readings exceeding the 60bpm. The observed readings point towards cardiovascular stress, with values approaching 100bpm indicative of tachycardia tendency. Sleep also presents an irregular frequency between 45 and 107 bpm, with a median of 76bpm. Values outside the typical 40 bpm to 60 bpm threshold, particularly those exceeding 100 bpm or falling below 40bpm, may lead to heart deficiency. Regarding minimum HR values, range between 30 and 61bpm with a median of 47bpm, while maximum HR values range between 99 and 174bpm with a median of 138bpm. The maximum HR suggests that the user engages in episodes of moderate to high intensity physical activity, potentially indicative of intensive exercise efforts or periods of struggle demand, given the nature of the user scenario.

The following Figure 3.17a illustrates distinctive ranges of water intake and step count range. The user has a healthy lifestyle, given the step count range of 1000 to 13.274 steps, which reduces the development of chronic diseases. Over a course of days, the step counts below 4.500 indicate periods of less activity, which correlates with the consumption of water. The analysis of the daily water intake shows that the user drinks between 0.7L and 2.9L of water, but there are some days the user fails to reach the 1.5L, which greatly affects the body hydration function.



(a) Healthy User (b) Sedentary User (c) User with CVD

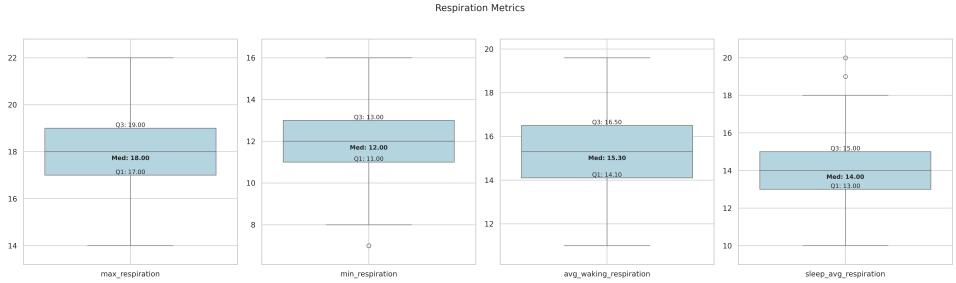
Figure 3.17: Movement and Water Intake

While the healthy profile demonstrates a healthy range of step counts and generally sufficient hydration, the sedentary scenario reveals a difference in range, as illustrated in Figure 3.17b. The movement ranged from 1.000 to 4.923 clearly shows that there are signs of sedentarism, despite some days reaching the optimal limit of 4.500. The presence of days with higher movement indicates that the user had a prolonged run or walk. The present water intake values also show a limited control

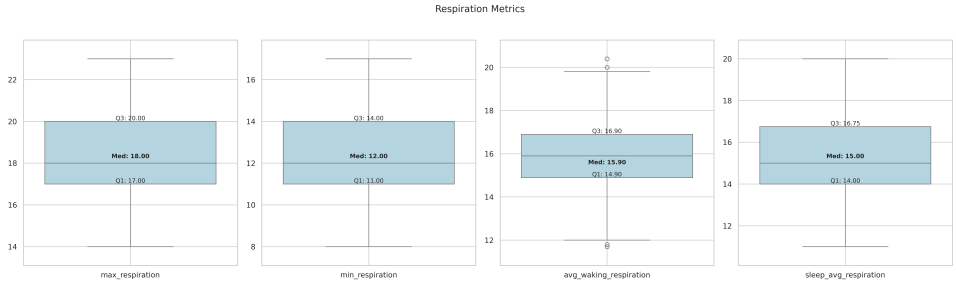
of hydration, with values ranging between 0.75L and 2.28L. The presence of multiple values below the 1.5L do not meet the general recommendations, since it affects the body health. These irregularities point towards a deficiency of fluid balance and a deprived active lifestyle.

Lastly, the following Figure 3.17c shows the movement and water intake metrics of a person with CVD symptoms. The user step counts, compared to previous scenarios, are lower, given the reduced levels of daily activity and movement. The step counts range stays below the recommended limit of 4.500, which reflects the influence of having CVD and the limitation of having a more active style. In parallel, the water consumption shows an inconsistent hydration habit with values ranging from 0.78L to 2.2L. The presence of outliers above the 2.3L suggests spikes in thirst, where the user feels the need to drink more water.

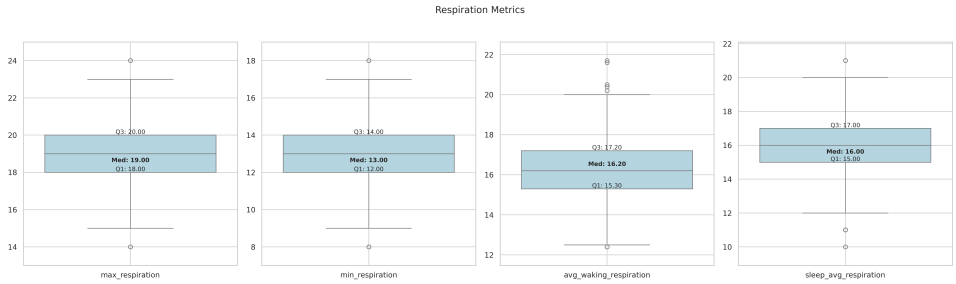
The following Figure 3.18a demonstrates various respiration rate metrics. Maximum RR values range from 14 to 22 breaths per minute, with elevated readings likely corresponding to periods of exercise, with induced respiratory demand. Minimum RR values span from 8 to 16rpm, however the data show an outlier value of 7rpm, possibly indicating bradyapnea provided by the overlapping data. Average waking RR values lie between 11 and $\approx$ 20rpm, with a median of 15rpm, showing a relatively balanced cardiorespiratory system within the optimal range of 12 to 20rpm, despite the minimum value being 11 rpm. The average sleep RR shows values within the range of 10 rpm and 18 rpm, with values outside the optimal range of sleep RR from 14 rpm to 16 rpm. These values may reflect sleep quality, stress recovery and potential respiratory inefficiencies. On the other hand, such reductions are occasionally observed in well trained individuals and may reflect enhanced oxygen delivery efficiency during rest. Although reduced respiratory rates may sometimes reflect improved recovery in trained individuals, they can also signal inefficiencies in sedentary users.



(a) Healthy User Respiration Rate



(b) Sedentary User Respiration Rates



(c) User with CVD Respiration Rates

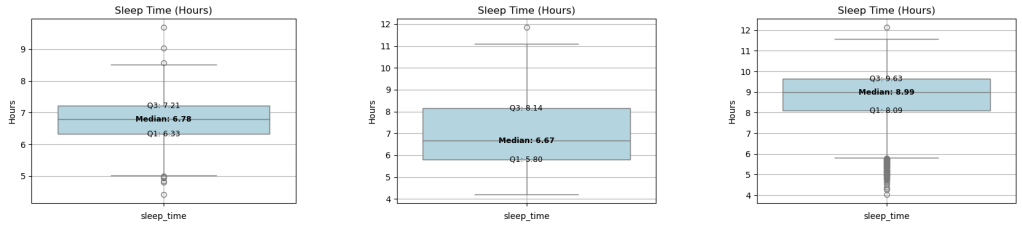
Figure 3.18: Respiration Metrics

Figure 3.18b shows this distinction, which presents RR measurements for a user with a sedentary lifestyle. The maximum RR values range from 14 to 23rpm, with various values surpassing 20rpm, indicates a higher strain on the body's cardiorespiratory system. The minimum RR values fall between 8 and 17rpm, with periods of notably reduced respiratory activity, potentially linked to rest or recovery phases. The daytime readings reflect somewhat stable respiratory patterns, from 12 to 20rpm, despite showing a few outliers above the 20rpm mark, and below the 12rpm mark. Regarding sleeping RR, the values show signs of respiratory irregularities, from 11 to 20 rpm, since most values fall outside the optimal range of 14 and 16rpm.

Lastly, Figure 3.18c illustrates several respiration rate metrics. Maximum RR displays values ranging between 15 and 23rpm with two outliers. One with 14rpm

and another with 24rpm. These values can be acknowledged, since low values indicate possible bradypnea during that day, while a high value can indicate a higher physical effort during exercise. Minimum RR values range between 9 and 15rpm, with outliers around 18rpm and another at 8rpm. In this particular case, given the nature of the user health status, having a low RR below 12rpm indicates the presence of bradypnea. In comparison, a minimum value of 18rpm may indicate the presence of tachypnea. Lastly, regarding average waking RR, most values are in an optimal range between ≈12 and 20rpm, with outliers between 20rpm and 22rpm. Lastly, the average sleep RR values range between 12 and 20 rpm. In this case, values below 14rpm may indicate signs of bradypnea or deep sleep respiration. On the other hand, we also observe an outlier of 21 rpm, which can be considered an episode of sleep disturbance.

Figure 3.19a illustrates sleep duration values from 5 to ≈9 hours, with a median around ≈7 hours. Instances falling below 5 hours suggest possible sleep deprivation, as highlighted by specific outlier data points. Conversely, three values exceeding 8 hours may indicate periods of sleep recovery.



(a) Healthy User (b) Sedentary User (c) User with CVD

Figure 3.19: Sleep Durations

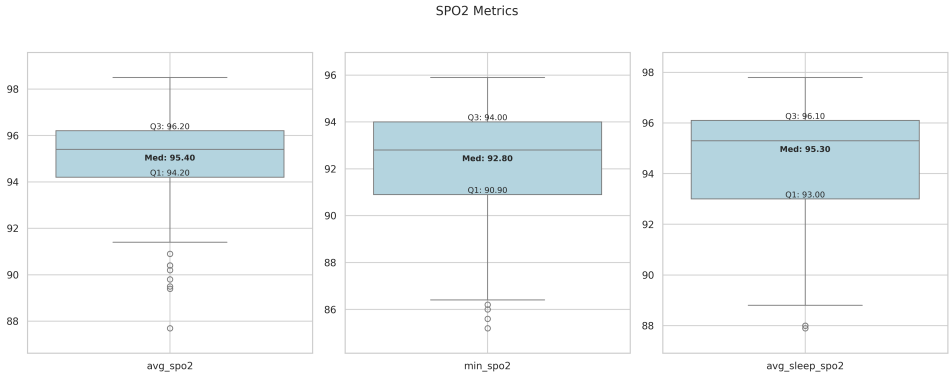
The following Figure 3.19b presents the sleep duration profile of an individual with a sedentary lifestyle. The user sleep ranges between ≈4 to 11 hours, with episodes of oversleeping, as well as sleep deprivation below the 6 hours. This extended sleep duration may reflect compensatory recovery from physical fatigue of low sleep duration or the presence of CVD.

Lastly, the illustrated Figure 3.19c shows the sleep duration between ≈6 and ≈12 hours, with a median of ≈9 hours. While most of the data fall within the recommended sleep duration for adults, observations below 6 hours suggest episodes of sleep deprivation. On the other hand, sleeping more than 8 hours indicates periods of compensatory recovery phase from periods of insufficient sleep or indications of CVD.

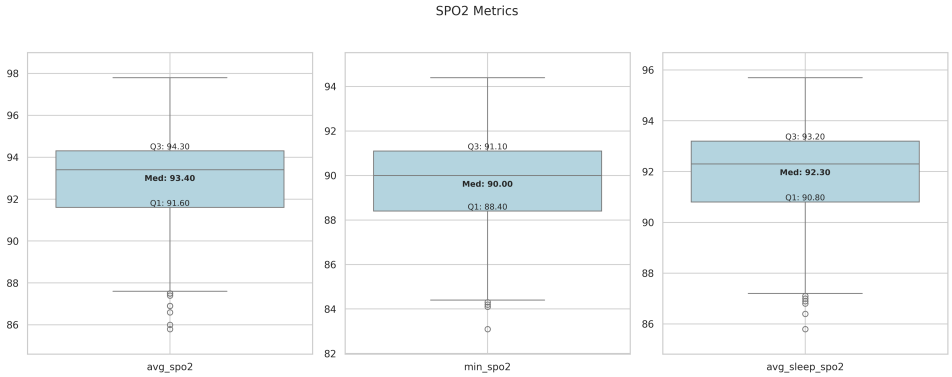
The data in Figure 3.20a presents daytime and nighttime oxygen saturation metrics. Average daytime SPO_2 values between ≈ 91% and 98% align well with the optimal range of adequate oxygenation in the body. We also observe occasional daytime dips reaching as low as 87%, although these are attributable to overlapping

desaturation from other profiles. These may also indicate situations where the user has sleeping positions that may show worse readings of SPO_2 levels, leading to situations of hypoxemia. Nighttime SPO_2 values span from 89% to 97.8%, with minimum readings ranging between 86.4% and 96.9% and a median of 92.8%. Readings below 90% are outliers representing data variability as a representation of hypoxemia.

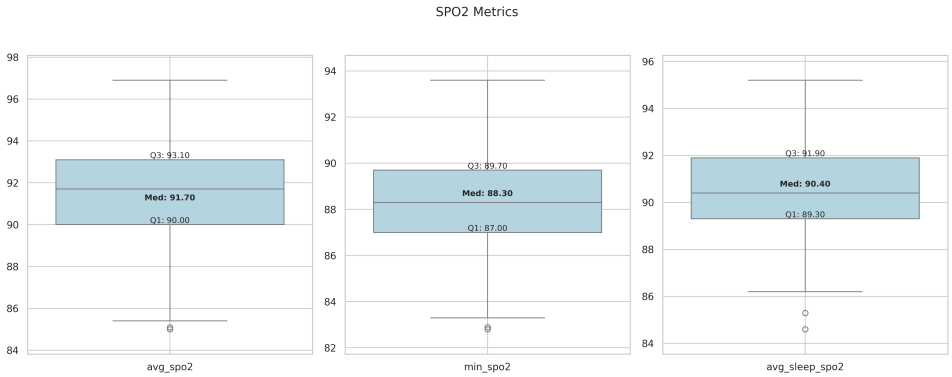
The sedentary profile, in contrast to the healthy scenario, as shown in Figure 3.20b, displays lower SPO_2 day and night readings. The SPO_2 during the day ranges between 87.7% to 97.8%, shows signs of hypoxemia below 90%, which limits the transport of oxygen over the body. Similarly, nighttime values trend slightly lower, with readings falling within the 87.4% and 95.7% range. Regarding minimum SPO_2, we also observe signs of hypoxemia with values ranging between 84.4% and 94.3%. Overall, nighttime and daytime dips suggest the presence of desaturation episodes and indicate the presence of hypoxemia.



(a) Healthy User



(b) Sedentary User



(c) User with CVD

Figure 3.20: Blood Oxygen Saturation

Lastly, Figure 3.20c presents daytime and nighttime SPO_2 metrics. Daytime SPO_2 values range from 85.6% to 96.9%, with a median of 91.7%. Several readings fall below the clinical threshold of 90%, suggesting instances of hypoxemia, observed, as well as, by two outliers values below 86%. During sleep, SPO_2 values range from 86.2% to 95.2%, again with two outliers between 84% and 86%. The recorded minimum SPO_2 values span from 83.6% to 93.7%. These patterns reflect episodes of hypoxemia, potentially attributable to the presence of CVD.

To validate and verify the rules we set on the CVD features, Figure 3.21 presents a heatmap of biometric correlations, derived from data collected from a smartwatch.

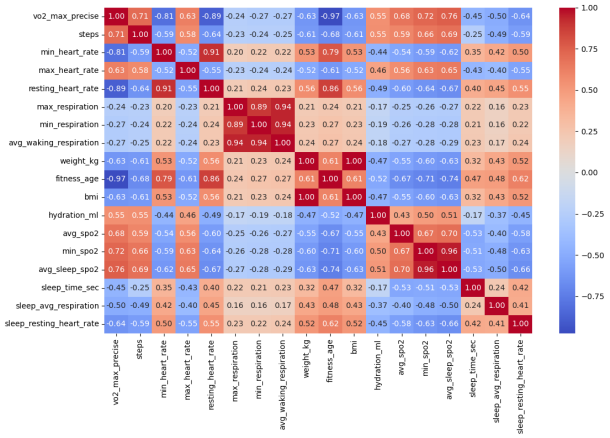


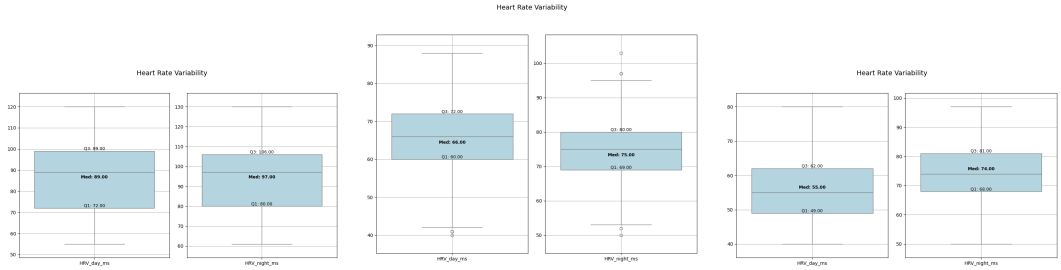
Figure 3.21: Heatmap Correlation

The heatmap shows a strong inverse correlation between VO_2max, resting heart rate, and fitness age, which conforms to the higher VO_2max scores being associated with lower resting heart rates and lower fitness age. Furthermore, people with a higher VO_2max score tend to engage with a higher step count to increase their performance score, which in turn demands a higher water intake. A positive correlation is also observed between steps and SPO_2, indicating that a higher intensity of physical activity contributes to improved blood oxygen saturation throughout the body. A positive correlation between steps and heart rate naturally indicates that a greater physical effort corresponds to an increased cardiac response. In addition, fitness age demonstrates a positive correlation with RHR, further reinforcing the link between cardiovascular efficiency and fitness ageing. A similar relationship is observed among weight, fitness age and BMI. Lower values across these metrics correspond to an improved body composition and overall health condition. Minimum SPO_2 also shows a strong positive correlation with average sleep SPO_2, as both are connected, according to Garmin. Finally, a positive correlation is present between minimum and maximum respiration rates and the average waking respiration rate, highlights the stability of respiratory patterns throughout different physiological conditions. Overall, the correlations observed in the heatmaps demonstrate the physiological relationships among the analysed variables and the real world physiological data. Overall, the correlations observed in the heatmaps demonstrate the physiological relationships among the analysed variables and the real world physiological data. These relationships confirm the CVD datasets reliability and consistency for their use in the application of ML algorithms to acquire good training/validation early prediction models.

3.4.2 T2D Data Profiles

Dataset Not Related on User Data

This dataset consists on selected metrics from the CVD dataset that is not related to the user data and, at the same time, are relevant to the assessment of T2D. The following Figure 3.22 demonstrates the HRV distributions during the day and night. Figure 3.22a demonstrates the HRV proxies for a healthy user. Given the age gap of the user, situated between 25-30 years, the user shows good metabolic efficiency with a higher variation between BPM in ms. During the day, the values range between 40 and 120 ms. These values align with the VO_2max values, where individuals with a higher CRF and lower RHR tend to exhibit a greater HRV, which further improves the production of insulin, reducing the chances of getting T2D condition. Regarding nighttime HRV, the range of values tends to be higher, due to periods of relaxed state of rest and recovery period, or the influence of deep sleep stages, which further contribute to this increase.



(a) Healthy User (b) Sedentary User (c) User with T2D

Figure 3.22: Heart Rate Variability Proxies

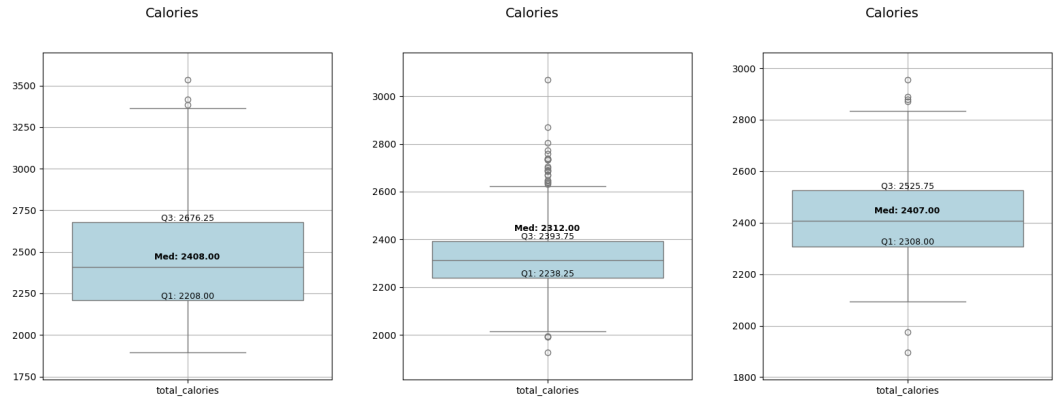
Regarding Figure 3.22b, shows the HRV data distribution for a user who follows a sedentary life. Compared to the previously healthy user scenario, the range of values is lower, which means that the user has a reduced level of activity that leads to the early presence of T2D leading to a higher resistance of insulin severity. This suggests that a sedentary life negatively impacts HRV, leading to a worsening of the CRF, given the values range of 42ms and 88ms. Nighttime HRV shows a stable range of variability between BPMs with a range of 53 and 95ms, despite a sleeping schedule outside of the recommended range of 6-8 hours, as previously observed in the CVD sedentary scenarios.

Lastly, Figure 3.22c shows HRV the data for a user with T2D. As mentioned before, the reduced physiological activity, as well as the reduced number of steps taken throughout the day leads to a worsening of the HRV. Given that the user has T2D, the range of BPMs is reduced, compared to the other scenarios. This indicates that the user, while having T2D, the number of BPMs shows a decrease of cardiovascular and metabolic efficiency, which leads to a higher resistance of insulin

production. Nonetheless, nighttime HRV remains high, with values ranging from 50 to 96ms, suggesting that some degree of cardiovascular recovery during rest or deep sleep is still preserved, despite the presence of T2D.

The following Figure 3.23 shows the data distributions of the total calories burned throughout the day. The user, with a healthy lifestyle, demonstrates a consistent energy expenditure pattern consistent with the expected values derived from the BMR formula. This estimation includes both the activity level attained with the attained VO_2max score, and the calories burned during exercise, which increases the normal production of insulin sensitivity. Overall, the user presents a good range of burned calories, from ≈1800 and ≈3125, which indicates a good and consistent metabolic efficiency.

Regarding Figure 3.23b, shows the total calories burned from a user with a sedentary lifestyle. As expected, the overall energy burned is significantly lower compared to the healthy scenario, despite the periodic energy expenditure outlined by the outliers above the 2600 calories. The reduced caloric output, ranging from ≈2050 and ≈2655, reflects minimal physical activity throughout the day, with most of the energy consumption derived from the user BMR rather than movement and exercise, which leads to a higher increase of insulin resistance. This limited number of calories burned suggests poor metabolic efficiency and a lack of cardiovascular stimulation, which, over time, can contribute to weight gain, reduced CRF and a higher risk of developing T2D.



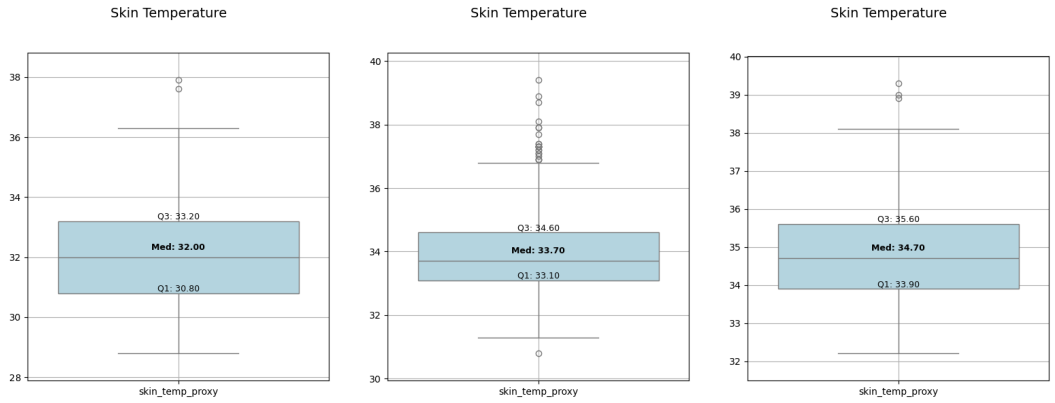
(a) Healthy User (b) Sedentary User (c) User with T2D

Figure 3.23: Total Burned Calories Proxies

Lastly, Figure 3.23c shows the total calories burned by a user diagnosed with T2D. In comparison with the previous scenarios, the user has burned more calories than when he was sedentary, while still being lower than when he had a healthy lifestyle. This increase is attributed to the elevated resting energy expenditure often associated with cardiac and respiratory system when the body requires more energy to maintain a normal physiological function, because of the worsening of glucose

and lipid metabolism. Additionally, light to moderate activity, can affect the caloric range, as seen by the caloric outliers above 2800, which contributes to a higher energy output, but most of the energy expenditure is supported by basic metabolic needs, instead of energy spent during exercise. Nonetheless, the values show a limited cardiovascular efficiency and reduced physical activity when compared to a healthy individual. Given the metabolic effects of T2D, the values emphasises the balance disruption of the metabolic and cardiovascular system, leading to a a higher insulin resistance.

The following Figure 3.24 illustrates the skin temperature regulation across the different user scenarios. Concerning Figure 3.24a, the user presents values ranging from ≈29Mestrado and 36.3Mestrado. This range indicates a stable thermoregulatory measure, typically observed in individuals with a healthy metabolic and peripheral circulation. The fluctuations within this interval correspond to natural variations, throughout the day, influenced by factors, like respiration rate, SPO_2, hydration level and physical activity, as defined in the formula. Specifically, higher respiration rates, lower step count and lower SPO_2, tend to increase the skin temperature due to elevated metabolic efficiency and reduced blood oxygen circulation. Adequate hydration also plays a vital role to moderate temperature increase by heat dissipation.



(a) Healthy User (b) Sedentary User (c) User with T2D

Figure 3.24: Skin Temperature Proxies

Regarding Figure 3.24b, illustrates the skin temperature for a user with a sedentary lifestyle with values ranging from ≈31Mestrado to 37Mestrado, with some outliers almost reaching 40Mestrado. These elevated periodic readings suggest irregular thermoregulation influenced by reduced physical activity and a limited CRF. According to the skin temperature estimation formula, lower steps with suboptimal RR and low SPO_2 can lead to higher skin temperatures which shows that the user has difficulties with heat dissipation and has metabolic issues. The presence

of outliers above 37Mestrado may also indicate periodic responses of dehydration, common in sedentary individuals that may start to show signs of T2D.

Lastly, Figure 3.24c shows the skin temperature for a user with T2D, whose values range between ≈32Mestrado and 38Mestrado, with periodic outliers above 38Mestrado. The elevated values and wider temperature range reflects difficulties in the thermoregulatory control due to direct consequences of poor circulation, glycation of vascular proteins and autonomic dysfunction, which hinders the heat dissipation. Overall, the broader range and higher temperature peaks highlight the difficulty of skin temperature control associated with T2D.

The following Figure 3.25 shows the correlation between metrics retrieved from CVD dataset, which are related to T2D, alongside four complementary metrics. Focusing on those four metrics and their relation with core T2D metrics, several patterns emerge. The skin temperature proxy shows a positive correlation with BMI and RHR, which indicates that a higher BMI and elevated RHR are associated with increased body temperature. Similarly, higher RRs also correspond to elevated temperature of the body because a higher activity and physiological stress can raise heat production and reduce the body ability to dissipate excess heat in an efficient way, leading to the presence of T2D conditions. Conversely, the higher the VO_2max and the greater the number of steps taken throughout the day, including those taken during exercise, the more stable is the skin temperature regulation, promoted by physical activity. Furthermore, decreases in HRV and SPO_2 are also associated with higher skin temperature, where reduced oxygen delivery and lower heart beat variability may contribute to body overheating. Regarding total calories, there is a positive correlation with BMI, RRs and RHR, both sleep and daytime, suggests that individuals with higher body mass or elevated RHR tend to spend more passive energy, even at rest. However, as the individual maintains a regular exercise activity and burn more calories, his RHR tends to decrease over time, which indicates an improvement of cardiovascular efficiency and exhibit a normal insulin production. In addition, consistent burn of calories contributes to skin temperature stabilization after a period of exercise, allowing calories to be used for activity rather than compensating for systemic stress.

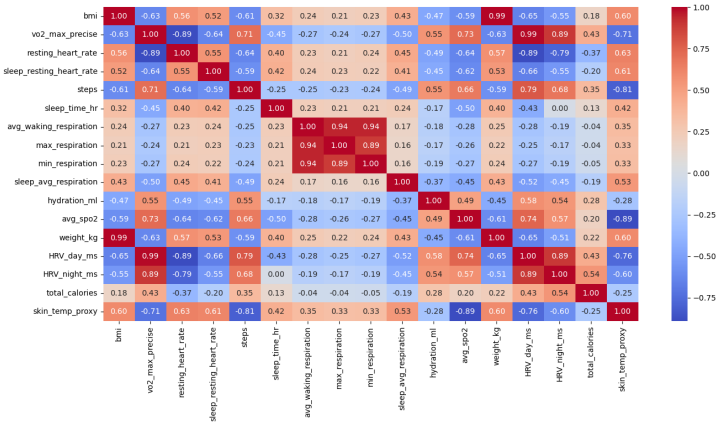


Figure 3.25: T2D Correlation Heatmap

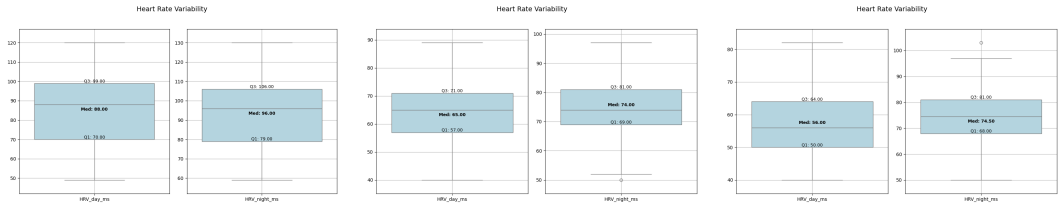
Lastly, regarding HRV metrics, there is a strong relationship with VO_2max, as individuals with higher CRF exhibit a better and sustained HRV, supported by a high number of daily steps. Hydration levels also correlate positively with HRV since it helps regulate the body fluids and gives a stable cardiovascular efficiency and improved insulin levels. Moreover, a higher blood oxygen saturation further enhances the heart beats variability which helps with the transport and circulation of oxygen and the attachment of proteins to sugar to reduce the levels of glucose and improve overall health.

Dataset Based on User Data

This dataset consists on selected metrics from the CVD dataset that is related to the user data and, at the same time, are relevant to the assessment of T2D. The following Figure 3.26 demonstrates the HRV data distribution across different scenarios of T2D. Figure 3.26a show the HRV data for a healthy user with values ranging between 49 and 120ms, with a median of 88ms. As mentioned in the previous section, given the user age gap, these HRV values fall within the expected range for individuals with a good CRF. Given this dataset, based on user data, the distribution further highlights the relation with individuals who have a lower RHR and a higher HRV values, showing a great cardiovascular resilience and insulin production. Furthermore, nighttime HRV values remain higher than daytime, indicating an effective recovery during rest.

Regarding Figure 3.26b, shows the data distribution of a user who follows a sedentary lifestyle with early signs of T2D. Compared to the healthy scenario, the number of heart rate variations are lower due to the worsening of VO_2max, which indicates a decline on cardiovascular fitness and a higher difficulty of insulin production. This decrease is also complemented by the decrease of RHR, since a lower RHR in sedentary conditions influence a diminished cardiovascular response, which

in turn affects the HRV condition. In terms of nighttime values, they tend to be higher than daytime. However, despite the user condition, the HRV still shows some degree of recovery during rest.



(a) Healthy User (b) Sedentary User (c) User with T2D

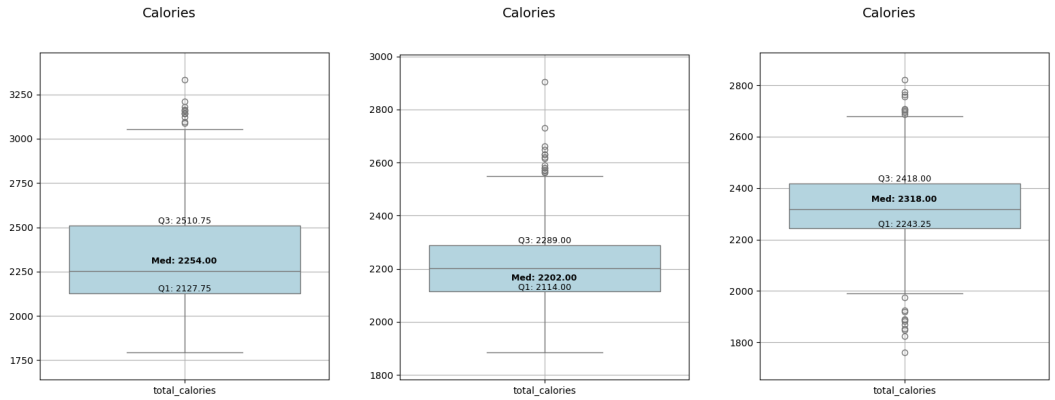
Figure 3.26: Heart Rate Variability Proxies

Finally, Figure 3.26c shows HRV of a user with T2D conditions. Compared to previous scenarios, the user's HRV are lower due to complications of T2D that influence the worsening of his health. This decrease is associated with a decrease of exercises activities leading to a higher RHR and a diminished flexibility. Moreover, lower exercise levels lead to reduced levels of VO_2max further limiting the heart to adapt to daily physiological demands. Nighttime values remain slightly higher than daytime, which suggest some indication of recovery periods during rest, but still below the levels observed in healthy or even sedentary scenarios. Overall, the combination of reduced HRV, higher resting heart rate, and lower exercise activities highlights the impact of T2D on cardiovascular and metabolic function.

Figure 3.27 shows the total calories burned on a daily basis. In this scenario in Figure 3.27a, the user daily caloric expenditure ranges ≈1800 and ≈3150 *Calories* (Kcal), with occasional outliers above 3100Kcal. The variations reflect the combined effects of BMR, exercise activity and intensity applied each day. Although the user maintains a median caloric expenditure of 2254Kcal, throughout the day, there are cases where the user, on the lower end of the interval, indicate periods of reduced activity, serving as recovery days followed by intense physical days of effort. Overall, the user has a healthy metabolic system, where the energy expenditure adapts to the varying activity demands with a normal insulin production and efficient glucose utilisation.

The following Figure 3.23b, presents the caloric distribution of a sedentary user. Compared to the healthy scenario, the interval of caloric expenditure is slightly lower values, which indicates a worse margin due to metabolic function. Furthermore, the upper end of the interval is lower, which indicates that the user has a reduced physical activity leading to such values and may indicate the start of signs of insulin resistance. Occasional outliers reach nearly 3000Kcal, likely to sporadic periods of increased activity intensity. Some of the present values, show related signs of insulin resistance, which are commonly observed in sedentary individuals or those at risk of T2D. Lower VO_2max limits the CRF, higher RHR shows a greater cardiovascular

stress and slight skin temperature increase, all of which contribute to inefficient energy utilisation. Combined with the BMR formula, these factors help explain the observed caloric distribution, resulting in the metabolic challenges that precede the evolution of T2D risk.



(a) Healthy User (b) Sedentary User (c) User with T2D

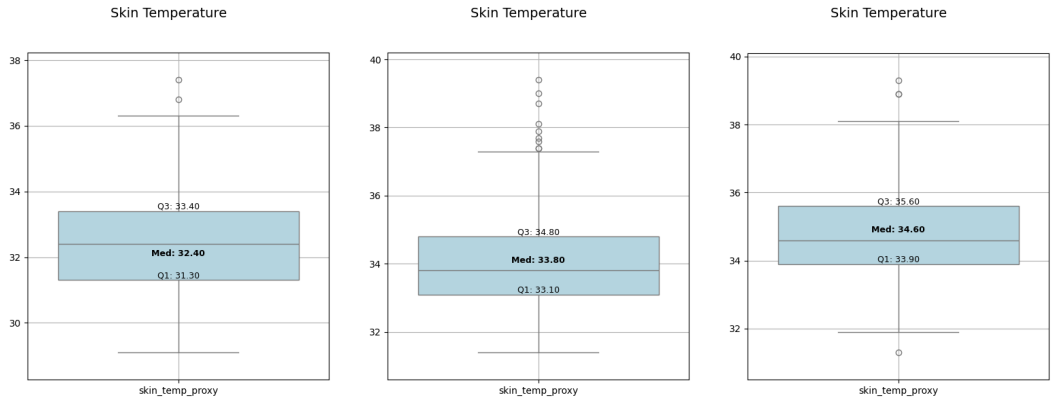
Figure 3.27: Total Burned Calories Proxy

Finally, 3.27c shows the calories burned of a user with T2D on a daily basis. Given the nature of this scenario, the caloric expenditure is higher than the sedentary scenario, because the body requires more energy to maintain the balance of the human body due to insulin resistance and metabolic inefficiency. Given these conditions, despite lower levels of physical activity the body demands lead to a higher increase of energy expenditure at rest. In addition, higher RHR and fluctuations in skin temperature reflect an increase of cardiovascular and thermoregulatory effort, while reduced VO_2max indicates a decreased of CRF, further contributing to the failure of energy usage.

Figure 3.28 shows the skin temperature over different T2D scenarios. Regarding 3.28a shows the user skin temperature ranging from 29Mestrado to ≈36Mestrado with a median on 32.4Mestrado. The observed values indicate an efficient regulation of skin temperature response, consistent with a healthy metabolic state. This skin temperature reflects a good CRF and RR activity, supported by adequate water intake and normal insulin sensitivity. This regulation helps maintain optimal heat dissipation during intense physical activity and during recovery periods.

Concerning a sedentary user scenario, Figure 3.28b presents a range of values from ≈31Mestrado to ≈37Mestrado, with a median of 33.8Mestrado. The presence of outliers above 37Mestrado shows periods of high skin temperature, indicating a thermoregulatory response over metabolic or intense activity exhaustion. Furthermore, these values can be attributed to an inadequate hydration and suboptimal SPO_2 and RR, which are leading indications of insulin resistance.

Lastly, Figure 3.28c shows the user skin temperature that has T2D. The present values exhibit a more diverse range close to 40Mestrado. Given the nature of this scenario, the user shows difficulties is balancing his temperature, which are direct consequences to poor oxygen circulation, glycation of vascular proteins and autonomic dysfunction which difficulties the heat dispersion. In addition, the user low physical activity and worsening of his heart and respirations rates, further influences the worsening of his health which is associated with T2D.



(a) Healthy User (b) Sedentary User (c) User with T2D

Figure 3.28: Total Burned Calories Proxy

The following Figure 3.29 illustrates the correlations between metrics associated with T2D, along with four additional complementary features. The examination of the interplay between these four features, with core T2D features reveals notable patterns that can be compared to real world data metrics.

Skin temperature proxy demonstrates a positive association with BMI and RHR, where a higher body mass and elevated resting heart rate are linked to the increase of skin temperature, a common factor of relevant to the development of T2D risk. Similarly, higher RR, shows physiological stress and heat production, which can challenge the body ability to dissipate the heat. In contrast, higher VO_2max with the inclusion of steps, support stable temperature regulation, which highlights the thermoregulatory benefits of physical activity.

Reduction in HRV and SPO_2 are likewise associated with higher skin temperature. The blockage of blood oxygen and proteins to manage sugar levels further contribute to overheating of the human body. Total calories burned also show positive correlations with VO_2max, BMI, steps and HRV across daytime and sleep time periods. Regular physical and intense activity helps with the caloric expenditure improving the cardiovascular and metabolic health of the user. Similarly, sustained energy expenditure further contributes to reduce the HRV, RHR and RR and skin temperature, post activity, allowing the body to effectively manage the stabilization of cardiovascular efficiency and better insulin regulation.

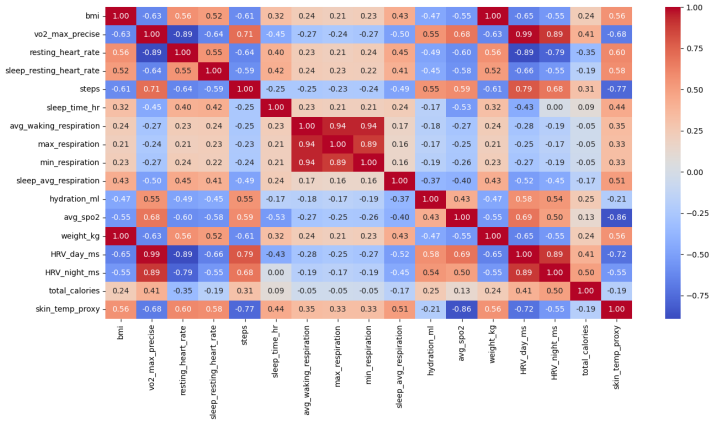


Figure 3.29: T2D Correlation Heatmap

Finally, HRV metrics are strongly correlated with VO_2max, where individuals with a higher CRF tend to have a prolonged time between heart beats, supported by the high number of steps on a daily basis. Adequate water intake also has a positive correlation with HRV, which helps balance the body fluids, regulate the skin temperature and helps with cardiovascular and metabolic function. Furthermore, a high SPO_2 enhances the HRV with the transport of oxygen and reducing glucose related protein modifications.

The following steps involve applying feature selection techniques to identify the most informative variables, to reduce dimension and eliminate redundant or less relevant features, thereby improving model performance and its interpretability in later applications for machine learning classification algorithms.

3.4.3 Data Standardization

In this section, we explain the application of a standardisation technique used to preprocess the datasets by adjusting the scale of the input features.

Standardisation of a dataset is a common preprocessing step for many machine learning algorithms. This process of feature scaling consists of transforming the variables in a dataset to a similar range and distribution so that machine learning estimators assume equal importance among features. The transformation process rescales the data to an interval between [0,1] or [-1,1], depending on the chosen technique. Although, the presence of outliers can often distort the sample mean and variance, leading to wrong representations of the data. In such cases, using the median and the interquartile range often provides a better alternative. To this end, we apply the Robust Scaler, a technique that scales features using robust statistics to outlier margins. Unlike standard scalers, Robust Scaler often results in a broader

range of transformed values. However, it preserves the integrity of the data distribution given such extremes. The following formula shows the application of the Robust Scale technique transformation over the datasets.

$$X_{\text{scaled}} = \frac{X - Q_2}{Q_3 - Q_1} \tag{3.2}$$

The formula subtracts the median of each record and divides it by the interquartile range, which reduces the effect of outliers while maintaining the distribution of non-outlier values. This formula can be divided into four parts:

- X = the original value on the dataset
- Q1 = the first quartile (25th percentile)
- Q2 = the second quartile (50th percentile)
- Q3 = the third quartile (75th percentile)

Using this formula, we standardise the features to balance their contribution equally, reduce their bias and improve the data quality to make it easier to understand and analyse. The scaling process in this example of code 3.2 starts by loading a CVD dataset, containing the consolidated data for their respective scenarios.

```
# Load dataset and separate the features from the target columns
df = pd.read_csv("cvd_dataset_A.csv")

df_process = df.drop(columns=["calendar_date"])
y_target = df_process['cvd_risk']
x_features = df_process.drop(columns=['cvd_risk'])

# Scale features
scaler = RobustScaler()
x_scaled = pd.DataFrame(scaler.fit_transform(x_features, y_target)
    , columns=x_features.columns)
```

Listing 3.2: Data Standardization using Robust Scaler

The `calendar_date` column is removed in this phase due to its categorical nature and the unnecessary scaling procedures for this attribute. Another reason to exclude this column is that the data will not be shuffled since it represents temporal data and the scaler can not interpret the date string format. With the removal of the date column, we separate the dataset between the features and the target `cvd_risk` column, to preserve the target variable for label encoding, while the scaling process is applied exclusively to the input features.

Parameter	Value	Description
with_centering	True	Centers the data around 0 before scaling. Useful when the scale and center vary a lot, especially with outliers.
with_scaling	True	Scales the data into the interquartile range.
quantile_range	(25.0, 75.0)	Defines the interquartile range to scale the data. The higher the range, the more sensitive it is to outliers.
copy	True	Creates a copy of the array with the scale data, instead of inplace scaling directly.
unit_variance	False	Scales data so that normally distributed features have a variance of 1.

Table 3.3: Robust Scaler Parameter List

Table 3.3 shows the delimited parameter list from which we run the `RobustScaler()` function, after splitting the data.

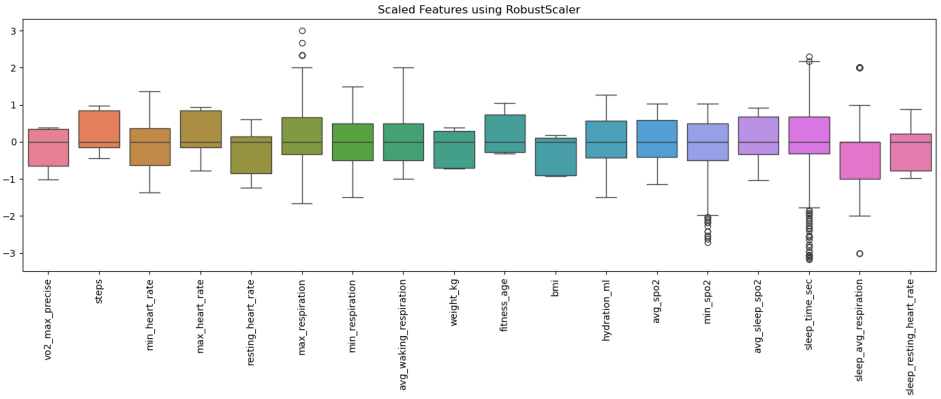
Robust scaler uses a method called `fit_transform()`, where it fits the transformer to the features and target column and returns the transformed version of the initial feature dataset. Table 3.4 summarises the parameters utilised during this process.

Parameter	**Value**	**Description**
X	x_features	Represents the input features
y	y_target	Represents the target variables. None parameter value if unsupervised transformation
**fit_params	None	Additional Parameters

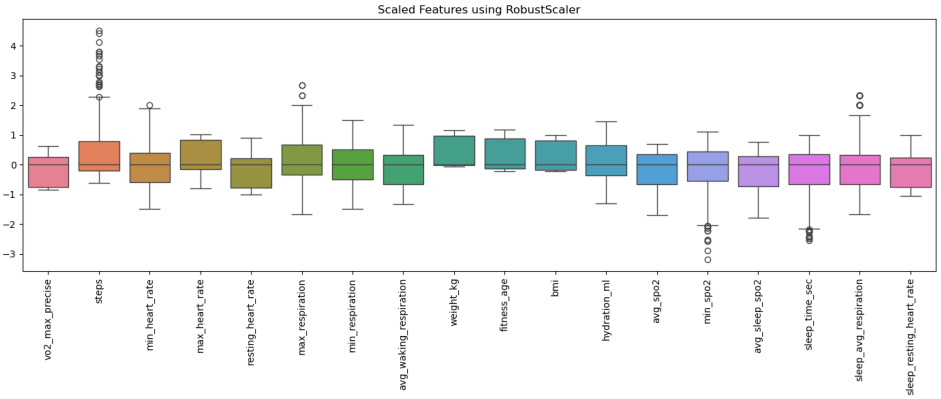
Table 3.4: Robust Scaler Fit Transform Parameter List

The application of the Robust Scaler is used in every CVD and T2D datasets to a common scale so that features with higher representation do not dominate the model learning process.

After the application of the Robust Scaler technique, Figure 3.30, shows an example on how the datasets transformation become standardised.



(a) CVD Dataset Not Related to User Data



(b) CVD Dataset Based on User Data

Figure 3.30: Standardisation with Robust Scaler

In Figures 3.30a and Figure 3.30b, we observe that the application of the Robust Scaler technique results in a broader data range, approximately between [-3, 4] variance. The results show a greater dispersion of the feature values, particularly influenced by the presence of outliers, which the scaler is designed to handle more effectively than standard standardisation methods. These outliers are acknowledged as they represent meaningful variations in the data rather than noise, and their influence is carefully managed without being entirely suppressed. The following step involves the application of a feature selection technique to identify and retain the most relevant variables for the predictive model, thereby improving performance and reducing potential overfitting.

3.5 Feature Selection

In this step, we explored different approaches to understand the most viable technique for this study, given the limited hardware discussed in previous chapters. While analysing several key studies that applied feature selection techniques, we

chose RFE. Recursive Feature Elimination is a technique used for supervised ML problems that applies an external estimator to evaluate the importance of input features.

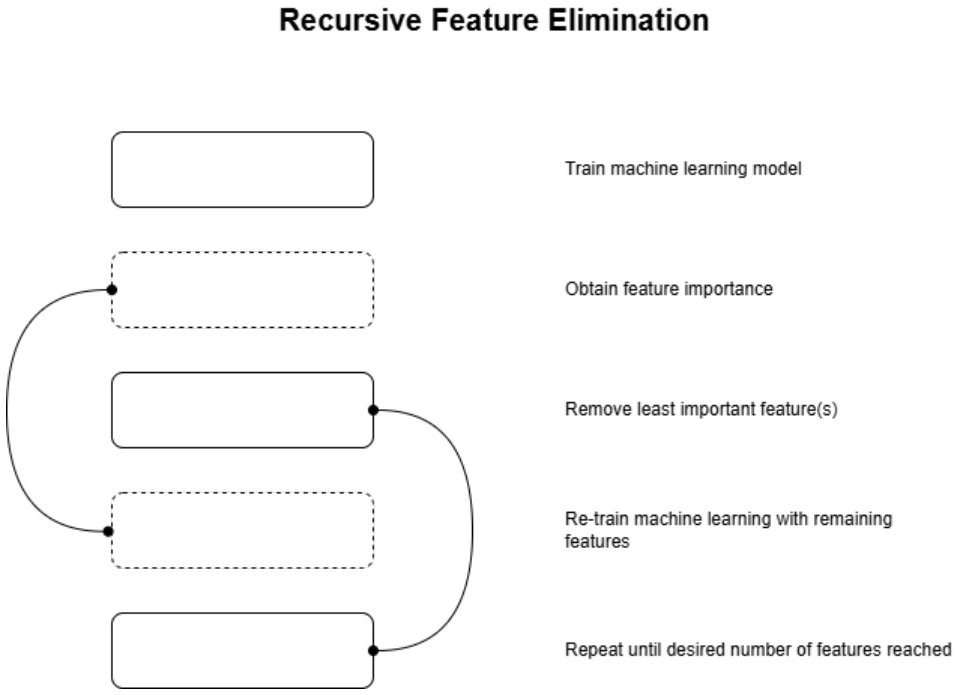


Figure 3.31: Recursive Feature Elimination Process

Figure 3.31 shows the steps for the implementation of the feature selection process. Using a recursive process, eliminates the least informative features to isolate a subset of features that influences the model's predictive capability. Feature importance is typically assessed via model attributes, such as `coef_` for linear estimators, or `feature_importances_` for tree based models. After ranking features based on their contribution to the target outcome, it discards the least significant feature, while the model retrains using the reduced feature space. This iterative procedure repeats until a predefined number of features remain. The recursive nature of RFE allows the selection algorithm to continually adapt to structural changes in the model, while enhances its ability to identify the best subset of predictive variables to reduce the computational cost by working with fewer features. The RFE presents challenges that include the possibility of discarding important features in the initial steps, particularly when individual variables show low significance but possess high correlations with other features. Each iteration of RFE requires retraining the model, and reevaluating its performance, which can be resource intensive if the dataset is not sufficiently balanced. However, despite its challenges, it consists of a powerful technique for feature selection due to its ability to improve model interpretability and performance by systematically eliminating irrelevant or redundant features. Moreover, in contexts where hardware capabilities are limited, reducing

the dimensionality of input data contributes to enhanced efficiency in model training and results in low resource usage. In this context, we selected the RF algorithm as the external estimator within the RFE frameworks based on statistical and practical considerations.

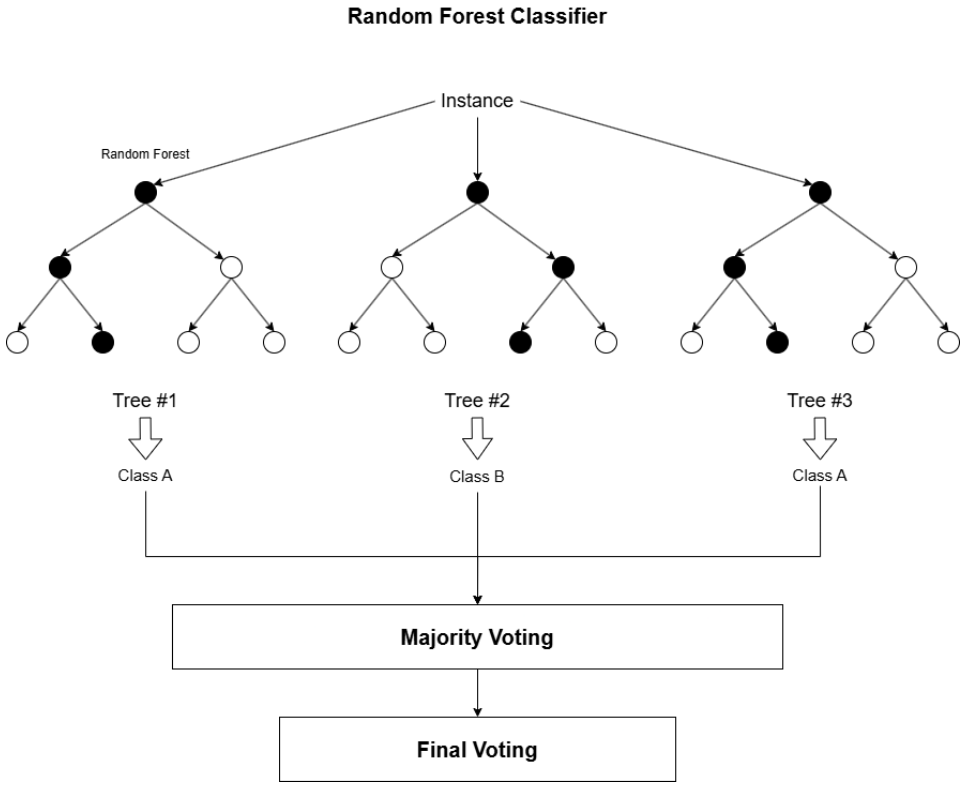


Figure 3.32: Random Forest Classifier

Figure 3.32 presents a simplified representation of the application of the RF classifier. From a statistical point of view, Random Forests are suited for classification problems due to their ability to handle multicollinearity and complex, non-linear decision boundaries. This algorithm builds multiple decision trees and aggregates their predictions using majority voting to generate the final output. Each decision tree trains using a distinct subset of features to reduce the correlation between trees and their susceptibility to overfitting.

From the practical perspective, Random Forest classifiers train efficiently by automatically selecting relevant features for prediction and their capability to deal with noise. Another aspect of RF is that it can work in parallel, which is ideal for classification scenarios constrained by limited computational resources. However, despite being identified as a good model, sometimes it can be challenging to interpret its results, given the large number of trees. With many decision trees, its complexity increases, making the model computationally expensive, specifically for large datasets. Additionally, during the random selection of features at each split,

if several features carry similar information, the model may consistently favour one over the others simply due to chance or slight differences in variance. Nonetheless, their built-in capability to estimate feature relevance makes it an effective tool for a recursive feature selection workflow. Its use can be beneficial for the early prediction of CVD and T2D due to its capability of capturing complex relations between multiple risk factors.

To further improve the workflow performance, we use the Randomised Search Cross Validation (RandomizedSearchCV) algorithm for hyperparameter tuning of RF. Rather than evaluating all possible combinations, it performs a randomised search over a specified parameter grid, evaluating a subset of parameter combinations using cross validation. Additionally, given the temporal structure of the dataset, we apply Time Series Split Cross Validation (TimeSeriesSplitCV) as the cross validation strategy to preserve chronological data integrity.

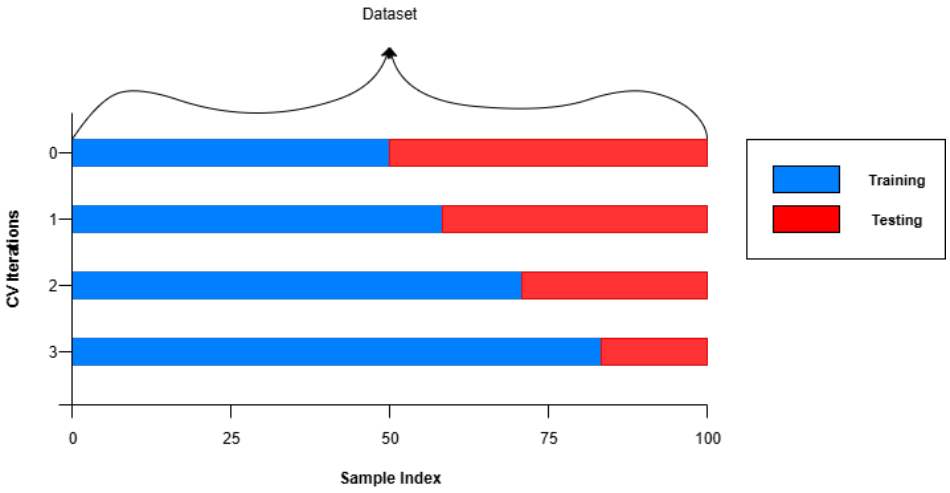


Figure 3.33: TimeSeriesSplit Cross Validation Process

In contrast to conventional cross validation techniques, TimeSeriesSplitCV mitigates data leakage by making sure that training data precedes the validation data in time, as seen in Figure 3.33. This approach conforms to this case study because it preserves temporal dependencies and provides a more realistic evaluation of a daily monitoring dataset.

The following code block 3.3 shows the implementation process of the feature selection. We split the data into 80% for training/validation and 20% for testing. The feature selection was performed solely on the training data. In this implementation, we use RFE with a RF as the base estimator. The RFE method recursively removes the least important features based on the estimator's feature importances, iteratively identifying the optimal subset of predictors, with iterations ranging from 25, 50, 75 and 100. To preserve the temporal integrity of the data and avoid look ahead bias, we use a TimeSeriesSplitCV strategy with various splits, ranging from

5, 7 and 10, for cross validation, to create more train/test combinations across time for performance evaluation over unseen data.

```
#divide the dataset into 80% for training and 20% for testing
x_train, x_test, y_train, y_test = train_test_split(
    x_scaled,
    y_target,
    test_size=0.2,
    shuffle=False  # important for time series to preserve order
)

rf = RandomForestClassifier()
rfe = RFE(estimator=rf)

tscv = TimeSeriesSplit(n_splits=5) # 5, 7, 10 splits

# Define the grid parameters to search
param_grid = {...}

scoring = {
    'f1_weighted': 'f1_weighted',
    'balanced_accuracy': 'balanced_accuracy',
    'precision_weighted': 'precision_weighted',
    'recall_weighted': 'recall_weighted'
}

random_search  = RandomizedSearchCV(
    estimator=rfe,
    param_distributions=param_grid,
    n_iter=25,
    cv=tscv,
    scoring=scoring,
    n_jobs=4,
    random_state=42,
    refit='f1_balanced'
)

start_time = time.time()

# Fit the model
random_search.fit(x_train, y_train)

end_time = time.time()
execution_time = end_time - start_time
```

Listing 3.3: Feature Selection and Hyperparameter Tuning using RFE with Random Forest optimised via RandomizedSearchCV and Time Series Cross Validation

The model hyperparameters are tuned using RandomizedSearchCV, which randomly samples a defined number of parameter combinations from a defined `param_grid` space, presented in Table 3.5.

Parameters	Description
feature_selection__n_features_to_select	Number of top features to select after ranking by the internal estimator (RandomForest). Smaller values reduce dimensionality and computational cost.
feature_selection__estimator__n_estimators	Number of trees in the RandomForest. More trees often improve stability and accuracy but increase training time.
feature_selection__estimator__max_depth	Maximum depth of each tree. Limits how "deep" the tree can go. Helps control overfitting. Avoiding None prevents overly complex trees and saves memory.
feature_selection__estimator__min_samples_split	Minimum number of samples required to split an internal node. Higher values make the model more conservative.
feature_selection__estimator__min_samples_leaf	Minimum number of samples required at a leaf node. Higher values smooth the model and prevent overfitting. Important in imbalanced datasets.
feature_selection__estimator__max_features	Number of features to consider when looking for the best split. 'sqrt' is default for classification, 'log2' or float fractions may boost performance but affect speed.
feature_selection__estimator__bootstrap	Setting True means each tree is trained on a random subset of the data, increasing diversity and robustness, instead of using the full dataset to train the tree.
feature_selection__estimator__max_samples	Fraction of the training samples used for fitting each base estimator if bootstrap=True. Controls diversity among trees and reduces overfitting.
feature_selection__estimator__criterion	Function to measure the quality of a split. 'gini' is faster and often preferred; 'entropy' uses information gain and may yield more precise splits.
feature_selection__estimator__class_weight	Helps with imbalanced datasets by giving more importance to minority classes, if 'balanced'.

Table 3.5: Description of Hyperparameters Space Used in Random Search for Feature Selection with Random Forest Based Recursive Feature Elimination

Returning to code block 3.3, we select multiple evaluation metrics for the scoring dictionary: balanced precision, balanced recall, balanced accuracy and balanced F1 score. We select the balanced F1 score as the refit metric, among the available metrics, since it combines precision and recall for each class. This choice allows a reliable assessment of the model's performance in a multiclass classification problem. Although the dataset is relatively unbalanced, the choice of `f1_weighted` over `balanced_accuracy` gives a fair view on how well the model predicts each class.

We trained the model on standardised features `x_train` and their corresponding target labels `y_target`, where we timed the execution time to assess the efficiency of the model. Upon completion, we extract the best feature subset from the fitted RFE model, including the selected features, best hyperparameters and overall cross validated performance. This approach ensures a robust feature selection and practical tuning model tailored to time dependent data.

Regarding decision tree based models, such as Random Forests, it is also important to determine the optimal way to split nodes to improve classification performance. This is done using impurity measures, with `Gini` and `Entropy` being the most common. Gini impurity measures the probability of incorrectly classifying a randomly chosen sample, assuming it is labelled randomly according to the class distribution in the node

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{3.3}$$

where p_i is the probability of the class i in the node.

Entropy, on the other hand, comes from information theory and quantifies the level of uncertainty or disorder in a node. When a node is split, the reduction in

entropy is called information gain

$$Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{3.4}$$

where p_i is the probability of the class i in the node.

Both functions aim to create pure subgroups of data. However, while Gini is generally faster to compute, entropy may be more sensitive to imbalanced class distributions.

With the optimal subset of features identified, the following section outlines the training and testing phase using the selected features obtained through the feature selection process. This phase encompasses the construction of predictive models, the evaluation of their performance across multiple validation metrics and the optimisation of hyperparameters to ensure generalisability and robustness.

3.6 Machine Learning Classification Methods

Following the application of feature selection, we implement two supervised classification algorithms, SVM and LR. As part of the SVM function, we use the Support Vector Classification (SVC) method, which is designed for classification problems. The SVC, as seen in Figure 3.34, aims to find the optimal hyperplane to separate data points into different classes with the maximum margin.

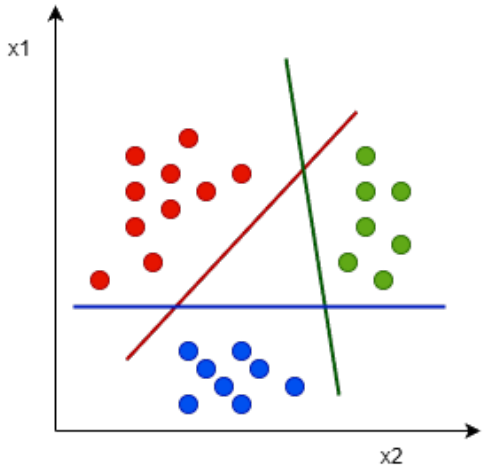


Figure 3.34: Support Vector Classification Data Separation

The margin is the distance between the hyperplane and the nearest data points, the support vectors, on each side. In cases where data is not linearly separable, kernel functions are used to project data into higher dimensions where separation is possible. The code segment in A.1 shows an example of the script for training

and tuning the SVC method to train and validate the CVD risk records using a TimeSeriesSplitCV strategy, with 5, 7 and 10 cross validation folds.

Parameters	Description
svm__C	Regularization parameter. A lower value leads to a simpler decision boundary by penalizing complexity, while a higher value tries to classify all training points correctly, risking overfitting
svm__kernel	Specifies the kernel type used in the algorithm; Linear creates a linear decision boundary; rbf (radial basis function) allows for non-linear classification.
svm__gamma	Defines how far the influence of a single training example reaches. scale (default) uses 1 / (n_features * X.var()), and auto uses 1 / n_features. It affects the decision region when using non-linear kernels.
svm__tol	Tolerance for stopping criteria. Lower values mean the algorithm runs longer and may converge to a better solution.
svm__shrinking	Whether to use the shrinking heuristic, which can improve optimization speed by ignoring less important support vectors during training.
svm__class_weight	Handles class imbalance. 'balanced' automatically adjusts weights inversely proportional to class frequencies, while None treats all classes equally.
svm__decision_function_shape	Specifies the strategy for multiclass classification. ovr (One-vs-Rest) trains one classifier per class.
svm__break_ties	In case of a tie during prediction, setting this to True uses decision function values to break the tie. Provides slightly more consistent predictions, especially for multiclass problems.

Table 3.6: Support Vector Classification Hyperparameter Distribution Space

The SVC classifier is initialised with a grid of hyperparameters `param_dist` to perform an extensive search. While RandomizedSearchCV offers faster results by sampling a subset of features, as applied during feature selection, we want to get the most optimal model by exploring all possible combinations in this phase, which can be seen in Table 3.6, with Grid Search Cross Validation (GridSearchCV). To preserve the temporal structure of the data, the code follows the TimeSeriesSplitCV for cross validation strategy to ensure that future data is not used to predict the past. To evaluate the model across different temporal segments, we employ 5, 7 and 10 splits. This variation will allow us to see how the model performs across time, while balancing the training set size and validation depth. To evaluate the model performance, we maintain the use of weighted metrics, given the rather balanced dataset mentioned before, which includes F1 score, precision, recall and accuracy. The training process is timed to assess computational efficiency and the grid search results are stored in a DataFrame for analysis. As a last step, the best performing model's parameters and the model are saved using the joblib Python package for future testing phases.

In terms of LR, the application of this function follows the same structure as the one applied to SVM, with some differences in the model configuration. In the code block A.2, we extend the LR function to a multiclass problem using the parameter `multi_class` with `multinomial` value. Multiclass logistic regression uses the `softmax` function as seen in Figure 3.35, instead of a `sigmoid` function that converts input values into probabilities ranging between 0 and 1 for binary outcomes.

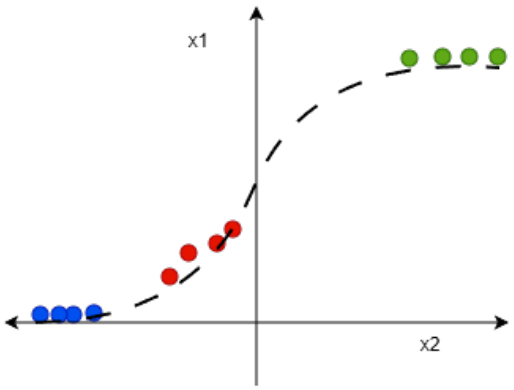


Figure 3.35: Multinomial Logistic Regression Data Representation

This function enables multiclass classification problems by generating a probability distribution over all possible classes, so the cumulative probability equals one. In addition to this multiclass configuration, LR is subjected to a hyperparameter optimisation distribution space, as shown in Table 3.7.

Hyperparameter	**Descrition**
lr__penalty	Controls the type of regularisation and prevents the model from overfitting
lr__tol	Controls the stopping criterion. Smaller values = more precision but slower training.
lr__C	Inverse regularization strength. Low values apply stronger regularisation, allowing some misclassifications to improve generalisation. Ideal for noisy data. High values reduce regularisation, forcing the model to fit the training data more closely, which may lead to overfitting.
lr__dual	Dual formulation is only implemented for l2 penalty with liblinear solver. Preferable when dual=False where n_samples >n_features.
lr__fit_intercept	Specifies whether to include an intercept term in the model. If set to True, the model learns a bias term to shift the decision boundary. If False, the model assumes the data is already centered and forces the decision boundary to pass through the origin.
lr__class_weight	Adjusts the weight assigned to each class during training to address class imbalance.
lr__solver	Algorithm to use in the optimization problem. Among the several solvers, only liblinear does not apply to multiclass problems.
lr__warm_start	If set to True, the model retrains its previous solution when .fit() is called, allowing iterative training or reuse of learned coefficients
lr__max_iter	Specifies the maximum number of iterations the solver will run during model optimization. Increasing this value can help with convergence, especially for complex models or poorly scaled data

Table 3.7: Logistic Regression Hyperparameter Distribution Space

This set of parameters are essential to control the model complexity and its behaviour in order to find the most optimal configuration.

On the other hand, during training, we convert the feature set into a C-contiguous format to enhance the efficiency of the SVC and LR execution, and to avoid data copy.

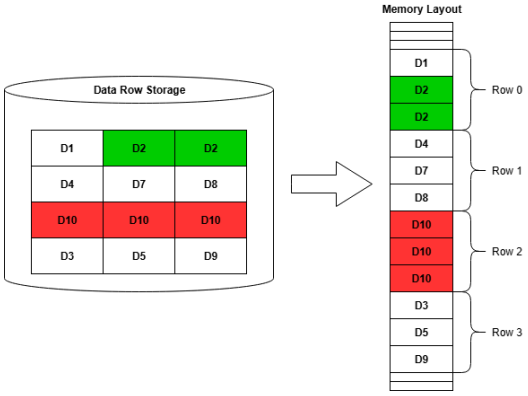


Figure 3.36: Visual Representation of C-Contiguous array storage in memory

Figure 3.36 shows a visual representation of how the array elements, in the memory layout, are stored in a row sequence, without gaps in between. This will allow an improvement in performance when accessing data in memory in the same location without being dispersed.

After identifying the best model from SVM and LR, we compare both performances and choose the one with the best results. In the code segment A.3, the final evaluation is applied on the remaining 20% test samples all of which are unseen during training. Then, we load the best trained model, saved in a pipeline object, along with the selected subset of features acquired during the feature selection phase. The model attempts to predict the risk categories, which are evaluated using balanced metrics such as accuracy, F1 score, precision and recall. Finally, we implement a confusion matrix to visualise how well the model performed and classified the various risk categories.

In the following chapter, we make a comprehensive analysis of the results, comparing model performance across different configurations and interpret the results and implications within the context of cardiovascular risk and type 2 diabetes prediction assessment.

Chapter 4

Results/Discussion

This section presents and discusses the results obtained from the experimental phase of this study. The analysis aims to evaluate how the implementation of feature selection techniques and machine learning models to address the early detection of CVD and T2D.

4.1 Cardiovascular Diseases Results

4.1.1 CVD Dataset not Related to User Data

Starting with the dataset where the generated data is not related to the user data, we observe that the feature selection process displayed in Table 4.1 for the CVD risk shows a strong degree of stability across all training configurations. In this observation, RFE with RF as its estimator, the algorithm repeatedly identified six to seven features as the most relevant to CVD risk.

Among the variables, steps, vo2_max_precise, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, and sleep_resting_heart_rate become apparent, which stand out as key metrics for predicting CVD. This may suggest that the model can identify stable and meaningful signals, especially on a diverse dataset with 2.000 records. This behaviour aligns with the moderate variability in the present values, which also shows strong physiological coherence between cardiovascular fitness, sleep and metabolic symptoms. Furthermore, despite the presence of correlated features, RF has effectively prioritised the most informative metrics without destabilising the

Iterations	Splits	Number of Features	Selected Features
25	**5**	**7**	**steps, vo2_max_precise, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr**
50	5	7	vo2_max_precise, steps, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
75	5	7	vo2_max_precise, steps, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
100	5	7	vo2_max_precise, steps, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
25	7	7	steps, vo2_max_precise, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
50	7	7	steps, vo2_max_precise, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
75	7	6	steps, vo2_max_precise, bmi, avg_SPO_2, sleep_time_sec, sleep_resting_hr
100	7	6	steps, vo2_max_precise, bmi, avg_SPO_2, sleep_time_sec, sleep_resting_hr
25	10	7	steps, vo2_max_precise, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
50	10	7	vo2_max_precise, steps, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
75	10	7	vo2_max_precise, steps, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr
100	10	7	vo2_max_precise, steps, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_hr

Table 4.1: Overview of Selected Features Across Different Iterations and Splits for the CVD Dataset Unrelated to the User Smartwatch Data

feature rankings. Since RF is robust to multicollinearity, the repeated selection of these variables also suggests that this dataset provides distinct and non-redundant information to predict CVD. However, such consistency could reflect a bias towards dominant or strong feature correlations, potentially overlooking other patterns. This is reinforced by the fact that the search space showed marginal variability in the selected features. Additionally, while the use of RandomizedSearchCV with a parameter grid of 25, 50, 75 and 100 iterations with TimeSeriesSplitCV splits of 5, 7 and 10 gives room for a moderate search space, there remains the possibility that it may have landed on a plateau, where it performs well on the data but may not generalise as strongly. Despite this caution, Table 4.2 further supports these observations among the different combinations of iterations and splits.

Iterations	Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Accuracy	Std. Dev. Accuracy	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Exec. Time (min)
25	**5**	**0.869077**	**0.005981**	**0.867722**	**0.007005**	**0.879472**	**0.008994**	**0.879966**	**0.006184**	**2.50**
50	5	0.866632	0.007738	0.864335	0.007285	0.876674	0.085853	0.879069	0.013905	6.42
75	5	0.863402	0.004440	0.865448	0.005978	0.870276	0.008265	0.87988	0.006299	9.08
100	5	0.865098	0.009905	0.867347	0.011549	0.870072	0.008385	0.872673	0.008827	12.10
25	7	0.865087	0.028681	0.866729	0.028961	0.870661	0.022213	0.865714	0.018866	3.53
50	7	0.863645	0.025233	0.866932	0.025881	0.87574	0.03444	0.866286	0.02907	8.30
75	7	0.863120	0.023192	0.865423	0.023792	0.871281	0.034497	0.872	0.030538	14.20
100	7	0.861164	0.027768	0.865356	0.028496	0.861824	0.024939	0.860571	0.021954	17.35
25	10	0.852984	0.029087	0.859619	0.030263	0.877498	0.030138	0.861878	0.031448	6.09
50	10	0.862809	0.031289	0.864596	0.033517	0.871164	0.03481	0.872376	0.034365	14.25
75	10	0.863075	0.034180	0.865113	0.036354	0.877418	0.03849	0.867956	0.038233	20.11
100	10	0.862858	0.035858	0.864757	0.037461	0.874132	0.030723	0.873481	0.031973	24.26

Table 4.2: Feature Selection Performance of Dataset Unrelated to User Smartwatch Data

The balance F1 score metric ranges between 86% and 87%, means that the model can maintain a good balance among the false positives and false negatives within the three risk classes. This balanced behaviour is particularly relevant in the early detection of CVD, since the use of balanced metrics in an unbalanced dataset prevents one class from being favoured at the detriment of another, ensuring that both low, medium and high risk patients are evaluated in a trusted way. In addition, the balanced accuracy, precision and recall follow this performance, reinforcing the model's strength. Balanced precision presents values between 86% and close to 88%

in several configurations, which means that the model can minimise the amount of false positives among risk classes. On the other hand, the balanced recall also shows solid results, between 86% and 88%, with a slight tendency to vary in scenarios with more splits. This shows that increasing the number of splits leads to a reduction of training samples per fold, which makes the evaluation more sensitive from each sample variability, which explains why it can present some oscillation in the ability to identify all records for all risk classes correctly. Another relevant aspect is that the analysis of the standard deviation results remain low, especially in simple configurations such as twenty five iterations with five splits. This suggests that the feature selection process shows stability with slight fluctuation across different cross validations folds. However, as the number of folds increases, there is a slight dispersion, since each split uses fewer samples, which makes the training/validation process more sensitive, while adding computational cost.

The following feature selection best estimator, in Listing 4.1, from 25 iterations and 5 splits, with an execution time of 2 minutes and 50 seconds, shows the parameter choice for the best performance.

```
Best estimator found: RFE(estimator=RandomForestClassifier(
    class_weight='balanced',
    criterion='entropy',
    max_depth=5,
    max_features=0.5,
    max_samples=0.5,
    min_samples_leaf=4,
    min_samples_split=6,
    n_estimators=200),
    n_features_to_select=7)
```

Listing 4.1: Feature Selection Best Estimator of CVD Dataset unrelated to the User Data

The process shows that the best estimator found by the process of RandomizedSearchCV, together with RFE, selected seven features as the most relevant. Table 4.3 shows the hyperparameter criteria list that contributed to the selection of features.

The number of selected features aligns with the predefined search space specified by n_features_to_select, ranging from five to nine features. This number represents a clear balance between the reduction of the feature dimension space and the preservation of the essential information for the prediction of CVD. Concerning the parameter n_estimators, the optimal value was 200, within the search space of 100, 200 and 300 estimators. This value ensured a higher robustness than 100 trees without incurring a higher computational cost of 300, which shows a balanced stability

Parameters	Values	Description
feature_selection__n_features_to_select	5, 6, 7, 8, 9	Number of features to select after ranking by the internal estimator. Smaller values reduce dimensionality and computational cost.
feature_selection__estimator__n_estimators	100, 200, 300	Number of trees in the RandomForest. More trees often improve stability and accuracy, but increase training time.
feature_selection__estimator__max_depth	3, 4, 5, 6	Maximum depth of each tree. Limits how deep the tree can go. Helps control overfitting. Avoiding None prevents overly complex trees and saves memory.
feature_selection__estimator__min_samples_split	6, 8, 10, 12	Minimum number of samples required to split an internal node. Higher values make the model more conservative.
feature_selection__estimator__min_samples_leaf	4, 5, 6, 8, 10	Minimum number of samples required at a leaf node. Higher values smooth the model and prevent overfitting. Important in imbalanced datasets.
feature_selection__estimator__max_features	sqrt, log2, 0.3, 0.5	Number of features to consider when looking for the best split. Sqrt is the default for classification. Log2 or float fractions may boost performance but affect speed.
feature_selection__estimator__bootstrap	True	Setting True means each tree is trained on a random subset of the data, increasing diversity and robustness, instead of using False to train the tree with the full dataset.
feature_selection__estimator__max_samples	0.1, 0.3, 0.5	Fraction of the training samples used for fitting each base estimator if bootstrap=True. Controls diversity among trees and reduces overfitting.
feature_selection__estimator__criterion	Gini, Entropy	Function to measure the quality of a split. Gini is faster and often preferred, while Entropy uses information gain and may yield more precise splits.
feature_selection__estimator__class_weight	None, balanced	Helps with imbalanced datasets by giving more importance to minority classes, otherwise use None.

Table 4.3: Hyperparameters List in RandomizedSearchCV for Feature Selection with Random Forest Based Recursive Feature Elimination

and efficiency. Regarding the max_depth parameter, the choice fell on five, within the range of 3, 4, 5 and 6 leaves. The model managed to capture patterns without creating a relatively complex structure that could lead to overfitting by expanding the nodes until all leaves are pure. This aspect is directly related to the observation of slight deviations in the results across different iterations and splits, given the fact that the controlled depth imposed limits to the extraction of complex patterns. As for the min_samples_split, the chosen value was six, between the range of 6, 8, 10 and 12, allowing a better flexibility in internal node divisions. Given the relatively balanced dataset, this is useful in increasing the exploration of relations between variables. To complement, the min_samples_leaf configuration choice was four, within the search space of 4, 5, 6, 8 and 10, which made the model more tolerant in accepting leaves with low samples. This can lead to small oscillations in the results, reported in Table 4.1, since leaf reduction can still help capture small patterns between iterations and folds. Within the tree configuration, the max_features parameter choice was 0.5, which means that half of the variables were considered in each division. Among the tested options, the model optimal choice offers better diversity among the trees, which explains the overall results consistency, despite introducing slight variations in the results. In the same way, the use of Bootstrap and max_samples value of 0.5 ensured that each tree trained with half of the available metrics in each unique subset of data. By exposing each tree to a distinct subset of data, ads variability among the trees. As a result, each tree learns different patterns and makes different errors, where the combination of these trees helps mitigate the risk of overfitting, while it strengthens the model's ability to generalise over unseen data. As for the criteria function, the choice was Entropy, in contrast to Gini. Entropy tends to produce more balanced and informative splits by evaluating the purity of nodes with greater sensitivity, which can be particularly advantageous in datasets with class imbalance, which is the case in this dataset. Furthermore, the class weight parameter optimal value was balanced. This configuration ensured classes with a

smaller representation were compensated in weight during the training/validation process, so that during the prediction of CVD, classes with smaller representation may not be neglected.

Having determined the best set of features, we apply two supervised machine learning techniques, SVM and LR, to assess the prevention of CVD risk with multiple classes, low, medium and high. In this phase, we apply GridSearchCV to identify the best possible combination of hyperparameters, in contrast to the RandomizedSearchCV applied in the feature selection stage, despite the present limitations of hardware capabilities. Nonetheless, the application of SVM in Table 4.4 shows a consistent, balanced F1 score of ≈83%, with balanced precision and recall, also maintaining high levels across different splits. In terms of balanced accuracy, the number of correct predictions is ≈82%, indicating that in the training phase the model managed to identify the majority of the class risks. Having a balanced F1 score close to accuracy further shows that the model is capable of handling prediction error across classes, which is a desirable behaviour for the early risk identification in multi class CVD risks. The configuration with 5 splits is selected as the optimal setup, despite a slightly higher mean of F1 score observed in 10 splits. The 5 split model achieved nearly identical balanced F1 score (82.9% vs. 83%), accuracy, precision and recall values, while maintaining a lower standard deviations across all metrics. This shows that the model is able to maintain a stable and consistent performance across the applied folds. In addition, as the number of folds increased, the execution time was approximately half of the 10 configuration splits, making it a more efficient model choice.

Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced Accuracy	Std. Dev. Accuracy	Exec. Time (min)
5	**0.8291**	**0.0149**	**0.8410**	**0.0103**	**0.8252**	**0.0165**	**0.8246**	**0.0158**	**14.22**
7	0.8273	0.0384	0.8336	0.0407	0.8257	0.0371	0.8229	0.0375	20.10
10	0.8303	0.0309	0.8386	0.0298	0.8276	0.0312	0.8245	0.0325	28.16

Table 4.4: SVC Training Results with GridSearchCV over Dataset unrelated to the User Smartwatch Data

During this process, with the SVC machine learning algorithm, the model reached its best configuration as seen in the code segment 4.2, supported by the selection of seven features from the feature selection phase.

```
Best parameters found: {
    'svm__C': 5,
    'svm__break_ties': True,
    'svm__class_weight': 'balanced',
    'svm__decision_function_shape': 'ovr',
    'svm__gamma': 'scale',
    'svm__kernel': 'rbf',
    'svm__shrinking': True,
```

```
    'svm__tol': 0.01
}
```

Listing 4.2: SVC Training Phase Best Parameters of CVD Dataset unrelated to the User Data

Amongst the list of hyperparameters from Table 4.5, the C parameter chose the value of five as its best result for an optimal margin. This choice shows that the model chose to penalise the classification errors quite strongly, by selecting a regularisation value that avoids extreme values of 10 and 100. Furthermore, the SVM classifier aimed to construct a decision boundary that correctly classifies most training points while allowing some flexibility to prevent overfitting. In this context, the model likely determined that smaller values resulted in a smooth boundary with a high bias, where the model is too simplistic to capture the true relationship patterns in the data, causing underfitting. On the other hand, larger values of the C parameter may lead the model to enforce a rigid separation of the values by classifying all training points correctly, where it tries to create a complex decision boundary with a narrower margin, leading to a higher risk of overfitting. The adoption of the rbf kernel further enhances this flexibility with the ability to capture nonlinear relationships between features and the CVD outcomes. With this choice, the model measures the similarity between data points in different dimensions and determines the classification of that data point by majority vote. To complement this, the gamma parameter choice of automatically scaling adjusts the influence of individual training subsets to ensure that none has an excessive impact, which is particularly important in health datasets where outliers may exist but should not dominate the decision boundary.

Parameter	**Values**	**Description**
svm__C	0.1, 0.2, 0.5, 1, 2, 5, 10, 100	Smaller values specify stronger regularization, while larger values aim to fit the training data more closely.
svm__kernel	linear, rbf	Specifies the kernel type. Linear creates a linear decision boundary, while rbf (radial basis function) creates a non-linear one.
svm__gamma	scale, auto	Kernel coefficient that defines the influence of training samples. 'scale' uses 1 / (n_features * X.var()), while 'auto' uses 1 / n_features.
svm__tol	1e-1, 1e-2, 1e-3, 1e-4	Stopping criterion tolerance. Smaller values mean more precise convergence but may increase training time.
svm__shrinking	True, False	Whether to use the shrinking heuristic, which can speed up training by ignoring less important support vectors.
svm__class_weight	None, balanced	Adjusts class weights to handle imbalanced datasets. 'balanced' uses the inverse class frequency.
svm__decision_function_shape	ovr	Determines how decision functions are calculated. 'ovr' (one-vs-rest) is the default.
svm__break_ties	True, False	If enabled, uses decision function values to break ties, when multiple classes have the same score. Only relevant with decision_function_shape='ovr'.

Table 4.5: SVC Hyperparameter List

Another critical parameter, in Table 4.5, is the balanced weight between the classes. This corrects potential class distribution differences by adjusting the penalty to the class frequency. Similarly, the decision function shape of One Versus Rest

(OVR) is well suited for multiclass problems, since it is capable of producing separate boundaries for each class against the others, by using a winner takes all strategy, in which the classifier with the highest output function assigns the class. This helps with the interpretation of results and distinguishes each risk from another. In addition, the model included break ties, which is relevant with the inclusion of OVR and number of classes > 2, will help solve draws in classification decision scores, especially in cases where patients present overlapping data that make the class boundaries less distinct. On the computational side, the presence of the shrinking heuristic optimises the model when the number of iterations are relatively large, as the shrinking can shorten the training time. However, if the model uses a large stopping tolerance, such as 1e-1 it allows the classifier to converge more quickly, since it stops once the improvement becomes very small. In this case, the choice for tolerance prioritises efficiency by reducing training time over precision compared to smaller tolerance values, like 1e-2, 1e-3 and 1e-4.

Following the SVC application, we evaluate the performance of LR classifier, using the same dataset and number of features, from the previous Table 4.1. As shown in Table 4.6, it presents lower performance results, with a balanced F1 score ranging between 78% and 79%, supported by the average precision of 79% and Recall of 78%. These results are noticeably lower than those obtained with SVC, which shows that this classifier, despite being a linear classifier, has a higher difficulty in capturing information that helps with the creation of class boundaries for this dataset. The slight higher precision compared to recall shows that the model favours correctness over sensitivity.

Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced Accuracy	Std. Dev. Accuracy	Exec Time (min)
5	**0.7896**	**0.0211**	**0.7979**	**0.0186**	**0.7862**	**0.0227**	**0.7810**	**0.0237**	**1.22**
7	0.7855	0.0166	0.7979	0.0202	0.7811	0.0164	0.7774	0.0164	1.51
10	0.7868	0.0173	0.7988	0.0204	0.7829	0.0169	0.7794	0.0162	2.02

Table 4.6: Logistic Regression Classifier Training with Grid Search Cross Validation of CVD Dataset unrelated to the User Data

In addition, the accuracy ranged closely to 78%, which confirms that the model has a higher difficulty in accurately separating the different risks for CVD. A possible conclusion to this performance may be that the dataset presents moderate variability and non-linear relationships among the selected variables that constrained the creation of linear decision boundaries. Another possible conclusion to this may be that the model used a simpler hyperparameter list, which evidently led to a less demanding resource model. An additional effect may be the use of only one penalty parameter, L2, which is the only available regularisation option supported by the chosen multi class solvers in this setup and might limit the flexibility of the model to adapt to the structure of the dataset. Despite this, the standard deviations across different folds remained low, which indicates the model is stable and has consistent

predictions on the validation process. Interestingly, the variability was slightly lower for the configurations with higher splits, compared to 5 splits. This improvement in stability may be attributed to a higher number of splits, where each training and validation subset performance becomes more reliable and less dependent on specific data partitions. This behaviour corresponds with what was previously observed in the feature selection process, where the dataset consistently showed stable results, except in configurations with high fold sizes. Furthermore, a major advantage of LR was its computational efficiency, with execution times under 3 minutes, even at the highest number of splits, making the use of this classifier a practical option in scenarios with limited resources, or when a fast retraining is needed.

The model optimal configuration choice is defined by the parameters list in Listing 4.3. According to the hyperparameter space presented in Table 4.7, the choice of the C parameter with 0.3 introduced a strong regularisation, which means that the model penalised strong coefficients and discouraged an overly complex decision boundary. This setting suggests the model prioritised generalisation and robustness over fitting every record into the training data, which aligns with the observed accuracy metric values, as an inverse regularisation can reduce the training accuracy while improves the stability across folds. However, this may also show that the model might have underfitted, because the model became simpler and might not have been able to capture real patterns accurately within the data. Furthermore, the use of balanced class weight compensated for the natural tendency of the model to favour the majority of the classes with higher representation by scaling the loss of contribution of underrepresented classes.

```
Best parameters found: {
    'lr__C': 0.3,
    'lr__class_weight': 'balanced',
    'lr__dual': False,
    'lr__fit_intercept': True,
    'lr__max_iter': 300,
    'lr__penalty': 'l2',
    'lr__solver': 'sag',
    'lr__tol': 0.1,
    'lr__warm_start': False
}
```

Listing 4.3: LR Training Phase Best Parameters of CVD Dataset unrelated to the User Data

Moreover, among the available options of solvers, only three algorithms stood out for multiclass problems, while being suitable for medium sized datasets. Among the available options, Stochastic Average Gradient (Sag) was chosen as the best

approach, since it efficiently manages medium/large datasets by iteratively updating the model parameters using the average of previously computed gradients.

Hyperparameter	Value	Description
lr__penalty	L2	Controls the type of regularization and prevents the model from overfitting
lr__tol	1e-1, 1e-2, 1e-3, 1e-4	Controls the stopping criterion. Smaller values = more precision but slower training.
lr__C	0.3, 0.5, 1, 2, 3	Inverse regularization strength. Low values apply stronger regularization, allowing some misclassifications to improve generalization. Ideal for noisy data. High values reduce regularization, forcing the model to fit the training data more closely, which may lead to overfitting.
lr__dual	False, True	Dual formulation is only implemented for l2 penalty with liblinear solver. Prefer dual=False when n_samples >n_features.
lr__fit_intercept	True, False	Specifies whether to include an intercept term in the model. If set to True, the model learns a bias term to shift the decision boundary. If False, the model assumes the data is already centered and forces the decision boundary to pass through the origin.
lr__class_weight	None, balanced	Adjusts the weight assigned to each class during training to address class imbalance.
lr__solver	lbfgs, newton-cg, sag	Algorithm to use in the optimization problem. Among the several solvers, only liblinear does not apply to multiclass problems.
lr__warm_start	True, False	If set to True, the model retrains its previous solution when .fit() is called, allowing iterative training or reuse of learned coefficients
lr__max_iter	100, 200, 300, 400	Specifies the maximum number of iterations the solver will run during model optimization. Increasing this value can help with convergence, especially for complex models or poorly scaled data

Table 4.7: Logistic Regression Hyperparamater List

This approach allows a fast convergence compared to traditional gradient descent methods. However, for a faster convergence speed, it is only guaranteed on features with approximately the same scale, a condition which is satisfied by the application of RobustScaler in the preprocessing phase. Furthermore, the only choice available for these algorithms was the L2 penalty. The primary purpose of this regularisation is to prevent overfitting by discouraging the model from assigning excessively large values to the coefficients in a way to shrink its values towards zero, but not forcing them to be exactly zero. This, in turn, will reduce the complexity of the model by finding a balance between the learning and regularisation rate, while improves the generalisation towards unseen data. Among the set range of tolerance values, the tolerance of 1e-1 is relatively high, compared to the default value of 1e-4, which means the optimisation stopped earlier once the improvements became marginal. This explains why the model converged faster while still delivering reasonable scores on a trade-off between F1 score and execution time. Regarding the maximum iterations of 300, it gave the solver enough iterations to guarantee convergence under a comfortable tolerance by avoiding an even earlier termination. The decision to keep fit_intercept, True, allowed the model to estimate a baseline logarithmic odd, which is particularly important, since not all patients start from the same baseline risk. Finally, the model by setting the dual parameter False, is the optimal approach because the dataset had more samples than features, which makes the computation cost cheaper, while a warm start parameter set to False allowed an unbiased optimisation for each run. Together, this set of hyperparameters shaped a model that prioritised a balanced dataset performance, with stability across correlated variables and computation efficiency.

When comparing both models, the SVC classifier achieved a higher performance across F1 score, precision, recall and accuracy during training. This shows that it may generalise better for CVD risk prediction over unseen data, because of the

dataset exhibiting non-linear patterns and feature interactions that are more effectively modelled by the SVC decision boundaries. Logistic Regression, while less performant, attempted to identify linear patterns among the features, which resulted in lower values of F1 score. Nevertheless, it demonstrated clear advantages in execution time and computational efficiency, making it a practical choice in scenarios where there are resource constraints or faster iteration cycles are needed. However, since these results were obtained in the training phase through cross validation, they should be interpreted as indicative trends rather than definitive conclusions.

4.1.2 CVD Dataset Related to User Data

Moving to the dataset, whose generated data is related to the user smartwatch data, the feature selection process followed the same methodology as with the previous dataset. The feature selection shows a strong consistency between different combinations of iterations and cross validations folds, as seen in Table 4.8. The number of features ranged between seven and eight selected metrics, with most scenarios consistently identifying seven features, vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec and sleep_resting_heart_rate. This combination appears repeatedly, which shows these variables hold substantial predictive value for the prediction of CVD in this dataset.

Iter	Splits	Number of Features	Selected Features
25	5	7	vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
50	5	8	vo2_max_precise, steps, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
75	5	8	vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
100	5	8	vo2_max_precise, steps, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
25	7	7	vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
50	7	8	vo2_max_precise, steps, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
75	7	7	vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
100	7	7	vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
25	**10**	**7**	**vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate**
50	10	7	vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
75	10	7	vo2_max_precise, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate
100	10	8	vo2_max_precise, steps, resting_heart_rate, bmi, avg_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_resting_heart_rate

Table 4.8: Overview of Selected Features Across Different Iterations and Splits for the CVD Dataset related to the User Smartwatch Data

The variable `steps` appears occasionally, but despite being relevant, it does not contribute enough information to surpass the other variables in importance. This behaviour mirrors the same instability observed in the previous dataset feature selection process, as it may have reached a plateau where different combinations and higher computational cost are not providing new information to change the feature ranking. Despite these occasional variabilities, Table 4.9 shows a solid performance among all configurations, with the optimal configuration of 25 iterations and 10 splits. The balanced F1 score of 87.9% and the highest balanced accuracy of 88.2% indicate that the model is capable of identifying a high number of risk classes with a balance between precision of 89% and recall of 88%.

This means that the model is capable of identifying most cases within each class while keeping the number of false positives and false negatives low. Notably, the

Iterations	Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Accuracy	Std. Dev. Accuracy	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Exec Time (min)
25	5	0.872582	0.022828	0.873496	0.024651	0.875151	0.020533	0.873273	0.024278	3.12
50	5	0.872055	0.026460	0.873077	0.028064	0.87555	0.022052	0.873874	0.023569	7.1
75	5	0.872504	0.026072	0.873475	0.027883	0.887183	0.022039	0.875075	0.023462	10.08
100	5	0.870270	0.023502	0.871517	0.024912	0.883769	0.02276	0.872673	0.025594	13.55
25	7	0.868841	0.031151	0.870473	0.031838	0.884816	0.023664	0.872571	0.02989	4.14
50	7	0.873409	0.038344	0.875707	0.038379	0.887144	0.026156	0.873714	0.032556	9.23
75	7	0.876393	0.030488	0.878581	0.031244	0.888426	0.027608	0.877143	0.032335	14.42
100	7	0.873511	0.031586	0.875218	0.031487	0.887982	0.025666	0.874857	0.033719	19.28
25	**10**	**0.879088**	**0.023979**	**0.882352**	**0.027643**	**0.889664**	**0.026985**	**0.88011**	**0.028713**	**5.58**
50	10	0.878698	0.027240	0.881272	0.029730	0.885678	0.031341	0.875138	0.031853	13.08
75	10	0.877638	0.025826	0.880163	0.028850	0.890501	0.024289	0.881215	0.025348	19.42
100	10	0.875321	0.026924	0.877842	0.029877	0.885349	0.025469	0.875691	0.027094	24.46

Table 4.9: Feature Selection Results of CVD Dataset with Data related to the User Smartwatch Data

narrow standard deviations show a relatively low variability across different temporal validation splits. This shows that the model managed to capture a broad portion of records without excessive bias toward specific classes, where the application of the Random Forest estimator seems to generalise well under different validation folds.

The process of RandomizedSearchCV applied to RFE with RF converged to an optimal estimator as seen in the Listing 4.4. With this set of configurations from the optimal estimator, using the same hyperparameter list from Table 3.5, the parameter n_estimators chose the value of 200, selected from the range of 100, 200 and 300.

```
Best estimator found:RFE(estimator=RandomForestClassifier(
    class_weight='balanced',
    criterion='entropy',
    max_depth=5,
    max_features=0.5,
    max_samples=0.5,
    min_samples_leaf=4,
    min_samples_split=6,
    n_estimators=200),
    n_features_to_select=7)
```

Listing 4.4: Feature Selection Best Estimator of CVD Dataset related to User Data

This choice led to a balanced computational cost and stability, without the need to add additional computational cost. The max_depth parameter selected the value of five, among the array of 3, 4, 5 and 6. The estimator believes that this choice leads to the capture of significant relationships between variables, without excessively going deeper into the tree itself, to the point of making every single node leaf pure. In terms of splits, the model used six sample splits and four sample leaves, both in the lower range of the grid parameter space. This allowed the trees to remain flexible in detecting meaningful patterns in the data without fragmenting into a finer granularity to distinguish the different levels of CVD risk. Regarding

max_features the model chose 0.5 from the array space, which means that it used half of the available variables at each split, promoting diversity among trees, which explains the consistency of the results over different sets of iterations and folds. Similarly, max_samples with a combination of the Bootstrap parameter allowed each tree to train with only half of the data on unique subsets of data, leading to an improved standard deviation. In this process, the model also chose Entropy over Gini criteria, which resulted in purer partitions, based on information gained. This choice enabled the need to capture subtle patterns in this dataset. Finally, the inclusion of weight balance further contributed to correctly identify unbalanced class risks, which contributed to a more reliable balanced F1 score and accuracy metric performance.

Following the feature selection stage for the dataset whose data is related to the user's smartwatch data, we applied the same supervised machine learning algorithms for the training process. Following the application of SVC, the results in Table 4.10 show consistent performance with a balanced F1 score between 81% and 82%, with the highest performance of 7 splits. The classifier was able to maintain an even trade-off between false positives and false negatives across the three CVD risk categories.

Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced Accuracy	Std. Dev. Accuracy	Exec Time (min)
5	0.8165	0.0198	0.8241	0.0212	0.8132	0.0200	0.8109	0.0214	15.24
7	**0.8223**	**0.0254**	**0.8256**	**0.0250**	**0.8211**	**0.0253**	**0.8155**	**0.0269**	**21.06**
10	0.8188	0.0298	0.8258	0.0319	0.8166	0.0291	0.8149	0.0307	32.75

Table 4.10: SVC Training Results with GridSearchCV for CVD Dataset with Data related to User Smartwatch Data

The balanced precision values ranged around 82% and 83%, demonstrates the ability to avoid a high number of false positive instances across all risks classes. On the other hand, the balanced Recall varied between 81% and 82%, shows the SVC model is reasonably good at identifying true positives across different splits. These result differences shows the decision model was more endorsed to determine the correctness of the predictions over the full capture of the positive cases.

The configuration with 5 splits presented a slightly lower standard deviation values compared to those with 7 or 10 splits. In this case, the higher variance may result from the smaller validation subsets used at higher split counts. Each fold contains fewer splits, which, in a multiclass setting with a relative unbalanced dataset, may have led to a less representative validation of subsets of data and a slight fluctuation of the metric performances. Finally, the execution time followed the expected behaviour, where a higher number of splits lead to a higher execution time, due to a higher number of training and validation cycles provided by GridSearchCV. However, while the additional splits can improve the reliability of the model, the computational cost rose significantly, making the 7 splits configuration a balanced option between stability and runtime.

In this process the optimal SVC estimator for this dataset in the Listing 4.5 supported the selection seven features, using the parameter settings in Table 4.5.

```
Best parameters found: {
    'svm__C': 100,
    'svm__break_ties': True,
    'svm__class_weight': None,
    'svm__decision_function_shape': 'ovr',
    'svm__gamma': 'scale',
    'svm__kernel': 'rbf',
    'svm__shrinking': True,
    'svm__tol': 0.01
}
```

Listing 4.5: SVC Best parameters estimator for CVD Dataset related to User Data

With this set of configurations, the model chose the value of 100 for its regularisation. This choice indicates that the model performed best under a weak regularisation, where it tried to correctly identify all training samples, rather than maintaining a broader separation between classes, even if it meant to allow for some misclassification of the training vectors. In addition, the model was capable of benefiting from a more complex decision boundary, which may indicate the case for overfitting if the dataset was small or had noisy data. The kernel, chosen as rbf, provides the flexibility to model non-linear relationships between features and the CVD outcomes. This choice is crucial because the relationships of CVD risk are often complex, multivariate and non-linear. This kernel contributed to a balanced accuracy values above 81%, confirming that the model generalised reasonably well between different classes. The gamma parameter, set to scale, adapts the influence of each support vector based on feature variance. This prevents excessively sharp decision boundaries and ensures that no single record dominates the classification. As a result, the model achieved stable recall values of 82% with seven splits, showing that it could consistently detect positive cases in multiple folds without missing too many risk instances. The class_weight, set to None, indicates that no adjustments were made for potential class imbalances. Interestingly, the results still show precision above 82% and recall above 81%, suggesting that the feature set already contained strong selective signals, despite the fact that the dataset is relatively unbalanced. The break_ties parameter, set to True, allowed the model to use the decision function values to resolve the ties between classes. Although subtle, this setting helps achieve slightly more consistent predictions across splits, supporting the low variance seen in metrics such as F1 score with a standard deviation of 0.0254. The shrinking heuristic, enabled to True, sped up training by ignoring less relevant support vectors

during optimisation. This allowed the model to converge more efficiently despite the high regularisation value, keeping execution times somewhat manageable, around 21 minutes with seven splits. Finally, the tolerance was set to 1e-2, which is relatively large compared to stricter options such as 1e-3 or 1e-4. This tolerance criterion reduced training time while still ensuring stable performance, as reflected by the consistent F1 score, precision, recall and accuracy scores across folds.

Regarding the LR classifier, the performance metrics were relatively lower, as shown in Table 4.11. The balanced F1 score around 75%, across all configurations, indicates the limited ability to capture more complex and less linear patterns in the dataset. This represents a noticeable decrease in performance compared to the SVC classifier.

Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced Accuracy	Std. Dev. Accuracy	Exec Time (min)
5	0.7498	0.020632	0.7701	0.025589	0.7423	0.020491	0.7406	0.0219	1.32
7	**0.7506**	**0.026166**	**0.7710**	**0.029059**	**0.7429**	**0.027357**	**0.7406**	**0.027615**	**1.43**
10	0.7495	0.023956	0.7642	0.028075	0.7431	0.02424	0.7417	0.025461	1.91

Table 4.11: Logistic Regression Classifier Training with Grid Search Cross Validation for CVD Dataset with Data related to the User Smartwatch Dataset

The balanced precision of ≈77% exceeded the recall of ≈74% by a small margin, which shows that the model was more conservative in making predictions. This behaviour implies a lower rate of false positives, but a lower tendency to miss a small number of true cases. Although this tendency reduced false positives, it also implies that the classifier misses a proportion of actual risk cases, which is an important limitation in risk assessment, where failing to identify true risk cases may lead to an underestimation of an individual's health. Furthermore, the balanced accuracy with a value ≈74% further highlights this limitation, which shows that the model struggled to maintain a consistent performance across all risk classes. The standard deviation remained low across all splits, where, although the model underperformed, in comparison with SVC, its behaviour remained stable during cross validation folds and was not too sensitive to variations in data partitions. However, despite these low values, LR demonstrated advantages in computational efficiency, requiring only 1 minute and 43 seconds with 7 splits, making it significantly faster than SVC. This efficiency suggests that LR could serve as a lightweight reference point or a quick evaluation model, but its reduced ability to capture complex non-linear relationships limits its practical use, particularly when dealing with derived non-invasive smartwatch physiological signals, which may encode more subtle or non-linear relationships.

From these observations, the best estimator for LR is as follows in Listing 4.6 with seven features.

```
Best parameters found: {
    'lr__C': 0.5,
    'lr__class_weight': None,
    'lr__dual': False,
    'lr__fit_intercept': True,
    'lr__max_iter': 200,
    'lr__penalty': 'l2',
    'lr__solver': 'sag',
    'lr__tol': 0.1,
    'lr__warm_start': True
   }
```

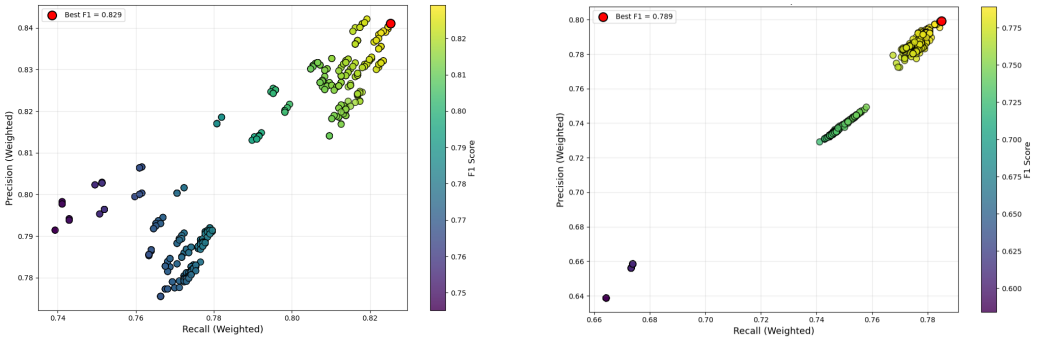
Listing 4.6: LR Best Parameters Estimator for Dataset with Data related to the User Data

Alongside the chosen parameters, Table 4.7 shows the available list of parameters space. Regarding regularisation, the model chose the C parameter of 0.5, which represents a relatively moderate regularisation compared to higher values. This choice allowed the model to control the magnitude of the coefficient to prevent overfitting, while being flexible in fitting the training data. The penalty was set to L2, which is consistent with the need to shrink coefficients, without eliminating them entirely, thus maintaining stability in the solution while handling correlated variables. The selected solver was Sag, an efficient method for large datasets and suitable for handling the L2 penalty, which also provides faster convergence compared to other solvers. The maximum number of iterations set to 200 indicates that the model did not require many iterations to reach an optimal cost of performance, compared to 100 iterations. Higher limits did not yield further performance gains, which suggests that the convergence was already attained with 200 iterations. Therefore, the chosen value shows a balance between computational cost and a fast convergence for the multiclass CVD risk. The tolerance of 1e-2, provided a balanced converged precision with training time, allowing the algorithm to stop earlier once an acceptable solution was reached. Additionally, the class_weight parameter was set to balanced, which ensures that minority classes coefficients received proportionally more importance, which helped improve recall and balanced accuracy in the presence of class imbalance. Finally, the presence of fit_intercept enabled the model to adjust for any systematic offset in the data, while the absence of warm_start ensured that the model started training from scratch in each run, avoiding bias from previous fits.

4.1.3 Comparison of CVD Datasets Training Models

In summary, the dataset whose data is not related to the user data, in Figure 4.1, showed stronger classification outcomes due to the clear class structured separability

and nature of the generated variables, which provide apparent patterns for the model to learn, but may limit its application to generalise to less controlled real world data.

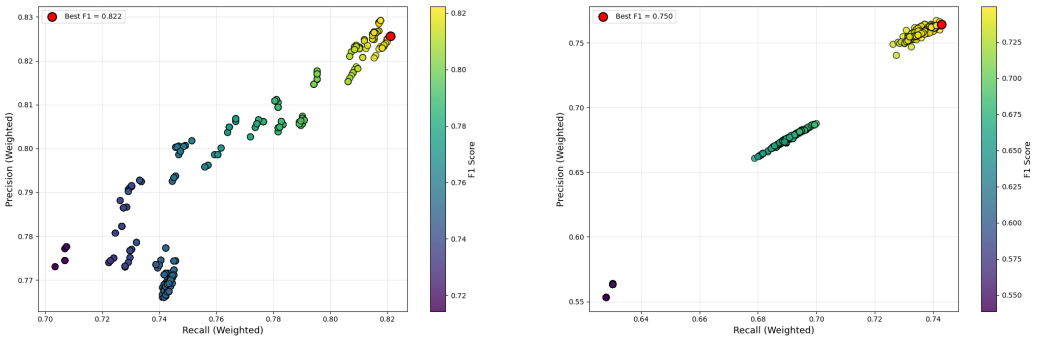


(a) SVC F1 Scores Distribution

(b) LR F1 Scores Distribution

Figure 4.1: Comparison of F1 Scores of SVC and LR Classifiers of Best Performance Results on the CVD Dataset whose Data is not related to the User Smartwatch Data

In contrast, the dataset whose generated data is related to the user smartwatch data, in Figure 4.2, showed more complex relationships and possibly non-linear feature patterns, consistent with real world physiological and behavioural data captured through smartwatches. The models had a slightly higher difficulty in capturing insights, which lead to a slightly lower performance.



(a) SVC F1 Scores Distribution

(b) LR F1 Scores Distribution

Figure 4.2: Comparison of F1 Scores of SVC and LR Classifiers of Best Performance Results on the CVD Dataset whose Data is related to the User Smartwatch Data

Nevertheless, the feature selection process showed a strong consistency with similar features being repeatedly selected across different iterations and folds. In the training phases, SVC performed best in both cases, where LR remained efficient but less effective in separating risk classes. However, with the assistance of feature selection, the models identified key predicted variables, highlighting their utility in the training models to acquire good performance values in the CVD classification outcomes.

4.2 Diabetes Type 2 Diseases Datasets Results

4.2.1 T2D Dataset not Related to the User Data

Having discussed the model performance on the CVD datasets, we shift the focus towards the T2D datasets, using the same methodology applied in the previous CVD datasets. Starting with the dataset where the data is derived from the CVD dataset, whose data is not related to the user smartwatch dataset, alongside additional T2D metrics, we observe the feature selection process displayed in Table 4.12. This table summarises the selected features across different iterations and cross validations splits settings for the T2D dataset.

Iter	Splits	Number of Features	Features Selected
25	5	7	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, total_calories, HRV_day_ms
50	5	8	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, total_calories, HRV_day_ms, HRV_night_ms
75	5	7	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, total_calories, HRV_day_ms
100	5	6	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, total_calories, HRV_day_ms
25	7	7	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
50	7	7	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
75	7	6	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms
100	7	6	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms
25	10	7	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
50	10	6	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms
75	**10**	**6**	**bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms**
100	10	6	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms

Table 4.12: Overview of Selected Features across Different Iterations and Splits for the T2D dataset unrelated to the User Data

The results show that there is a consistent subset of predictive variables in almost all configurations. In particular, bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, and HRV_day_ms, which are repeated as core predictors of T2D. These variables emerged as highly influential with strong or moderate correlations with the target class, from which the RF importance rankings consistently determined these values as the top physiological contributors in this dataset, towards T2D. Secondary variables such as avg_SPO_2, total_calories and HRV_night_ms provide additional context of relevance, although less dominant, towards the T2D assessment. The model selected these variables in configurations where it favours broader tree structures of deeper feature utilisation, where their contribution becomes relevant. Across all configurations, the number of selected features varies between 6 and 8 variables, which shows that the RF classifier managed to identify meaningful metrics by assigning a higher importance to those metrics. This observation shows that the model converged on a subset of variables that reflect a strong predictive influence on T2D predictive analysis. Additionally, in contrast to the CVD datasets, it appears that the model has not reached a plateau where the feature selection process stabilises. With the use of RandomizedSearchCV and the assistance of TimeSeriesSplitCV, the model has explored a region of the search space that has not reached full convergence, which may require additional iterations to reach an optimal focal point. To this end, the best optimal configuration lies with

75 iterations and 10 splits, supported by the results in Table 4.13. The model converged towards a stable and relevant physiological feature set, with the integration of metabolic, cardiovascular and sleep related variables, which are strongly associated with T2D risk.

Iter	Splits	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced F1 score	Std. Dev. F1 score	Balanced Accuracy	Std. Dev. Accuracy	Exec. Time (min)
25	5	0.87553	0.01094	0.86006	0.00819	0.86328	0.00857	0.86105	0.00865	3.06
50	5	0.87296	0.01859	0.86306	0.01640	0.86530	0.01686	0.86262	0.01752	7.14
75	5	0.87703	0.01134	0.86126	0.00837	0.86465	0.00852	0.86238	0.00866	10.32
100	5	0.87553	0.01710	0.86186	0.01531	0.86484	0.01557	0.86221	0.01617	13.58
25	7	0.87946	0.03036	0.86343	0.02791	0.86664	0.02773	0.86363	0.02955	3.32
50	7	0.87979	0.03150	0.86400	0.03024	0.86723	0.02978	0.86442	0.03128	8.42
75	7	0.87752	0.03080	0.86514	0.02810	0.86782	0.02805	0.86357	0.02932	11.36
100	7	0.88080	0.03004	0.86857	0.02674	0.87099	0.02682	0.86699	0.02783	17.56
25	10	0.87717	0.03451	0.86133	0.03567	0.86433	0.03474	0.86107	0.03631	5.40
50	10	0.87693	0.03906	0.86519	0.03470	0.86712	0.03548	0.86372	0.03714	12.21
75	**10**	**0.88180**	**0.03527**	**0.86961**	**0.03616**	**0.87190**	**0.03564**	**0.86901**	**0.03702**	**18.36**
100	10	0.87774	0.03934	0.86575	0.03780	0.86791	0.03767	0.86505	0.03930	22.59

Table 4.13: T2D Feature Selection Performance of Dataset derived from CVD dataset with Data not related to User Smartwatch Data

The results show that the model performance gradually improved as the number of iterations and cross validation folds increased, with the ability to refine the importance of feature coefficients towards features with patterns, relevant to T2D. With the best configuration of 75 iterations and 10 splits, the model presents the highest balanced F1 score of 87.2%, supported by a mean precision of 89.2% and recall of 87%. This indicates that the model avoids overestimating higher risk labels unless supported by coherent evidence across the selected physiological variables. The balanced accuracy of 86.0% further confirms this performance across all target T2D risk labels, shows the model ability to distinguish the different T2D classes. Furthermore, the standard deviations remain moderate, around ≈0.035 between iterations, which shows that the model is reliable with minor fluctuations in performance occurring in different temporal folds. This stability is partially attributed to the strong and consistent core of features, which dominated the importance rankings and were selected repeatedly during the search RandomizedSearchCV process. Finally, the execution time for this optimal configuration was ≈18 minutes, which led the model to make a deeper exploration of the search space list. Nevertheless, the performance gains justify the additional computational cost with highlights for a stable compromise between execution time, precision, recall and accuracy over the detection of T2D risk. To validate the results, the Listing 4.7 presents the feature selection best estimator.

```
Best estimator found: RFE(estimator=RandomForestClassifier(
    class_weight='balanced',
    criterion='gini',
    max_depth=6,
    max_features=0.5,
```

```
6       max_samples=0.5,
7       min_samples_leaf=5,
8       min_samples_split=10,
9       n_estimators=300),
10      n_features_to_select=6
11  )
```

Listing 4.7: Features Selection Best Estimator of T2D Dataset derived from CVD Dataset not related to User Data

The process shows that RandomizedSearchCV, together with RFE, selected six features as the most relevant. The following Table 4.3 shows the hyperparameter search space list that contributed to the selection of the most predominant features. Concerning the number of trees, the model chose 300 estimators from the range of 100, 200 and 300. The increase of estimators enhances the stability and robustness of the feature importance ranking, which lead to a higher computational time, as seen by the optimal execution time of ≈18 minutes from the best configuration and high accuracy of 86%. By limiting the depth of the trees, the model chose the value of 6 for the max_depth parameter. This value provided sufficient depth to capture important patterns among physiological variables, without becoming complex or prone to overfitting. By controlling the tree complexity to this depth, the RFE model avoids an overly excessive decision boundary, since deep trees could inflate the importance of less generalised features. The parameters for min_samples_split and min_samples_leaf are tuned to the ranges of 6 to 12 and 4 to 10, respectively, to constrain the tree growth and smooth the decision boundaries. Regarding min_samples_split, the model determined that each node needs to have at least 10 samples before it can be split, which introduces another layer of regularisation. This helps in reducing overfitting and contributes to the interpretability of feature contributions, where meaningful statistical variables are flagged before creating new splits. As for min_samples_leaf, in the context of RFE, the model says it needs at least five leaves per node to make sure each decision is supported by a sufficient number of observations. These options are particularly beneficial in imbalanced datasets because they prevent the model from creating branches or leaves dominated by a single class, thereby promoting a more balanced feature selection training across all target risk categories. Among the options of max_features, the model chose 0.5, where each tree considers at least 50% of the available features to find the best split. This promotes the model to reduce the correlation among features and encourages diversity among trees to improve the feature coefficient importance scores. In addition, it prevents dominant variables from clouding informative features. With the bootstrap parameter set to True, it ensures that each tree is trained on a random subset of samples to promote diversity and reduce variance. The complementary max_samples set to 0.5 further enhances generalisation by reducing the dependency of specific

data subsets. In this dataset, the model selected Gini over Entropy criteria. Even though Gini impurity measurement is faster and less intensive, it still provides sufficient information to evaluate the quality of each split. Finally, the use of balanced weight mitigates the class imbalance effects by increasing the weight of minority class coefficient vectors during training. These metrics configuration shows that this model supports a reliable and generalised feature selection process across multiple folds with low variance. Using the selected features obtained from the feature selection process, we applied the SVC classifier with the GridSearchCV to assess the detection of T2D risks. The following Table 4.14 shows a strong performance across all validation settings, with gradual improvements as the number of folds increases.

Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced Accuracy	Std. Dev. Accuracy	Execution Time (min)
5	0.83318	0.00635	0.84033	0.00761	0.83063	0.00674	0.82840	0.00686	14.14
7	0.83571	0.02033	0.84160	0.02160	0.83371	0.02027	0.82974	0.02120	19.76
10	**0.84016**	**0.01595**	**0.84509**	**0.01609**	**0.83923**	**0.01647**	**0.83422**	**0.01631**	**28.36**

Table 4.14: SVC Training Results with GridSearchCV of T2D Dataset derived from the CVD Dataset with Data not related to User Smartwatch Data

The observed performance results, with the best configuration of 10 folds, show the highest average F1 score of 84% with a precision of 85% and a balanced recall of 83.9%. The model maintained a strong equilibrium between the correct positive predictions and the ability to identify true positive cases across all T2D label risks. The slightly higher precision shows that the model preferred to be cautious in prioritising confidence over sensitivity. Notably, the variability of the performance metrics decreases as the number of splits increases, with the standard deviation for the F1 score dropping from 0.0203 at 7 splits to 0.0159 at 10 folds. This decrease demonstrates that the model validation predictions became more stable and reliable, when trained with a higher number of splits during the data distribution of the cross validation process. The balanced accuracy of 83.4% supports the model performance across all risk levels, despite the slight class imbalance present in the dataset. However, this stability improvement showed a trade-off with execution time. The execution process increased from 14 minutes with 5 folds to 28 minutes with 10 splits, which demonstrates an additional demand for more thorough validations.

In this process, the optimal SVC estimator parameters for this dataset in the Listing 4.8 supported 6 features.

```
Best parameters found: {
    'svm__C': 100,
    'svm__break_ties': True,
    'svm__class_weight': None,
    'svm__decision_function_shape': 'ovr',
    'svm__gamma': 'auto',
```

```
    'svm__kernel': 'rbf',
    'svm__shrinking': True,
    'svm__tol': 0.01
}
```

Listing 4.8: SVC Best Estimator for T2D Dataset derived from CVD Dataset not related to User Smartwatch Data

With this set of configurations, Table 4.5 shows a regularisation strength of 100, where the model benefited from a more flexible decision boundary to better fit the training data, as accurately as possible. This weaker regularisation improved the model's ability to distinguish subtle variations among data samples, although at the cost of penalizing misclassification quite heavily. The model choice for rbf kernel confirms the action of capturing nonlinear complex interactions within the core feature set of cardiovascular, metabolic and sleep related variables linked to T2D outcomes. It allowed the classifier to create a smooth decision boundary to accommodate the most informative variations that may differ from each risk case. The gamma parameter is set to auto scaled, while combined with the high regularisation value and rbf kernel choice, the model was able to automatically adjust the curvature of its decision boundaries according to the spread of the data without manual intervention. This prevented the model from becoming overly sensitive to noise, which is common when using a large regularisation value, where it encourages the model to classify all training points correctly. The model tries to accurately separate the low, medium and high T2D risk categories, while maintaining a stable and consistent performance when evaluated using cross validation. The inclusion of break_ties to True ensured that each decision where the model has difficulty in determining which data point corresponds to which class, with the same score, it assists in breaking that tie. When using OVR, in a multi class setting, break_ties guarantees that these situations are handled appropriately so that each sample is assigned to the correct class. Similarly, the use of shrinking helped the optimisation process by ignoring less important support vectors. A tolerance value of 1e-2 provided a modest stopping criteria, which requires more iterations to find a near optimal accuracy, which eventually lead to a relatively high execution time of 14 minutes. Finally, the model determined that there was no need to assign a class weight, to appoint categories with less presence a compensatory weighting. However, since the dataset is relatively imbalanced, the aid of other parameters were sufficient to correctly identify minority classes, while maintaining a balanced performance across all risk T2D categories.

Regarding the LR classifier, the following Table 4.15 presents the training results, where the best performance was achieved with 5 splits. The model reached a mean balanced F1 score of 78.5%, supported by a precision of 79.3% and a mean recall of 78.1%. These values show that the model maintained a reasonable balance between identifying correctly positive cases and avoiding false positives across all T2D risk

categories. The slightly higher precision to recall shows the model was more cautious with assigning positive predictions, where it prioritised correctness over maximising sensitivity. This behaviour aligns with earlier observations of the feature distribution and correlation structure. The selected variables related to metabolic, cardiovascular and sleep exhibited non-linear boundaries, which were more effectively captured by the SVC model. In contrast, the linear LR algorithm is less capable of separating the T2D risk labels where feature information overlapped. This partially explains the lower scores, when compared to SVC, where the power of certain features appears to depend on non-linear iterations.

Splits	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced F1 score	Std. Dev. F1 score	Balanced Accuracy	Std. Dev. Accuracy	Exec Time (min)
5	**0.79289**	**0.02304**	**0.78138**	**0.02523**	**0.78652**	**0.02436**	**0.77499**	**0.02670**	**1.02**
7	0.78911	0.02632	0.78114	0.02250	0.78259	0.02346	0.77455	0.02225	1.43
10	0.79345	0.02306	0.78066	0.02471	0.78401	0.02341	0.77477	0.02498	2.04

Table 4.15: Logistic Regression Training Results with GridSearchCV of T2D Dataset derived from the CVD Dataset with Data not related to User Smartwatch Data

Furthermore, the balanced accuracy of 77.5% confirms that the model had a higher difficulty in distinguishing the three risk categories, given the imbalances in the dataset. Across all standard deviations, the values are relatively low, which demonstrates that the model was stable and able to be consistent across different cross validation folds, without major fluctuations in specific training/validation subsets of data. This stability also corresponds to the feature distributions observed in the feature selection exploratory analysis. In addition, increasing the number of splits from 7 to 10 did not lead to any substantial improvements in performance. In fact, with 10 splits, while it showed a higher precision of 79.4% and a relatively equally balanced F1 score of 78.6%, the recall actually decreased and the execution time doubled in time. This demonstrates that using 5 splits with the GridSearchCV technique provided the most efficient model in terms of performance, stability and computational cost.

With this performance, the optimal LR estimator parameters for this dataset in the Listing 4.9, supported 6 features.

```
Best parameters found: {
    'lr__C': 1,
    'lr__class_weight': 'balanced',
    'lr__dual': False,
    'lr__fit_intercept': True,
    'lr__max_iter': 200,
    'lr__penalty': 'l2',
    'lr__solver': 'sag',
    'lr__tol': 0.1,
    'lr__warm_start': False
```

```
}
```

Listing 4.9: Logistic Regression Best Estimator for T2D Dataset derived from CVD Dataset not related to User Data

To further validate the options selected by the estimator, Table 4.7 shows the parameter list space. Regarding the regularisation strength, the model chose the value of 1. This allows the model to balance the misclassification of some points, to improve generalisation, but at the same time, it forces the model to fit most of the training data accurately. In practise, it maintains a reasonable margin between classes, where it avoids underfitting and overfitting, leading to a somewhat generalised decision boundary across T2D risk categories. In combination with the L2 penalty, it helps with stabilising the model across the various subsets of cross validation folds, while at the same time, it prevents the model from overfitting by penalising large coefficients of extreme values. The choice of a balanced class weight helps minority classes by giving them a higher importance so that they are not ignored during training. This helps the relatively imbalance dataset maintain a fair training/validation performance across all risk categories. Setting dual formulation as False shows that the number of samples surpasses the number of features, which simplifies the computation and memory usage, as observed by the fast execution times. Meanwhile, adding an intercept term allows the model to capture trends and mean shifts in the data, rather than misclassifying samples simply because the features have different scales and distributions. If it was set to False the model would be forced to make the decision boundary pass through the origin, where all features values are zero, which could lead to a poor performance if the data was not centered or normalised. In terms of the solver, the model chose Sag. This solver is efficient for medium and large datasets where it updates the model coefficients using the average of the past gradients computed over small randomly selected subsets of data from the training data. Additionally, by keeping track of the previously gradients it reuses them to converge fast, particularly in datasets with many samples and less features, which is observed by the choice of dual parameter set to False and with L2 regularisation to maintain a smooth function. The stopping tolerance of 1e-1 allows the model to stop the training early once the improvement between iterations is smaller than 1e-1. Smaller values could yield more precise improvements, at the cost of a higher training time, which could potentially risk overfitting by chasing minor fluctuations in the training process. The maximum number of iterations defines how many iterations the solver needs to perform while it searches for the optimal coefficients. By setting 200 iterations, it ensures the model keeps the improvement between iterations falls below the defined tolerance threshold or the number of iterations reaches its defined maximum value. In these settings, since the tolerance is high, the solver tends to converge before the number of iterations limit is reached.

Finally, the warm start was set to False to prevent the model from reusing learned coefficients so that each cross validation train is not biased.

The analysis of both SVC and LR classifiers shows notable differences in performance, stability and suitability of certain data characteristics. Although both models were trained using the same feature set, the SVC clearly outperformed the LR classifier across all evaluated metrics. This superiority aligns with the exploratory analysis, where several features exhibited non-linear relationships with the risk categories and feature correlations. These characteristics benefit models capable of modelling non-linear decision boundaries, such as SVC with rbf kernel, whereas LR, restricted to linear separability with engineered features, struggled to capture complex class distinctions. In terms of performance behaviour, both models showed stable cross validations splits, as evidenced by the consistent low standard deviations. However, SVC showed greater improvements in performance and stability as the number of splits increased, despite the additional variability introduced by having more folds. Compared to SVC, LR gained little from increasing the number of folds, as noticed by the recall that did not improve with a higher number of splits, as well as the emphasis of the model limitation in identifying all true positive cases in a multiclass, relatively balanced dataset. On the other hand, the execution time shows an additional practical consideration in future implementations. The LR is the most computational efficient, with runtimes between 1 and 2 minutes, while the SVC required more time, from 14 to 18 minutes. This difference reflects the computational cost of the kernel methods used as well as the chosen parameter values to acquire the best possible configuration performance. However, despite this increased cost, the balanced performance metrics acquired by the SVC justify its choice for the application of testing datasets to determine if the model can, in fact generalise to unseen data and accurately identify risk classes under real world conditions.

4.2.2 T2D Dataset Related to the User Data

Regarding the T2D dataset, whose data is derived from the CVD dataset, where the data is related to the user smartwatch data, Table 4.16 shows the evolution of the feature selection outcomes across various iterations and splits. Compared to the previous dataset, the model became steadier, where it consistently identified BMI, vo2_max_precise, HRV_day_ms, sleep_time_hr and sleep_resting_heart_heart as the most dominant informative variables for T2D.

Additionally, metrics such as avg_SPO_2, resting_heart_rate and total_calories appeared in some settings, suggests the model's ability to identify complementary informative metrics that help distinguish T2D risk levels. Their selection is consistent with their data distribution, where these variables show a higher variation in inter feature correlation.

Iter	Splits	Number of Features	Features Selected
25	5	7	bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms
50	5	8	bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
75	5	7	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
100	5	6	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms
25	7	7	bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, HRV_day_ms, total_calories
50	7	8	bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
75	7	8	bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
100	7	8	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, HRV_night_ms, total_calories
25	10	7	bmi, vo2_max_precise, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms, total_calories
50	10	7	bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms
75	10	7	bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, HRV_day_ms, total_calories
100	**10**	**7**	**bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, sleep_time_hr, avg_SPO_2, HRV_day_ms**

Table 4.16: Overview of selected Features across different Iterations and Splits of T2D Dataset derived from the CVD Dataset with Data related to User Smartwatch Data

Most configurations converged to 7 or 8 features, with one instance yielding 6 features, while the best performing setting used 100 iterations with 10 splits. The results, given the selected metrics, show that the model likely reached a plateau where most of the informative metrics were consistently identified, in contrast to the additional complementary metrics, which may converge to a limited value gain. These results are expected, since this dataset is derived from smartwatch sensors, which provide cleaner and more consistent data for the measurement of metrics such as HRV, sleep time, and SPO_2, providing clear boundaries between the T2D risk categories.

Table 4.17 summarises the performance of the feature selection with various training/validation settings. The results show a progressive improvement towards model performance as the number of iterations and validation folds increased, where the process benefited from a deeper exploration of the hyperparameter list space. As mentioned above, the best configuration achieved with 100 iterations and 10 splits reached a balanced F1 score of 87.3%, supported by a precision of 88.2% and a recall of 87%.

Iter	Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced Accuracy	Std. Dev. Accuracy	Execution Time (min)
25	5	0.86347	0.02278	0.87435	0.02087	0.86006	0.02399	0.86047	0.02487	2.47
50	5	0.86876	0.02099	0.87688	0.02050	0.86607	0.02171	0.86605	0.02320	5.58
75	5	0.87010	0.01931	0.87909	0.01876	0.86727	0.01987	0.86712	0.02071	8.43
100	5	0.86654	0.02286	0.87484	0.02131	0.86366	0.02392	0.86379	0.02475	11.47
25	7	0.86598	0.02694	0.87594	0.02416	0.86286	0.02834	0.86254	0.02861	3.40
50	7	0.86556	0.03063	0.87452	0.02920	0.86286	0.03148	0.86248	0.03248	7.37
75	7	0.87045	0.03043	0.87919	0.02837	0.86800	0.03135	0.86783	0.03172	12.21
100	7	0.87106	0.03428	0.88003	0.03138	0.86857	0.03542	0.86886	0.03574	16.50
25	10	0.86553	0.02569	0.87538	0.02507	0.86243	0.02700	0.86433	0.02778	4.55
50	10	0.87047	0.02830	0.88044	0.02615	0.86740	0.02975	0.86991	0.03049	11.29
75	10	0.87183	0.02563	0.88106	0.02498	0.86906	0.02673	0.87114	0.02805	17.28
100	**10**	**0.87293**	**0.02796**	**0.88166**	**0.02633**	**0.87017**	**0.02926**	**0.87174**	**0.02963**	**23.14**

Table 4.17: T2D Feature Selection Performance of Dataset derived from CVD dataset with Data related to the User Smartwatch Data

The balance suggests that the model was not only able to identify most of the risk labels but also reduced the risk of false positives and false negatives, which maintained an equilibrium between precision and sensitivity. Furthermore, the balanced accuracy of 87.2% reinforces the model's ability to accurately identify the right

support vectors risk labels, which reinforces its ability to generalise its predictions evenly among the T2D risk classes. In terms of variability, the standard deviations remain moderate with a difference between ≈0.028 to 0.03 across all metrics, which indicates that the model was stable and reliable across temporal folds, even as the validation level of complexity increased. Finally, the execution time increased as the number of iterations and splits increased, from 2.5 minutes to 23 minutes for the best configuration. This reinforces that the model computational demand was necessary to make a deeper search into the parameter list space to uncover the patterns of informative metrics to distinguish each support vector T2D risk assessment.

Taking this into account, the following Listing 4.10 presents the selected parameter option values, as the most optimal setting. With this set of configurations, the model was able to achieve its highest results by iteratively identifying different combinations of settings as observed in Table 4.3. The model used the entropy criteria to assess the quality of information gain with each split.

```
Best estimator found: RFE(estimator=RandomForestClassifier(
    criterion='entropy',
    max_depth=6,
    max_features=0.5,
    max_samples=0.5,
    min_samples_leaf=8,
    min_samples_split=6,
    n_estimators=300,
    class_weight=None,
    bootstrap=True),
    n_features_to_select=8))
```

Listing 4.10: Features Selection Best Estimator of T2D Dataset derived from CVD Dataset related to User Data

Compared to Gini, it produces a more balanced and informative partitions, which is beneficial when different distinctions between metrics can indicate different T2D risk categories. The maximum tree depth was limited to the value of 6. This allowed the model to determine how detailed each decision tree could become, which limited the risk of overfitting by avoiding overly complex trees that memorise the training data. Similarly, the number of samples required at each split and the number of samples at each leaf were both set to moderate, high values to encourage, in a way, a smoother and more generalised decision boundary. Regarding the maximum number of features and samples, both set at 0.5, it ensured that each tree is trained using half of the available features and half of the training samples. This promotes diversity among the trees and the ability to generalise over unseen data. Moreover, setting bootstrap to true ensured each tree is trained on a random subset of data to further reinforce the model's robustness and reduce the risk of overfitting. In

terms of capacity, the model used 300 trees, which provided a sufficient stability without excessively increasing computation time. A larger number of trees generally enhances model reliability by averaging multiple independent learners, reducing variance and improving predictive consistency. Finally, the model did not need to set class weight, which meant that the somewhat imbalanced dataset was not severe enough to require compensatory weighting.

The performance of the SVC classifier, in Table 4.18, shows the training/validation process. Across the various cross validation configurations, the model demonstrated a consistent performance, with marginal variations between the split settings. The configuration with 7 splits achieved the best overall balance, with a balanced F1 score of 82.4%, supported by a precision of 83.2% and a recall of 82.1%. The model managed to maintain an equilibrium between precision and sensitivity in identifying true positives T2D risk categories and avoiding false detections.

Splits	Balanced F1 score	Std. Dev. F1 score	Balanced Precision	Std. Dev. Precision	Balanced Recall	Std. Dev. Recall	Balanced Accuracy	Std. Dev. Accuracy	Execution Time (min)
5	0.821833	0.006149	0.828693	0.008077	0.819219	0.006952	0.816808	0.007221	16.18
7	**0.823883**	**0.027367**	**0.832002**	**0.027357**	**0.820571**	**0.027666**	**0.817308**	**0.028859**	**23.08**
10	0.821914	0.025732	0.829355	0.027931	0.819337	0.024960	0.818147	0.026646	31.28

Table 4.18: SVC Training Results with GridSearchCV of T2D Dataset derived from the CVD Dataset with Data related to User Smartwatch Data

Furthermore, the balanced accuracy of 81.7% reinforces the model with the capability to generalise well across classes, without disproportionately favouring any specific class. Regarding standard deviations, the model performance slightly fluctuated across folds. However, the 7 split choice managed to outperform both 5 and 10 splits, despite having similar scores, which implies that the level of cross validation provided and the optimal trade-off between optimisation and generalisation. With an execution time of ≈23 minutes, compared to the other split configuration times, the 7 split setup demonstrated an efficient use of resources to capture informative data patterns.

The following Listing 4.11 presents the most optimal parameter choices that lead to the best performance of SVC with 7 splits.

```
Best parameters found: {
    'svm__C': 100,
    'svm__break_ties': True,
    'svm__class_weight': 'balanced',
    'svm__decision_function_shape': 'ovr',
    'svm__gamma': 'scale',
    'svm__kernel': 'rbf',
    'svm__shrinking': True,
    'svm__tol': 0.01
}
```

Listing 4.11: SVC Best Estimator for T2D Dataset derived from CVD Dataset not related to User Data

With this configuration set, Table 4.5 shows the full configuration choices to interpret its choices. The model chose the rbf kernel with a regularisation of 100 and gamma scale of auto. The model chose a weak regularisation since it prioritised to correctly classify the training data, while at the same to attempt to capture non-linear patterns present in physiological and behavioural health data, between metrics such as HRV, sleep time and metabolic variables associated with T2D. Concerning the gamma parameter set to scale, it automatically adjusts the influence of each training sample based on the data variance. This ensures a balanced sensitivity towards the decision boundary, while it tries to avoid overfitting if the gamma was too high. Regarding class weight, the model opted for the balanced option, since it felt the need to compensate underrepresented classes by assigning higher importance to the coefficients, to prevent the model from being biased towards the majority of the outcomes. Moreover, the choice for OVR allowed the model to make binary decisions for each class independently, while break ties set to true ensured that in ambiguous cases, the model chose the class with the highest confidence score, to improve prediction stability. Finally, the shrinking set to true enabled a faster and more efficient convergence of the optimisation process by employing a heuristic to reduce the number of active support vectors during training. With the assistance of a tolerance value of 1e-1, it defined a small stopping criteria for the solver to find a convergence without requiring a higher training time to find additional patterns that may bring minimal improvements.

Finally, regarding LR classifier performance, Table 4.19 shows a modest variation in performance across different cross validation splits. The best results were obtained with 7 splits, with a balanced F1 score of 74.3%, supported by the balanced precision of 75.8% and a recall of 73.3%. The model maintains a reasonable trade-off between predicting correct positives cases and avoiding false positives within the different T2D risk labels. The balanced accuracy score of 73.3% further emphasises the difficulties in distinguishing the risk cases of T2D. This result is expected since LR relies on linear decision boundaries to capture more complex data patterns, whereas the data in this dataset present non-linear patterns.

Splits	**Balanced F1 score**	**Std. Dev. F1 score**	**Balanced Precision**	**Std. Dev. Precision**	**Balanced Recall**	**Std. Dev. Recall**	**Balanced Accuracy**	**Std. Dev. Accuracy**	**Exec Time (min)**
5	0.74241	0.02544	0.75981	0.02337	0.73514	0.02729	0.73332	0.02789	1.10
7	**0.74250**	**0.03142**	**0.75805**	**0.03765**	**0.73600**	**0.03054**	**0.73273**	**0.02968**	**1.37**
10	0.74000	0.03229	0.75415	0.03280	0.73481	0.03164	0.73199	0.02660	2.12

Table 4.19: Logistic Regression Training Results with GridSearchCV of T2D Dataset derived from the CVD Dataset with Data related to User Smartwatch Data

However, compared to the SVC classifier, the execution speed of LR is superior. With execution times between 1 and 2 minutes, LR was substantially faster than SVC, which required between 16 and 31 minutes, depending on the number of splits. Nonetheless, the execution time has its advantage, varying between 1 and 2 minutes, which makes it faster than the SVC algorithm, which required between 17 and 23 minutes.

A notable advantage of LR lies in its computational efficiency. With execution times ranging from approximately 1 to 2 minutes, LR was substantially faster than SVC, which required between 16 and 31 minutes depending on the number of splits. This makes LR an attractive option when computational resources or time constraints are critical, even if its predictive performance is lower.

The following Listing 4.12 shows the optimal estimator with 8 variables. This configuration, chosen by the parameter list space in Table 4.7, shows a regularisation strength of 0.3. This indicates that the model applied a strong regularisation, which prevents the model from overfitting and allows it to generalise better, even if it means misclassifying some points. This is useful given the variability present in physiological features in the LR training/validation results.

```
Best parameters found: {
    'lr__C': 0.3,
    'lr__class_weight': 'balanced',
    'lr__dual': False,
    'lr__fit_intercept': True,
    'lr__max_iter': 300,
    'lr__penalty': 'l2',
    'lr__solver': 'sag',
    'lr__tol': 0.1,
    'lr__warm_start': True
}
```

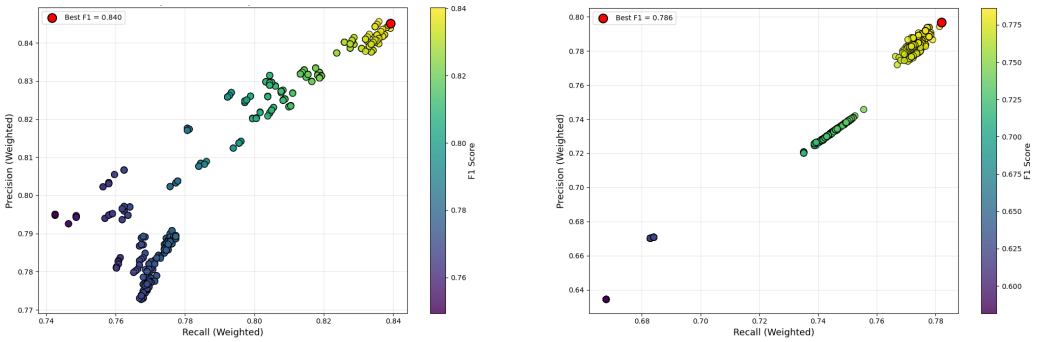
Listing 4.12: Logistic Regression Best Estimator for T2D Dataset derived from CVD Dataset related to User Data

The penalty set to L2 supports this regularisation, where it penalises the features that dominate the model, so that it can smooth the decision boundary and improve the strength across splits. In combination with the regularisation, it prevents the model from overfitting. Furthermore, setting the intercept term to true allows the model to learn from bias and make the decision boundary shift away from the origin so that the values are properly separated between classes without forcing them to be near zero. The model also chose the choice of balanced class weight since it felt the need to compensate classes with lower representation with higher coefficients. This ensures the model does not ignore minority classes, which in turn tends to improve the recall performance of the model. The dual formulation is appropriate

in this case, since it's only implemented with the use of L2 penalty, but also when the number of samples exceeds the number of features. Furthermore, it improves the efficiency of the execution time, as seen by the 1-2 minute runs. Moreover, the Sag solver choice for optimisation is particularly effective in medium and large datasets with L2 regularisation since it allows a faster convergence compared to classical solvers. By combining this solver with 300 iterations, to make sure the solver had enough iterations to converge to a stable solution, while a tolerance of 1e-1 assists, as well, with the stopping criteria, which matches the convergence performance with computational speed. Finally, by choosing the warm start to true, the model reuses the previous solution when the fit function is called again. This may be one of the reasons for this decrease in performance in training, because by using the learned coefficients from the previous function call, it could have acquired a suboptimal solution, which could have biased the new function call. This may have prevented the function from reaching its best possible solution if it had been fitted from scratch instead. As a result, the model may have converged faster, but it may have affected the performance metrics results.

4.2.3 Comparison of T2D Datasets Training Models

Both datasets provide substantial information that influenced the feature selection process and the training/validation performance of the SVC and LR algorithms. Across the datasets, Figure 4.3 and Figure 4.4, the SVC classifier consistently outperformed LR, which validates its ability to capture non-linear relationships within the T2D risk categories datasets.



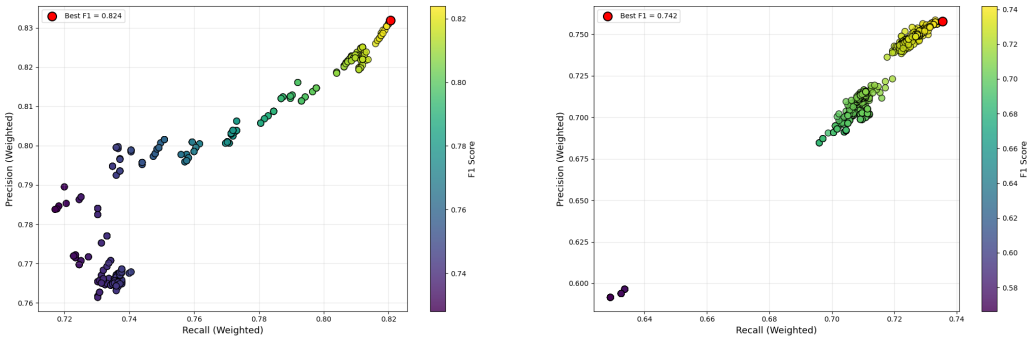
(a) SVC F1 Scores Distribution

(b) LR F1 Scores Distribution

Figure 4.3: Comparison of F1 Scores from SVC and LR Classifiers with Best Performance Results on the T2D Dataset whose Data is derived from the CVD Dataset which is not related to the User Smartwatch Data

However, the performance levels varied, either in performance, stability or cross validations configurations. In the dataset whose data is derived from the CVD dataset, but not related to the user smartwatch, demonstrated a stronger training

performance in both classifiers, with lower standard deviations and better outcomes from deeper cross validations configurations.



(a) SVC F1 Scores Distribution

(b) LR F1 Scores Distribution

Figure 4.4: Comparison of F1 Scores from SVC and LR Classifiers of Best Performance Results on the T2D Dataset whose Data is derived from the CVD dataset which is related to the User Smartwatch Data

Another reason for this performance is due to the generated data being stable, structured and with a higher variability that is normally consistent in synthetic datasets. This helps models such as SVC to detect consistent patterns between features. In contrast, the dataset whose data is derived from the CVD dataset and is related to the collected user smartwatch data, exhibited a lower performance, but had a lower variability in metrics such as HRV, sleep metrics, RHR and VO_2max. With the increase in the number of cross validation folds, the presence of these derived smartwatch sensor variables lead the model to apply a higher effort to find complex relationships between the features and risk labels. With this in mind, the choice of the best model falls on the SVC algorithm applied on the dataset whose data is not related to the user's smartwatch data.

4.3 Testing Phase of CVD and T2D Datasets

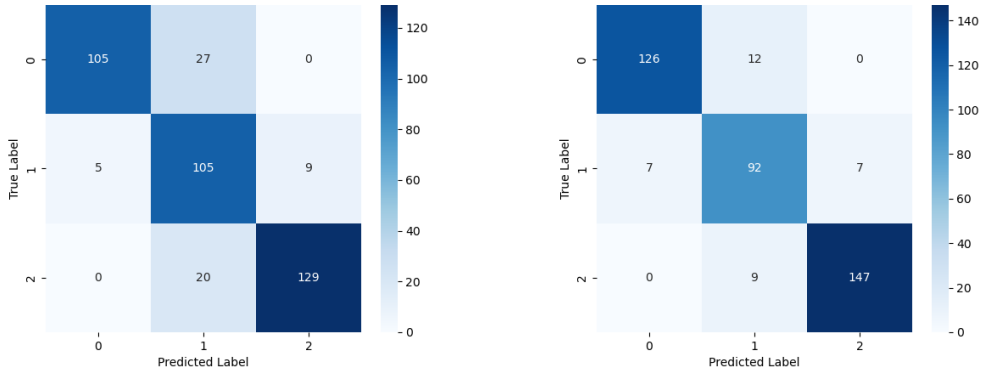
Having benchmarked each CVD and T2D datasets over two machine learning algorithms configurations, we select the best performing models, including the selected features, over the remaining 20% of the testing data, which corresponds to 400 records.

The evaluation of the performance of the CVD datasets, in Table 4.20, presents the testing results acquired from the application of the best model performance on the remaining unseen data. The SVC model, it shows promising results with a greater emphasis on the CVD synthetic dataset related to the user smartwatch data over the non-user related dataset across all metrics. The balanced F1 score increases from 85% to 91.3%, showing a more reliable balance between precision and recall.

Datasets	Balanced F1 Score	Balanced Precision	Balanced Recall	Balanced Accuracy	Execution Time (seconds)
CVD synthetic dataset unrelated to the user smartwatch data	0.852	0.869	0.848	0.848	2.6
CVD synthetic dataset related to the user smartwatch data	**0.913**	**0.915**	**0.912**	**0.908**	**2.4**

Table 4.20: Comparison of CVD Testing Results with the Application of SVC Best Model

The Balanced precision increases from 86.9% to 91.2%, shows a reduction in false positive predictions where the model correctly identifies a high proportion of CVD cases that do not correspond to the actual instances, therefore leading to an increase in true positive cases. Furthermore, the balanced recall increased from 84.8% to 91.2%, shows that the model was able to identify a large number of the various cases of CVD risk, which in medical terms is critical to the detection and timely intervention of healthcare professionals. Lastly, the increase in balanced accuracy from 84.8% to 90.8% indicates that the SVC model was able to effectively identify a large majority of CVD risk classes, even with the potential of class imbalance that can influence the prediction of certain risk classes. By analysing each risk group, Figure 4.5 shows the number of risk cases predicted on a total of 400 records. The number of risk classes, among the 400 records in Figure 4.5a, has 132 records classified as low risk, 119 as medium risk and 149 as high risk. On the other hand, in Figure 4.5b, it has 138 as low risk, 106 as medium risk and 156 as high risk of CVD.



(a) CVD synthetic dataset unrelated to the user smartwatch data

(b) CVD synthetic dataset related to the user smartwatch data

Figure 4.5: Testing results over unseen data on CVD datasets

The dataset unrelated to the user smartwatch data, shown in Figure 4.5a, indicates that the model was able to identify the vast majority of CVD cases. Of 138 low risk records, 105 were correctly predicted, with 27 cases being misclassified as medium risk. In a clinical setting, these 27 cases must be taken into account, as these flagged records could lead to unnecessary follow-ups and patient anxiety, which can lead to an increased use and cost of clinical resources. Furthermore, the

model correctly predicted 105 records belonging to the medium risk group, with 5 being misclassified as low risk and 9 as high risk. Regarding low risk misclassifications, it can be problematic since the model is underestimating the actual patient health status, which can delay preventive care of the CVD risk. Moreover, with an additional 9 cases being flagged as high risk can cause additional stress since the patient needs to undergo closer monitoring and additional interventions to prevent a sudden surge of CVD. Lastly, regarding the high risk cases, the model was able to correctly predict 129, with 20 being flagged as medium risk. It is also worth noting that the model was able to prevent extreme misclassifications from low risk and high risk being assigned as opposite categories, which could be harmful for making decisions in medical assessments.

On the other hand, in Figure 4.5b, the misclassifications are evidently reduced in all risk categories, except for the increase of 2 cases from medium risk being interpreted as low risk. Interestingly, the SVC model acquired from the training phase achieves a clear separation of risk categories and is less prone to errors that could affect interpretation or lead to incorrect clinical analysis. Moreover, the execution time remained nearly identical in both cases, which shows that the enhanced predictive performance did not come at the cost of computational efficiency. Unlike the fully generated dataset, the model was able to capture real world patterns using derived realistic non-invasive smartwatch metrics, which enabled the model to learn meaningful distinctions of low, medium and high CVD risk levels, making it possible to generalise over unseen data.

The RFE feature selection improved the performance by reducing the dimensionality and complexity of the datasets. The selection of metrics such as VO_2max, BMI, oxygen saturation, sleep duration, and sleep HR serves as non-invasive metrics fundamental to the detection of CVD risks. The benchmarking of the CVD datasets shows that non-invasive metrics provide rich physiological information to support the accurate progression of CVD risk, and justify the use of both datasets for future application of machine learning models to assess the cardiovascular risk.

The subsequent analysis focuses on the T2D datasets and the corresponding predictive performance shown in Table 4.21. Across both T2D datasets, the SVC classifier shows consistent results, with a difference of 2% in performance. The application of the RFE combined with the TimeSeriesSplitCV proved effective in identifying a subset of predictive features that resulted in a performance above 85%. Various features, including VO_2max, HRV, total calories, sleep metrics and BMI served as the base for relevant indicators of T2D assessment. These metrics align with existing clinical evidence, as they play a central role in T2D progression.

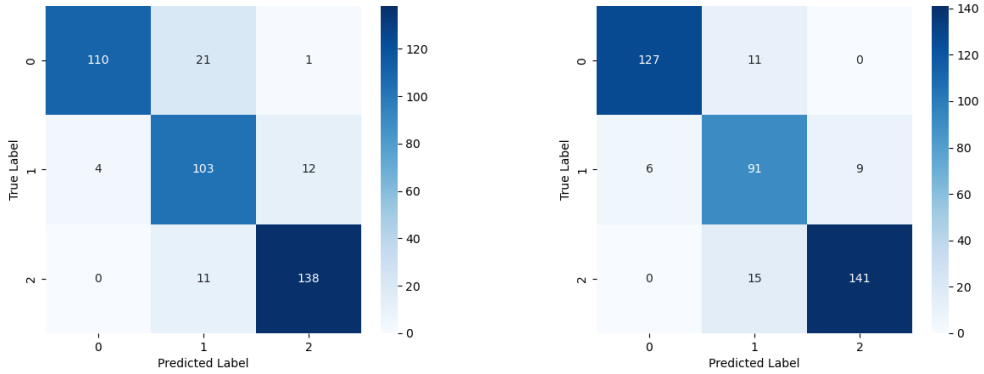
The performance of both datasets shows interesting insights. As previously mentioned in the T2D training assessment, the best performing model emerged from the dataset not related to the user smartwatch data, with SVC as the best algorithm.

Datasets	**Balanced F1 Score**	**Balanced Precision**	**Balanced Recall**	**Balanced Accuracy**	**Execution Time (seconds)**
T2D synthetic dataset unrelated to the user smartwatch data	0.879	0.886	0.877	0.875	4.27
T2D synthetic dataset related to the user smartwatch data	**0.899**	**0.902**	**0.897**	**0.894**	**4.3**

Table 4.21: Comparison of T2D Testing Results with the Application of SVC Best Model

The fully synthetic datasets have a more controlled variability, which allows the model to easily detect relationships during the hyperparameter optimisation. The structured nature of the fully synthetic dataset may have helped with the identification of non-linear patterns. However, despite the training results, the dataset derived from real smartwatch data ultimately outperformed the testing phase with the highest F1 score of 89.9% and overall metrics. This outcome proves that the model was capable of finding the patterns inherent in derived smartwatch data that enhanced the model's generalisation over unseen data, even when the model was originally optimised on a more idealistic synthetic dataset.

A closer analysis of the predicted results in the confusion matrix, shown in Figure 4.6, reveals important insights regarding the classifier behaviour across the prediction of the risk categories. The dataset in Figure 4.6a corresponding to the unrelated smartwatch data shows that out of 132 low risk cases, 110 were classified correctly, with 21 misclassified as medium risk, and 1 as high risk. This high risk case shows that the model had a higher difficulty in detecting this case as low risk, because the features showed characteristics close to those typically associated with medium or high risk, leading to an overestimation of the risk. Furthermore, another reason for this overestimation may be due to the overlapping data that made the distinction of the risk categories more challenging. In terms of medium risk assessment, 103 cases were correctly predicted out of 119, with minor misclassifications of 4 cases being predicted as lower risk, and 12 cases as medium risk. Misclassification of the lower risk could delay the initial interventions to prevent the early development of T2D, although not so harmful as the high risk cases, since the potential to develop immediate complications typically tends to be slower. Of 149 high risk cases, 138 were accurately detected, with 11 being misclassified as medium risk. These results indicate the model's ability to predict T2D risk cases with minimal misclassifications, which minimises the likelihood of underestimating cases that require urgent interventions.



(a) T2D synthetic dataset unrelated to the user smartwatch data

(b) T2D synthetic dataset related to the user smartwatch data

Figure 4.6: Testing results over unseen data on T2D datasets

The dataset in Figure 4.6b shows similar patterns with slightly improved correct predictions of T2D. Of 138 low risk cases, 127 were classified correctly, while 11 were misclassified as medium risk. In the medium risk category, 91 of 106 records were accurately identified as medium risk cases, with 6 as low risk and 9 as high risk. The misclassification of medium risk cases as low risk may delay the start of preventive interventions, while high risk overestimation could lead to even more intensive tests, potentially increasing the patient burden and healthcare costs. Lastly, the model predicted 141 high risk cases of 156, with 15 misclassifications to the medium risk level. This model shows that it's capable of identifying early risk cases of T2D. Furthermore, no extreme misclassifications occurred, which is desirable in healthcare risk assessment, while medium risk cases tend to be slightly prone to overestimation. These results reinforces the relevance of non-invasive features derived from real world smartwatch data related to T2D, including VO_2max, sleep time, BMI, and estimated metrics including HRV and total calories, all of which were selected by the RFE feature selection as strong predictors of early T2D progression.

Table A.2 demonstrates how the proposed architecture pipeline addressed the objectives and research questions that guided this project. The study identified key non-invasive metrics such as VO_2max, BMI, sleep time, heart rate variability, oxygen saturation and activity metrics that contribute to the early detection of CVD and T2D. The safe acquisition of smartwatch data from the Garmin Web Portal using an open source API project confirmed the feasibility of wearable technologies for continuous health monitoring, with concerns that align with *RQ1*. The application of feature selection methods and machine learning algorithms, particularly SVC, allowed to capture non-linear physiological relationships among the variables, which addresses *RQ2-RQ4*. The validation of the predictive power of derived smartwatch measurements and the model's ability to minimise misclassifications across CVD and

T2D risk groups confirm that the developed models offer a foundation for chronic disease assessment. Furthermore, the benchmarking and comparative analysis between synthetic datasets and derived smartwatch datasets revealed that, despite the synthetic data facilitating the model training, the real world physiological structure present in smartwatch based datasets proved to be better at generalisation during testing. These outcomes show the potential use in actionable, real time assessments in early chronic disease predictions.

Chapter 5

Conclusion

The main objective of this work centered around the acquisition of real time physiological data from the Garmin Venu 2 Plus smartwatch, particularly VO_2max and related cardiometabolic indicators, that could help with the early detection of CVD and T2D. In this context, the project proposed the construction of a pipeline that allowed the safe acquisition of real time data from a smartwatch, the implementation of feature selection techniques and the development of predictive models capable of identifying early risk patterns.

The presented results obtained throughout this project conclude that the initially established goals were partially achieved. The implementation of a real data acquisition framework with the help of a Garmin Connect API demonstrated a secure extraction of data from a commercial smartwatch through the use of Consumer and Secret Key credentials. These ensured the necessary authentication and access to the user information during the transfer process. This mechanism allows the data integrity, privacy and compliance through the secure handling of personal health information.

Given the limited amount of real world data collected over a period of five months, and the implementation of a risk stratification strategy, supported by validated scientific works, that initially classified all smartwatch records with no risk, we constructed synthetic datasets to support the implementation of the feature selection and model development. Consequently, we set the goal of implementing two datasets, for both CVD and T2D, which distinguish those directly related to the user

smartwatch data and those generated independently. This allowed the experimentation and benchmarking under three risk distribution conditions for comparative analysis.

The correlation analysis and the implementation of the RFE feature selection technique with TimeSeriesSplitCV, over the CVD and T2D synthetic datasets, confirmed that VO_2max, heart rate metrics, BMI, sleep, respiration metrics, oxygen saturation and activity variables have a strong representation, in the progression of CVD and T2D assessment. With RF as the RFE estimator, it was particularly advantageous due to its ability of multicollinearity to capture linear and non-linear relationships among health metrics. Furthermore, RandomizedSearchCV allowed an efficient search space of the hyperparameter list, which reduced the computational burden compared to exhaustive methods. Although the performance gains may have reached a plateau in some cases, after a certain number of sample configurations, its ability to identify meaningful metrics, despite hardware capabilities, proved particularly beneficial in this project.

The selection of SVM and LR algorithms for the model evaluation, with the assistance of GridSearchCV, enabled reliable optimisation models for both the CVD and T2D risk classification. Among the applied ML models, SVM was able to achieve better results among all performance metrics in both chronic disease datasets. This model was able to capture non-linear relationships among the selected variables, in comparison to LR that, despite finding linear connections, provided a lower performance throughout the entire hyperparameter cross validation search conducted by GridSearchCV. The final SVM models demonstrated encouraging results in the testing phase, confirming that the proposed approach can effectively distinguish the low, medium and high risk categories of non-invasive derived smartwatch data. In addition, the performance metrics, such as balanced accuracy, F1 score, balanced precision and balanced recall, assisted the model in capturing relevant patterns to effectively determine the CVD and T2D progression. Lastly, although the best training models emerged from the synthetic datasets not related to user smartwatch data, the highest testing performance was obtained from datasets derived from real world data, which shows the importance of realistic health metrics distributions in achieving reliable model generalisation.

Taken together, the relevance of this study supports the feasibility of using non-invasive smartwatch metrics for the CVD and T2D disease progression and their eventual integration into clinical decision support systems.

Chapter 6

Limitations and Future Work

This project presents several limitations that should be acknowledged.

The smartwatch Venus 2 Plus used for data collection lacked certain health metrics such as HRV, skin temperature, insulin and glucose levels. These parameters could provide additional insights into the assessment of CVD and T2D conditions. Furthermore, the stress level metric obtained from the smartwatch did not align with real world physiological responses, which prevented the reliability of related stress analysis.

Another limitation concerns the size of the initial dataset obtained by the smartwatch. Although the data collected over a period of five months provided insights into the CVD and T2D conditions, the dataset remained relatively small. Even with the inclusion of a criteria strategy to determine the risk cases for each record, the sample size would still be insufficient to guarantee model generalisation. To mitigate this limitation, we applied synthetic data generation algorithms to create multiple datasets for CVD and T2D. However, the use of synthetic datasets limits the representation of real world data over the application of the ML models.

From a hardware standpoint, the study was conducted on a virtual machine with limited resources, which constrained the efficiency of the model training and testing. One of these limitations was the inability to compute additional parameter settings that could have revealed deeper relationships between health metrics and disease progression, over a higher search space. Moreover, the application of the AUC evaluation metric significantly increased the processing time, with executions going over several hours.

Finally, the target column creation relied on scientific validated research studies and journal combinations of metrics, but lacked clinical validation from medical professionals, which represents a gap between technical implementation and medical applicability.

For future work, we aim to address these limitations through several improvements. One of the priorities lies in the acquisition of a newer generation smartwatch, which offers enhanced sensing capabilities and additional biometrics that can further support the identification of early CVD and T2D risk scenarios. Expanding the dataset, by either extending the data collection period or by including additional participants, could improve the model robustness and generalisation. Furthermore, having access to a more powerful computing hardware would allow a faster and more efficient model training and testing.

In addition, future research should explore the implementation of alternative feature selection techniques and test a wider range of machine learning algorithms to assess their performance in the early prediction of chronic diseases. Finally, the collaboration with healthcare professionals is essential to validate the proposed criteria stratification, to ensure the risk classification over the developed models aligns with clinical standards and contributes to medical decision support systems.

References

[Abbas et al., 2024] Abbas, U., Shah, S. A., Babar, N., Agha, P., Khowaja, M. A., Nasrumminallah, M., Arif, H. E., Hussain, N., Hasan, S. M., and Baloch, I. A. (2024). Cardiorespiratory dynamics of type 2 diabetes mellitus: An extensive view of breathing and fitness challenges in a diabetes prevalent population. *Plos one*, 19(7):e0303564. [Cited on page 3]

[Abdul Basith Khan et al., 2020] Abdul Basith Khan, M., Hashim, M. J., King, J. K., Govender, R. D., Mustafa, H., and Al Kaabi, J. (2020). Epidemiology of type 2 diabetesâĂŤglobal burden of disease and forecasted trends. *Journal of epidemiology and global health*, 10(1):107–111. [Cited on page 3]

[Ahmadi et al., 2024] Ahmadi, M. N., Rezende, L. F., Ferrari, G., Cruz, B. D. P., Lee, I.-M., and Stamatakis, E. (2024). Do the associations of daily steps with mortality and incident cardiovascular disease differ by sedentary time levels? a device-based cohort study. *British Journal of Sports Medicine*, 58(5):261–268. [Cited on page 39]

[Alhaddad et al., 2022] Alhaddad, A. Y., Aly, H., Gad, H., Al-Ali, A., Sadasivuni, K. K., Cabibihan, J.-J., and Malik, R. A. (2022). Sense and learn: recent advances in wearable sensing and machine learning for blood glucose monitoring and trend-detection. *Frontiers in Bioengineering and Biotechnology*, 10:876672. [Cited on page 21]

[Almutairi et al., 2024] Almutairi, A. H., Almutairi, N. S., Mousa, N., Elsayed, A., El-Sehrawy, A., and Elmetwalli, A. (2024). Aerobic exercise as a non-pharmacological intervention for improving metabolic and hemodynamic profiles in type 2 diabetes. *Irish Journal of Medical Science (1971-)*, pages 1–10. [Cited on page 4]

[Amin et al., 2021] Amin, T., Mobbs, R. J., Mostafa, N., Sy, L. W., and Choy, W. J. (2021). Wearable devices for patient monitoring in the early postoperative period: a literature review. *Mhealth*, 7. [Cited on page 5]

[Babu et al., 2024] Babu, M., Lautman, Z., Lin, X., Sobota, M. H., and Snyder, M. P. (2024). Wearable devices: implications for precision medicine and the future of health care. *Annual Review of Medicine*, 75(1):401–415. [Cited on page 4]

[Balducci et al., 2023] Balducci, S., Haxhi, J., Vitale, M., Mattia, L., Sacchetti, M., Orlando, G., Cardelli, P., Iacobini, C., Bollanti, L., Conti, F., et al. (2023). Sustained increase in physical fitness independently predicts improvements in cardiometabolic risk profile in type 2 diabetes. *Diabetes/Metabolism Research and Reviews*, 39(6):e3671. [Cited on page 3]

[Banack et al., 2025] Banack, H. R., Kim, C. D., Cook, C. E., Wasser, A., Kaufman, J. S., and Stovitz, S. D. (2025). B mi-for-age percentile curves for older adults. *Obesity*. [Cited on page 37]

[Barthel et al., 2013] Barthel, P., Wensel, R., Bauer, A., Müller, A., Wolf, P., Ulm, K., Huster, K. M., Francis, D. P., Malik, M., and Schmidt, G. (2013). Respiratory rate predicts outcome after acute myocardial infarction: a prospective cohort study. *European heart journal*, 34(22):1644–1650. [Cited on page 40]

[Baumert et al., 2019] Baumert, M., Linz, D., Stone, K., McEvoy, R. D., Cummings, S., Redline, S., Mehra, R., and Immanuel, S. (2019). Mean nocturnal respiratory rate predicts cardiovascular and all-cause mortality in community-dwelling older men and women. *European Respiratory Journal*, 54(1). [Cited on page 40]

[Bayes, 1968] Bayes, T. (1968). Naive bayes classifier. *Article Sources and Contributors*, pages 1–9. [Cited on page 16]

[Bayoumy et al., 2021] Bayoumy, K., Gaber, M., Elshafeey, A., Mhaimeed, O., Dineen, E. H., Marvel, F. A., Martin, S. S., Muse, E. D., Turakhia, M. P., Tarakji, K. G., et al. (2021). Smart wearable devices in cardiovascular care: where we are and how to move forward. *Nature Reviews Cardiology*, 18(8):581–599. [Cited on page 5]

[Belson, 1959] Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8(2):65–75. [Cited on page 16]

[Bergstra and Bengio, 2012] Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2). [Cited on page 19]

[Berkson, 1953] Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, 48(263):565–599. [Cited on page 12]

[Bernfort et al., 2020] Bernfort, L., Husberg, M., Wiréhn, A.-B., Rosenqvist, U., Gustavsson, S., Karlsdotter, K., and Levin, L.-Å. (2020). Disease burden and

healthcare costs for t2d patients with and without established cardiovascular disease in sweden: a retrospective cohort study. *Diabetes Therapy*, 11:1537–1549. [Cited on page 2]

[Bogue-Jimenez et al., 2022] Bogue-Jimenez, B., Huang, X., Powell, D., and Doblas, A. (2022). Selection of noninvasive features in wrist-based wearable sensors to predict blood glucose concentrations using machine learning algorithms. *Sensors*, 22(9):3534. [Cited on pages 14 and 23]

[Bouqentar et al., 2024] Bouqentar, M. A., Terrada, O., Hamida, S., Saleh, S., Lamrani, D., Cherradi, B., and Raihani, A. (2024). Early heart disease prediction using feature engineering and machine learning algorithms. *Heliyon*, 10(19). [Cited on pages 18 and 24]

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32. [Cited on page 12]

[Butt et al., 2021] Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., and Sherazi, H. H. R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of healthcare engineering*, 2021(1):9930985. [Cited on page 20]

[Cade, 2022] Cade, B. E. (2022). Incident cardiovascular disease risk prediction using extensive oximetry patterns. [Cited on page 39]

[Carrier et al., 2023] Carrier, B., Helm, M. M., Cruz, K., Barrios, B., and Navalta, J. W. (2023). Validation of aerobic capacity (vo2max) and lactate threshold in wearable technology for athletic populations. *Technologies*, 11(3):71. [Cited on page 4]

[Caserman et al., 2024] Caserman, P., Yum, S., Göbel, S., Reif, A., and Matura, S. (2024). Assessing the accuracy of smartwatch-based estimation of maximum oxygen uptake using the apple watch series 7: Validation study. *JMIR Biomedical Engineering*, 9:e59459. [Cited on page 5]

[Castiglione, 2013] Castiglione, F. (2013). Simulating th onset of type 2 diabetes integrating genetic, metabloic and nutritional data. `http://kraken.iac.rm.cnr.it/T2DM`. [Accessed 23-11-2024]. [Cited on page 20]

[Cheng et al., 2021] Cheng, Y., Wang, K., Xu, H., Li, T., Jin, Q., and Cui, D. (2021). Recent developments in sensors for wearable device applications. *Analytical and bioanalytical chemistry*, 413(24):6037–6057. [Cited on page 4]

[Choi et al., 2024] Choi, Y., Kim, G., Yoon, J., and Kim, Y. S. (2024). Association of resting heart rate and physical activity with cardiovascular mortality:

A population-based cohort study of korean adults. *Journal of sports sciences*, 42(16):1529–1537. [Cited on page 37]

[Control and Agency, 1998] Control, K. D. and Agency, P. (1998). Korea National Health and Nutrition Examination Survey. [Cited on page 12]

[Cortes, 1995] Cortes, C. (1995). Support-vector networks. *Machine Learning*. [Cited on page 12]

[Cover and Hart, 1967] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27. [Cited on page 12]

[Devijver and Kittler, 1982] Devijver, P. A. and Kittler, J. (1982). Pattern recognition: A statistical approach. *(No Title)*. [Cited on page 18]

[Di Cesare et al., 2023] Di Cesare, M., Bixby, H., Gaziano, T., Hadeed, L., Kabudula, C., McGhie, D. V., Mwangi, J., Pervan, B., Perel, P., Piñeiro, D., et al. (2023). World heart report 2023: Confronting the worldâĂŹs number one killer. *World Heart Federation: Geneva, Switzerland*. [Cited on page 3]

[Ding and Peng, 2005] Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205. [Cited on page 16]

[Dmitrieva et al., 2022] Dmitrieva, N. I., Liu, D., Wu, C. O., and Boehm, M. (2022). Middle age serum sodium levels in the upper part of normal range and risk of heart failure. *European Heart Journal*, 43(35):3335–3348. [Cited on page 39]

[Ducharme and Gibson, 2021] Ducharme, J. B. and Gibson, A. L. (2021). Efficacy of estimating vo 2 max with the heart rate ratio method in middle-aged and older adults. *European Journal of Applied Physiology*, 121:3431–3436. [Cited on page 3]

[Dunn et al., 2021] Dunn, J., Kidzinski, L., Runge, R., Witt, D., Hicks, J. L., Schüssler-Fiorenza Rose, S. M., Li, X., Bahmani, A., Delp, S. L., Hastie, T., et al. (2021). Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nature medicine*, 27(6):1105–1112. [Cited on pages 13 and 23]

[Ehlers et al., 2021] Ehlers, L. H., Lamotte, M., Monteiro, S., Sandgaard, S., Holmgaard, P., Frary, E. C., and Ejskjaer, N. (2021). The cost-effectiveness of empagliflozin versus liraglutide treatment in people with type 2 diabetes and established cardiovascular disease. *Diabetes Therapy*, 12:1523–1534. [Cited on page 2]

[Ehrenwald et al., 2019] Ehrenwald, M., Wasserman, A., Shenhar-Tsarfaty, S., Zeltser, D., Friedensohn, L., Shapira, I., Berliner, S., and Rogowski, O. (2019).

Exercise capacity and body mass index-important predictors of change in resting heart rate. *BMC cardiovascular disorders*, 19:1–8. [Cited on page 37]

[El-Sofany, 2024] El-Sofany, H. F. (2024). Predicting heart diseases using machine learning and different data classification techniques. *IEEE Access*. [Cited on page 16]

[Fisher, 1970] Fisher, R. A. (1970). Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer. [Cited on page 16]

[Fix, 1985] Fix, E. (1985). *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine. [Cited on page 12]

[Fleuret, 2004] Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(9). [Cited on page 16]

[Galicia-Garcia et al., 2020] Galicia-Garcia, U., Benito-Vicente, A., and Jebari (2020). Pathophysiology of type 2 diabetes mellitus. *International journal of molecular sciences*, 21(17):6275. [Cited on page 3]

[Ghazizadeh et al., 2020] Ghazizadeh, H., Mirinezhad, S. M. R., Asadi, Z., Parizadeh, S. M., Zare-Feyzabadi, R., Shabani, N., Eidi, M., Mosa Farkhany, E., Esmaily, H., Mahmoudi, A. A., et al. (2020). Association between obesity categories with cardiovascular disease and its related risk factors in the mashad cohort study population. *Journal of Clinical Laboratory Analysis*, 34(5):e23160. [Cited on page 37]

[Ghosh et al., 2021] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., and De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, 9:19304–19326. [Cited on pages 15 and 23]

[Goldbloom, 2010] Goldbloom, A. (2010). Kaggle datasets. `https://www.kaggle.com/datasets`. Accessed: 2025-07-23. [Cited on page 19]

[Gomez and Giang, 2024] Gomez, O. N. and Giang, N. T. (2024). Aerobic control. *A Unified System Fitness Design: Concepts of Holistic and Inclusive Fitness Framework*, pages 143–145. [Cited on page 3]

[Gonzales et al., 2023] Gonzales, T. I., Jeon, J. Y., Lindsay, T., Westgate, K., Perez-Pozuelo, I., Hollidge, S., Wijndaele, K., Rennie, K., Forouhi, N., Griffin, S., et al. (2023). Resting heart rate is a population-level biomarker of cardiorespiratory fitness: The fenland study. *Plos one*, 18(5):e0285272. [Cited on pages 38 and 40]

[Group, 2021] Group, T. S. (2021). Long-term complications in youth-onset type 2 diabetes. *New England Journal of Medicine*, 385(5):416–426. [Cited on page 3]

[Gudhmundsson, 2024] Gudhmundsson, J. O. (2024). *Breathing frequency and breathing volume of recreational cyclists and runners*. PhD thesis, School of Social Sciences. [Cited on page 4]

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422. [Cited on page 23]

[Hacker, 2024] Hacker, K. (2024). The burden of chronic disease. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, 8(1):112–119. [Cited on page 4]

[Harber et al., 2024] Harber, M. P., Myers, J., Bonikowske, A. R., Muntaner-Mas, A., Molina-Garcia, P., Arena, R., and Ortega, F. B. (2024). Assessing cardiorespiratory fitness in clinical and community settings: Lessons and advancements in the 100th year anniversary of vo2max. *Progress in Cardiovascular Diseases*, 83:36–42. Cardiorespiratory Fitness and Physical Activity: An Update of Evidence, Global Status and Recommendations. [Cited on pages 3 and 4]

[Häusler et al., 2019] Häusler, N., Marques-Vidal, P., Heinzer, R., and Haba-Rubio, J. (2019). How are sleep characteristics related to cardiovascular health? results from the population-based hypnolaus study. *Journal of the American Heart Association*, 8(7):e011372. [Cited on page 38]

[Himi et al., 2023] Himi, S. T., Monalisa, N. T., Whaiduzzaman, M., Barros, A., and Uddin, M. S. (2023). Medai: A smartwatch-based application framework for the prediction of common diseases using machine learning. *IEEE Access*, 11:12342–12359. [Cited on pages 12 and 23]

[Ho, 1995] Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE. [Cited on page 12]

[Hossain et al., 2020] Hossain, M. E., Uddin, S., Khan, A., and Moni, M. A. (2020). A framework to understand the progression of cardiovascular disease for type 2 diabetes mellitus patients using a network approach. *International Journal of Environmental Research and Public Health*, 17(2):596. [Cited on page 2]

[Huang et al., 2025] Huang, X., Schmelter, F., Seitzer, C., Martensen, L., Otzen, H., Piet, A., Witt, O., Schröder, T., Günther, U. L., Marshall, L., et al. (2025). Digital biomarkers for interstitial glucose prediction in healthy individuals using wearables and machine learning. *Scientific Reports*, 15(1):30164. [Cited on page 22]

[Hughes et al., 2023] Hughes, A., Shandhi, M. M. H., Master, H., Dunn, J., and Brittain, E. (2023). Wearable devices in cardiovascular medicine. *Circulation research*, 132(5):652–670. [Cited on page 4]

[Iversen and Fogelholm, 2023] Iversen, P. O. and Fogelholm, M. (2023). Fluid and water balance: a scoping review for the nordic nutrition recommendations 2023. *Food & Nutrition Research*, 67:10–29219. [Cited on page 39]

[Jabeen et al., 2019] Jabeen, F., Maqsood, M., Ghazanfar, M. A., Aadil, F., Khan, S., Khan, M. F., and Mehmood, I. (2019). An iot based efficient hybrid recommender system for cardiovascular disease. *Peer-to-Peer Networking and Applications*, 12:1263–1276. [Cited on pages 17 and 23]

[James et al., 2021] James, D. E., Stöckli, J., and Birnbaum, M. J. (2021). The aetiology and molecular landscape of insulin resistance. *Nature Reviews Molecular Cell Biology*, 22(11):751–771. [Cited on page 3]

[Janosi, 1989] Janosi, Andras, S. (1989). Heart Disease. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52P4X. [Cited on page 16]

[Javaid et al., 2022] Javaid, A., Zghyer, F., Kim, C., Spaulding, E. M., Isakadze, N., Ding, J., Kargillis, D., Gao, Y., Rahman, F., Brown, D. E., et al. (2022). Medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology. *American Journal of Preventive Cardiology*, 12:100379. [Cited on page 5]

[Jiang et al., 2021] Jiang, H., Mao, H., Lu, H., Lin, P., Garry, W., Lu, H., Yang, G., Rainer, T. H., and Chen, X. (2021). Machine learning-based models to support decision-making in emergency department triage for patients with suspected cardiovascular disease. *International Journal of Medical Informatics*, 145:104326. [Cited on pages 22 and 24]

[Jiang et al., 2023] Jiang, J., Sun, X., Chen, R., Su, Y., Xu, W., Cheng, C., and Zhang, S. (2023). Association between nighttime heart rate and cardiovascular mortality in patients with implantable cardioverter-defibrillator: A cohort study. *Heart Rhythm*, 20(12):1682–1688. [Cited on page 40]

[Kaminsky et al., 2019] Kaminsky, L. A., Arena, R., Ellingsen, Ø., Harber, M. P., Myers, J., Ozemek, C., and Ross, R. (2019). Cardiorespiratory fitness and cardiovascular disease-the past, present, and future. *Progress in cardiovascular diseases*, 62(2):86–93. [Cited on page 3]

[Khan and Algarni, 2020] Khan, M. A. and Algarni, F. (2020). A healthcare monitoring system for the diagnosis of heart disease in the iomt cloud environment using msso-anfis. *IEEE access*, 8:122259–122269. [Cited on pages 23 and 24]

[Kim, 2021] Kim, M.-J. (2021). Building a cardiovascular disease prediction model for smartwatch users using machine learning: Based on the korea national health and nutrition examination survey. *Biosensors*, 11(7):228. [Cited on pages 12 and 23]

[Kira and Rendell, 1992a] Kira, K. and Rendell, L. A. (1992a). The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the tenth national conference on Artificial intelligence*, pages 129–134. [Cited on page 15]

[Kira and Rendell, 1992b] Kira, K. and Rendell, L. A. (1992b). A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier. [Cited on page 15]

[Klinkien, 2020] Klinkien, R. (2020). Python: Garmin connect. `https://github.com/cyberjunky/python-garminconnect`. Accessed: 2025-10-20. [Cited on pages 28 and 32]

[Kohavi, 1995] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Morgan Kaufman Publishing*. [Cited on page 13]

[Kuhn, 2008] Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, 28:1–26. [Cited on page 12]

[Kumar et al., 2022] Kumar, A. K., Ritam, M., Han, L., Guo, S., and Chandra, R. (2022). Deep learning for predicting respiratory rate from biosignals. *Computers in biology and medicine*, 144:105338. [Cited on page 38]

[Kumar et al., 2023] Kumar, S., Victoria-Castro, A. M., Melchinger, H., OâĂŹConnor, K. D., Psotka, M., Desai, N. R., Ahmad, T., and Wilson, F. P. (2023). Wearables in cardiovascular disease. *Journal of Cardiovascular Translational Research*, 16(3):557–568. [Cited on page 4]

[Laranjo et al., 2024] Laranjo, L., Lanas, F., Sun, M. C., Chen, D. A., Hynes, L., Imran, T. F., Kazi, D. S., Kengne, A. P., Komiyama, M., Kuwabara, M., et al. (2024). World heart federation roadmap for secondary prevention of cardiovascular disease: 2023 update. *Global heart*, 19(1). [Cited on page 2]

[Lee et al., 2022] Lee, S., Chu, Y., Ryu, J., Park, Y. J., Yang, S., and Koh, S. B. (2022). Artificial intelligence for detection of cardiovascular-related diseases from wearable devices: a systematic review and meta-analysis. *Yonsei medical journal*, 63(Suppl):S93. [Cited on page 5]

[Lee and Lee, 2020] Lee, S. M. and Lee, D. (2020). Healthcare wearable devices: an analysis of key factors for continuous use intention. *Service Business*, 14(4):503–531. [Cited on page 4]

[Lehmann et al., 2023] Lehmann, V., Föll, S., Maritsch, M., van Weenen, E., Kraus, M., Lagger, S., Odermatt, K., Albrecht, C., Fleisch, E., Zueger, T., et al. (2023). Noninvasive hypoglycemia detection in people with diabetes using smartwatch data. *Diabetes Care*, 46(5):993–997. [Cited on page 14]

[Li et al., 2019] Li, B., Ding, S., Song, G., Li, J., and Zhang, Q. (2019). Computer-aided diagnosis and clinical trials of cardiovascular diseases based on artificial intelligence technologies for risk-early warning model. *Journal of medical systems*, 43(7):228. [Cited on pages 17 and 24]

[Li et al., 2020] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., and Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE access*, 8:107562–107582. [Cited on pages 16 and 23]

[Liang et al., 2023] Liang, Y. Y., Feng, H., Chen, Y., Jin, X., Xue, H., Zhou, M., Ma, H., Ai, S., Wing, Y.-K., Geng, Q., et al. (2023). Joint association of physical activity and sleep duration with risk of all-cause and cause-specific mortality: a population-based cohort study using accelerometry. *European Journal of Preventive Cardiology*, 30(9):832–843. [Cited on page 37]

[Liu et al., 2023] Liu, Y., Herrin, J., Huang, C., Khera, R., Dhingra, L. S., Dong, W., Mortazavi, B. J., Krumholz, H. M., and Lu, Y. (2023). Nonexercise machine learning models for maximal oxygen uptake prediction in national population surveys. *Journal of the American Medical Informatics Association*, 30(5):943–952. [Cited on page 17]

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc. [Cited on page 14]

[Mathur et al., 2020] Mathur, P., Srivastava, S., Xu, X., and Mehta, J. L. (2020). Artificial intelligence, machine learning, and cardiovascular disease. *Clinical Medicine Insights: Cardiology*, 14:1179546820927404. [Cited on page 5]

[Matthews et al., 1985] Matthews, D. R., Hosker, J. P., Rudenski, A. S., Naylor, B., Treacher, D. F., and Turner, R. (1985). Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *diabetologia*, 28:412–419. [Cited on page 24]

[Metwally et al., 2025] Metwally, A. A., Heydari, A. A., McDuff, D., Solot, A., Esmaeilpour, Z., Faranesh, A. Z., Zhou, M., Savage, D. B., Heneghan, C., Patel, S., et al. (2025). Insulin resistance prediction from wearables and routine blood biomarkers. *arXiv preprint arXiv:2505.03784*. [Cited on page 20]

[Mifflin et al., 1990] Mifflin, M. D., St Jeor, S. T., Hill, L. A., Scott, B. J., Daugherty, S. A., and Koh, Y. O. (1990). A new predictive equation for resting energy expenditure in healthy individuals. *The American journal of clinical nutrition*, 51(2):241–247. [Cited on page 45]

[Migliaccio et al., 2024] Migliaccio, G. M., Padulo, J., and Russo, L. (2024). The impact of wearable technologies on marginal gains in sports performance: An integrative overview on advances in sports, exercise, and health. *Applied Sciences*, 14(15):6649. [Cited on page 5]

[Miller et al., 2022] Miller, D. J., Sargent, C., and Roach, G. D. (2022). A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. *Sensors*, 22(16):6317. [Cited on page 5]

[Mohajan and Mohajan, 2023] Mohajan, D. and Mohajan, H. (2023). Long-term regular exercise increases vo2max for cardiorespiratory fitness. *none*. [Cited on page 3]

[Molina-Garcia et al., 2022] Molina-Garcia, P., Notbohm, H. L., Schumann, M., Argent, R., Hetherington-Rauth, M., Stang, J., Bloch, W., Cheng, S., Ekelund, U., Sardinha, L. B., et al. (2022). Validity of estimating the maximal oxygen consumption by consumer wearables: a systematic review with meta-analysis and expert statement of the interlive network. *Sports Medicine*, 52(7):1577–1597. [Cited on page 5]

[Mosenzon et al., 2021] Mosenzon, O., Alguwaihes, A., Leon, J. L. A., Bayram, F., Darmon, P., Davis, T. M., Dieuzeide, G., Eriksen, K. T., Hong, T., Kaltoft, M. S., et al. (2021). Capture: a multinational, cross-sectional study of cardiovascular disease prevalence in adults with type 2 diabetes across 13 countries. *Cardiovascular Diabetology*, 20:1–13. [Cited on page 2]

[Moshawrab et al., 2023] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., and Raad, A. (2023). Smart wearables for the detection of cardiovascular diseases: a systematic literature review. *Sensors*, 23(2):828. [Cited on page 4]

[Myers et al., 2021] Myers, J., Kokkinos, P., Arena, R., and LaMonte, M. J. (2021). The impact of moving more, physical activity, and cardiorespiratory fitness: why we should strive to measure and improve fitness. *Progress in cardiovascular diseases*, 64:77–82. [Cited on page 3]

[Neshitov et al., 2023] Neshitov, A., Tyapochkin, K., Kovaleva, M., Dreneva, A., Surkova, E., Smorodnikova, E., and Pravdin, P. (2023). Estimation of cardiorespiratory fitness using heart rate and step count data. *Scientific Reports*, 13(1):15808. [Cited on page 4]

[Nurmi and Lohan, 2023] Nurmi, J. and Lohan, E. S. (2023). Machine-learning-based diabetes prediction using multisensor data. *IEEE Sensors Journal*, 23(22):28370–28377. [Cited on page 22]

[O'Driscoll et al., 2024] O'Driscoll, B. R., Kirton, L., Weatherall, M., Bakerly, N. D., Turkington, P., Cook, J., and Beasley, R. (2024). Effect of a lower target oxygen saturation range on the risk of hypoxaemia and elevated news2 scores at a university hospital: a retrospective study. *BMJ open respiratory research*, 11(1):e002019. [Cited on page 39]

[O'Grady et al., 2024] O'Grady, B., Lambe, R., Baldwin, M., Acheson, T., and Doherty, C. (2024). The validity of apple watch series 9 and ultra 2 for serial measurements of heart rate variability and resting heart rate. *Sensors*, 24(19):6220. [Cited on page 5]

[Ometov et al., 2021] Ometov, A., Shubina, V., Klus, L., Skibińska, J., Saafi, S., Pascacio, P., Flueratoru, L., Gaibor, D. Q., Chukhno, N., Chukhno, O., et al. (2021). A survey on wearable technology: History, state-of-the-art and current challenges. *Computer Networks*, 193:108074. [Cited on page 4]

[Organization, 2023] Organization, W. H. (2023). *Advancing the global agenda on prevention and control of noncommunicable diseases 2000 to 2020: looking forwards to 2030*. World Health Organization. [Cited on page 3]

[Organization et al., 2019] Organization, W. H. et al. (2019). Classification of diabetes mellitus. *World Health Organization*. [Cited on page 3]

[Paluch et al., 2023] Paluch, A. E., Bajpai, S., Ballin, M., Bassett, D. R., Buford, T. W., Carnethon, M. R., Chernofsky, A., Dooley, E. E., Ekelund, U., Evenson, K. R., et al. (2023). Prospective association of daily steps with cardiovascular disease: a harmonized meta-analysis. *Circulation*, 147(2):122–131. [Cited on page 39]

[Pearson, 1900] Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175. [Cited on page 16]

[Petersen et al., 2022] Petersen, M. H., de Almeida, M. E., Wentorf, E. K., Jensen, K., Ørtenblad, N., and Højlund, K. (2022). High-intensity interval training combining rowing and cycling efficiently improves insulin sensitivity, body composition and vo2max in men with obesity and type 2 diabetes. *Frontiers in endocrinology*, 13:1032235. [Cited on page 3]

[Pittaras et al., 2024] Pittaras, A., Kokkinos, P., Faselis, C., Grassos, C., Doumas, M., Kallistratos, E., Manolis, A., and Samuel, I. B. (2024). Resting heart rate and mortality risk in hypertensive patients with no atrial fibrillation. *Journal of Hypertension*, 42(Suppl 1):e50. [Cited on page 38]

[Pleil et al., 2021] Pleil, J. D., Wallace, M. A. G., Davis, M. D., and Matty, C. M. (2021). The physics of human breathing: flow, timing, volume, and pressure parameters for normal, on-demand, and ventilator respiration. *Journal of breath research*, 15(4):042002. [Cited on page 38]

[Qadri et al., 2023] Qadri, A. M., Raza, A., Munir, K., and Almutairi, M. S. (2023). Effective feature engineering technique for heart disease prediction with machine learning. *IEEE Access*, 11:56214–56224. [Cited on pages 19 and 24]

[Qureshi et al., 2020] Qureshi, K. N., Din, S., Jeon, G., and Piccialli, F. (2020). An accurate and dynamic predictive model for a smart m-health system using machine learning. *Information Sciences*, 538:486–502. [Cited on pages 13 and 23]

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386. [Cited on page 12]

[Roth et al., 2020] Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., et al. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American college of cardiology*, 76(25):2982–3021. [Cited on page 2]

[Rückert-Eheberg et al., 2025] Rückert-Eheberg, I.-M., Steger, A., Müller, A., Linkohr, B., Barthel, P., Maier, M., Allescher, J., Sinner, M. F., Rizas, K. D., Rathmann, W., et al. (2025). Respiratory rate and its associations with disease and lifestyle factors in the general population–results from the kora-ff4 study. *PloS one*, 20(3):e0318502. [Cited on page 39]

[Rutters et al., 2024] Rutters, F., den Braver, N. R., Lakerveld, J., Mackenbach, J. D., van der Ploeg, H. P., Griffin, S., Elders, P. J., and Beulens, J. W. (2024). Lifestyle interventions for cardiometabolic health. *Nature Medicine*, pages 1–13. [Cited on page 3]

[Saad et al., 2024] Saad, H. S., Zaki, J. F., and Abdelsalam, M. M. (2024). Employing of machine learning and wearable devices in healthcare system: tasks and challenges. *Neural Computing and Applications*, 36(29):17829–17849. [Cited on page 5]

[Schnor, 2023] Schnor, N. P. (2023). Pima Indians Diabetes Database — kaggle.com. `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`. [Accessed 23-11-2024]. [Cited on page 20]

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423. [Cited on page 16]

[Speed et al., 2023] Speed, C., Arneil, T., Harle, R., Wilson, A., Karthikesalingam, A., McConnell, M., and Phillips, J. (2023). Measure by measure: Resting heart rate across the 24-hour cycle. *PLOS Digital Health*, 2(4):e0000236. [Cited on page 40]

[Stolfi et al., 2020] Stolfi, P., Valentini, I., Palumbo, M. C., Tieri, P., Grignolio, A., and Castiglione, F. (2020). Potential predictors of type-2 diabetes risk: machine learning, synthetic data and wearable health devices. *BMC bioinformatics*, 21:1–19. [Cited on pages 20 and 23]

[Sun et al., 2009] Sun, Y., Todorovic, S., and Goodison, S. (2009). Local-learning-based feature selection for high-dimensional data analysis. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1610–1626. [Cited on page 16]

[Tatli et al., 2024] Tatli, D., Papapanagiotou, V., Liakos, A., Tsapas, A., and Delopoulos, A. (2024). Prediabetes detection in unconstrained conditions using wearable sensors. *Clinical Nutrition Open Science*, 58:163–174. [Cited on page 15]

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288. [Cited on page 15]

[Timmis et al., 2022] Timmis, A., Vardas, P., Townsend, N., Torbica, A., Katus, H., De Smedt, D., Gale, C. P., Maggioni, A. P., Petersen, S. E., Huculeci, R., et al. (2022). European society of cardiology: cardiovascular disease statistics 2021. *European heart journal*, 43(8):716–799. [Cited on page 2]

[Uth et al., 2004] Uth, N., Sørensen, H., Overgaard, K., and Pedersen, P. K. (2004). Estimation of vİĞ o 2max from the ratio between hr max and hr rest–the heart rate ratio method. *European journal of applied physiology*, 91:111–115. [Cited on page 25]

[Wang et al., 2020] Wang, W., Huang, C., Hsu, P., Lin, C., Wang, Y., Din, Y., Liou, T., Wang, Y., Huang, S., Lu, T., et al. (2020). Association of sleep duration and cardiovascular events. *European Heart Journal*, 41(Supplement_2):ehaa946–1324. [Cited on page 37]

[Wang et al., 2023] Wang, X., Ren, J., Ren, H., Song, W., Qiao, Y., Zhao, Y., Linghu, L., Cui, Y., Zhao, Z., Chen, L., et al. (2023). Diabetes mellitus early warning and factor analysis using ensemble bayesian networks with smote-enn and boruta. *Scientific Reports*, 13(1):12718. [Cited on page 20]

[Weeldreyer et al., 2025] Weeldreyer, N. R., De Guzman, J. C., Paterson, C., Allen, J. D., Gaesser, G. A., and Angadi, S. S. (2025). Cardiorespiratory fitness, body mass index and mortality: a systematic review and meta-analysis. *British journal of sports medicine*, 59(5):339–346. [Cited on page 37]

[Weston and Guyon, 2012] Weston, J. and Guyon, I. (2012). Support vector machineâĂŤrecursive feature elimination (svm-rfe). US Patent 8,095,483. [Cited on page 17]

[Yang et al., 2024] Yang, L., Amin, O., and Shihada, B. (2024). Intelligent wearable systems: Opportunities and challenges in health and sports. *ACM Computing Surveys*, 56(7):1–42. [Cited on page 4]

[Yu et al., 2021] Yu, B., Li, C., Sun, Y., and Wang, D. W. (2021). Insulin treatment is associated with increased mortality in patients with covid-19 and type 2 diabetes. *Cell metabolism*, 33(1):65–77. [Cited on page 3]

[Zhang et al., 2018] Zhang, B., Pei, C., Zhang, Y., Sun, Y., and Meng, S. (2018). High resting heart rate and high bmi predicted severe coronary atherosclerosis burden in patients with stable angina pectoris by syntax score. *Angiology*, 69(5):380–386. [Cited on page 37]

[Zhang et al., 2024] Zhang, T. Y., Du, Y. J., Hou, Y. Z., Du, Q., Dou, H. R., and Gao, X. M. (2024). Heart/breathing rate ratio (hbr) as a predictor of mortality in critically ill patients. *Heliyon*, 10(10). [Cited on page 38]

[Zheng et al., 2018] Zheng, Y., Ley, S. H., and Hu, F. B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature reviews endocrinology*, 14(2):88–98. [Cited on page 3]

Appendix A

Appendix

A.1 Garmin Connect Menu Options

Python: Garmin Connect

```
$ ./example.py
*** Garmin Connect API Demo by cyberjunky ***

Trying to login to Garmin Connect using token data from directory '~/.garmi

1 -- Get full name
2 -- Get unit system
3 -- Get activity data for '2024-11-10'
4 -- Get activity data for '2024-11-10' (compatible with garminconnect-ha)
5 -- Get body composition data for '2024-11-10' (compatible with garminconn
6 -- Get body composition data for from '2024-11-03' to '2024-11-10' (to be
7 -- Get stats and body composition data for '2024-11-10'
8 -- Get steps data for '2024-11-10'
9 -- Get heart rate data for '2024-11-10'
0 -- Get training readiness data for '2024-11-10'
- -- Get daily step data for '2024-11-03' to '2024-11-10'
/ -- Get body battery data for '2024-11-03' to '2024-11-10'
! -- Get floors data for '2024-11-03'
? -- Get blood pressure data for '2024-11-03' to '2024-11-10'
. -- Get training status data for '2024-11-10'
a -- Get resting heart rate data for '2024-11-10'
b -- Get hydration data for '2024-11-10'
c -- Get sleep data for '2024-11-10'
d -- Get stress data for '2024-11-10'
e -- Get respiration data for '2024-11-10'
f -- Get SpO2 data for '2024-11-10'
g -- Get max metric data (like vo2MaxValue and fitnessAge) for '2024-11-10'
h -- Get personal record for user
i -- Get earned badges for user
j -- Get adhoc challenges data from start '0' and limit '100'
k -- Get available badge challenges data from '1' and limit '100'
l -- Get badge challenges data from '1' and limit '100'
m -- Get non completed badge challenges data from '1' and limit '100'
n -- Get activities data from start '0' and limit '100'
o -- Get last activity
p -- Download activities data by date from '2024-11-03' to '2024-11-10'
r -- Get all kinds of activities data from '0'
s -- Upload activity data from file 'MY_ACTIVITY.fit'
t -- Get all kinds of Garmin device info
u -- Get active goals
v -- Get future goals
w -- Get past goals
y -- Get all Garmin device alarms
x -- Get Heart Rate Variability data (HRV) for '2024-11-10'
z -- Get progress summary from '2024-11-03' to '2024-11-10' for all metrics
A -- Get gear, the defaults, activity types and statistics
B -- Get weight-ins from '2024-11-03' to '2024-11-10'
C -- Get daily weigh-ins for '2024-11-10'
D -- Delete all weigh-ins for '2024-11-10'
E -- Add a weigh-in of 89.6kg on '2024-11-10'
F -- Get virtual challenges/expeditions from '2024-11-03' to '2024-11-10'
G -- Get hill score data from '2024-11-03' to '2024-11-10'
H -- Get endurance score data from '2024-11-03' to '2024-11-10'
I -- Get activities for date '2024-11-10'
J -- Get race predictions
K -- Get all day stress data for '2024-11-10'
L -- Add body composition for '2024-11-10'
M -- Set blood pressure "120,80,80,notes='Testing with example.py'"
N -- Get user profile/settings
O -- Reload epoch data for '2024-11-10'
P -- Get workouts 0-100, get and download last one to .FIT file
R -- Get solar data from your devices
S -- Get pregnancy summary data
T -- Add hydration data
U -- Get Fitness Age data for '2024-11-10'
V -- Get daily wellness events data for '2024-11-03'
W -- Get userprofile settings
Z -- Remove stored login tokens (logout)
q -- Exit
Make your selection:
```

Figure A.1: Overview of Garmin Connect API

A.2 Comparison of CVD and T2D Datasets

Dataset	Origin	Records	Variables	Data Type	Purpose
Cardiovascular Diseases (CVD)	Synthetic dataset based on real time physiological smartwatch data	2000	calendar_date, vo2_max_precise, steps, min_heart_rate, max_heart_rate, resting_heart_rate, max_respiration, min_respiration, avg_waking_respiration, weight_kg, fitness_age, bmi, hydration_ml, avg_SPO_2, min_SPO_2, avg_sleep_SPO_2, sleep_time_sec, sleep_avg_respiration, sleep_resting_heart_rate	Numerical dataset with continuous features and an ordinal target variable	Used for machine learning 80% training and validation and 20% for testing for CVD risk detection benchmark
Cardiovascular Diseases (CVD)	Synthetic dataset not related to real time physiological smartwatch data	2000	—	Fully simulated from smartwatch metrics with numerical continuous features and an ordinal target variable	Used for machine learning 80% training and validation and 20% for testing for CVD risk detection benchmark
Diabetes Type 2 (T2D)	Derived subset of features from the CVD dataset with additional parameters relevant to T2D	2000	calendar_date, bmi, vo2_max_precise, resting_heart_rate, sleep_resting_heart_rate, steps, sleep_time_hr, avg_waking_respiration, max_respiration, min_respiration, sleep_avg_respiration, hydration_ml, avg_SPO_2, weight_kg, HRV_day_ms, HRV_night_ms, total_calories, skin_temp_proxy	Continuous numerical features with ordinal target value	Used for machine learning 80% training and validation and 20% for testing for T2D risk detection benchmark
Diabetes Type 2 (T2D)	Derived subset of features from the CVD dataset not related to the real world smartwatch data with additional parameters relevant to T2D	2000	—	Continuous numerical features with ordinal target value	Used for machine learning 80% training and validation and 20% for testing for T2D risk detection benchmark

Table A.1: Comparison of CVD and T2D datasets

A.3 Comprehensive Summary of Objectives, Methods, Results and Research Questions

Objectives	Methods Applied	Results	Evidence
O1 - Explore the correlation between vo2max and additional health metrics with the early progression of CVD and T2D risks by identifying predictive indicators.	Literature Review on VO_2max and additional smartwatch non-invasive physiological biomarkers; Risk stratification criteria using scientific studies to combine multiple biomarkers and assign weighted risk levels to the datasets. Use of correlation heatmaps between derived and fully generated smartwatch datasets: Exploratory analisys in CVD and T2D datasets.	Strong correlation between VO_2max, BMI, sleep, heart rate metrics, oxygen saturation and total calories as risk indicators for CVD nad T2D	Correlation Matrices Figures; Scientific studies to align multiple and individual metrics for risk assigning; Tables with selected metrics from training and testing phases.
O2 - Familiarise with the technical requirements of wearable healthcare devices and retrieve real time data to apply machine learning algorithms.	Study of Garmin Venus 2 Plus Smarwatch and its wearable capabilities on its daily use; Secure data Extraction using Garmin Connect API with Consumer and Secret keys; Construction of structured CVD and T2D datasets.	Secure acquisition of biometric data, using the Garmin Connect Open Source Application Interface; Construction of strucutred datasets of CVD and T2D to benchmark machine learning algorithms.	API access logs; Smartwatch data array; CVD and T2D related and fully generated smartwatch non-invasive metrics datasets for benchmarking; Code Segments with implementation of architecture.
03 - Apply feature selection techniques and ML algorithms to analyse the impact of VO_2max and additional health metrics in the early progression of CVD and T2D.	Feature Scaling using Robust Scaler; Time Series Cross Validation to preserve chronological data; Recursive Feature Elimination selection with Random Forest estimator; Support Vector Machine and Logistic Regression with Grid Search Cross Validation.	Identification of optimal non-invasive predictors; SVM model demonstrated superior ability to capture non-linear relations in both diseases;	RFE metrics ranking; Hyperparameters and best models logs;
O4 - Evaluate predictive models using performance metrics relevant to CVD and T2D risk assessment.	Performance metrics: Balanced F1 Score, Balanced Accuracy, Balanced Precision, Balanced Recall Confusion Matrics; Split of datasets with 80% for training/validation and 20% for testing.	High predictive reliability for low, medium and high risk classes; Smartwatch derived models generalise better in testing, despite synthetic models scoring higher in training.	Training and Testing tables and plots; Confusion matrices from CVD and T2D datasets.

Table A.2: Summary to Connect Objectives, Methods, Results and Research Questions

A.4 Implementation of SVC Training

```
....

svm = SVC()

param_dist = {...}

tscv = TimeSeriesSplit(n_splits=5) # 5, 7, 10 splits

scoring = {
    'f1_weighted': 'f1_weighted',
    'precision_weighted': 'precision_weighted',
    'recall_weighted': 'recall_weighted',
    'balanced_accuracy': 'balanced_accuracy'
}

grid_search = GridSearchCV(
    estimator=svm,
    param_grid=param_dist,
    scoring=scoring,
    refit='f1_weighted',
    cv=tscv,
    pre_dispatch=4,
    return_train_score=True
)

# Track training time
start_time = time.time()

grid_search.fit(np.ascontiguousarray(x_train, dtype=np.float64),
    y_train)

execution_time = time.time() - start_time

# Save the best model
joblib.dump(grid_search.best_estimator_, os.path.join(
    results_folder, 'best_svm_pipeline.pkl'))
```

Listing A.1: Application of SVC Training with Hyperparameter Optimization using GridSearchCV with TimeSeriesSplitCV

A.5 Implementation of LR Training

```
1
2  ...
3
4  lr = LogisticRegression(multi_class='multinomial')
5
6  param_dist = {...}
7
8  scoring = {
9      'f1_weighted': 'f1_weighted',
10     'precision_weighted': 'precision_weighted',
11     'recall_weighted': 'recall_weighted',
12     'balanced_accuracy': 'balanced_accuracy'
13 }
14
15 # Time series cross validation
16 tscv = TimeSeriesSplit(n_splits=5)
17
18 grid_search = GridSearchCV(
19     estimator=lr,
20     param_grid=param_dist,
21     scoring=scoring,
22     refit='f1_weighted',
23     cv=tscv,
24     n_jobs=4
25 )
26
27 # Track training time
28 start_time = time.time()
29 grid_search.fit(np.ascontiguousarray(x_train, dtype=np.float64),
       y_train)
30 execution_time = time.time() - start_time
31
32 # Save best estimator
33 joblib.dump(best_estimator, os.path.join(results_folder, '
       best_lr_pipeline.pkl'))
```

Listing A.2: Application of Logistic Regression Classifier Training with Hyperparameter Optimization using GridSearchCV with TimeSeriesSplitCV

A.6 Testing Phase Example

```
...

try:
    best_model = joblib.load("Database_B/
        svm_results_point_B_grid_search_7_contiguous/
        best_svm_pipeline.pkl")
    print("Loaded successfully!")
except Exception as e:
    print("Error loading model:", e)

# Load selected features
with open("Database_B/results_point_B__25_10/selected_features
    .json", "r") as f:
    selected_features = json.load(f)

y_pred = best_model.predict(X_test)

# Compute balanced metrics
balanced_acc = balanced_accuracy_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred, average='weighted')
precision = precision_score(y_test, y_pred, average='weighted'
    )
recall = recall_score(y_test, y_pred, average='weighted')

# Display results
metrics = {
    "Balanced Accuracy": balanced_acc,
    "F1-score (weighted)": f1,
    "Precision (weighted)": precision,
    "Recall (weighted)": recall
}

print("Balanced Metrics on Test Set:")
for k, v in metrics.items():
    print(f"{k}: {v:.3f}")

print(classification_report(y_test, y_pred))
```

Listing A.3: Testing Phase Example over the Unseen Data on a CVD Dataset