Using Polysomnography-Derived Parameters to Predict Cardiovascular Outcome

Siying He

A thesis submitted to fulfil the requirements of the degree of
Doctor of Philosophy

School of Biomedical Engineering, Faculty of Engineering
The University of Sydney

2025

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged as follows:

1. The original research topics, title, and outlines were developed in discussion with Professor Philip de Chazal and Professor Peter Cistulli.
2. The research content and progress were discussed with members of the Sleep Research Group at the Charles Perkins Centre (CPC), The University of Sydney.
3. Advice and editorial assistance on thesis writing were provided by Professor Philip de Chazal.

Siying He

Acknowledgment

Looking back, this PhD journey has been one of the most challenging yet rewarding experiences of my life, and it would not have been possible without the support, guidance, and encouragement of many wonderful people.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Professor Philip de Chazal, for his unwavering support, patience, and kindness throughout my PhD. I am deeply thankful for his steady guidance whenever I faced research challenges, his detailed feedback on my papers, and his careful proofreading. His willingness to explain even the most fundamental methodological concepts with clarity, together with his encouragement to embrace every opportunity for presentations and academic connections, has been truly inspiring and motivating. His support in exploring teaching roles and his efforts in introducing me to new opportunities have profoundly shaped both my academic and personal growth. Beyond academic mentorship, his genuine care for my wellbeing as an international student studying alone has meant a great deal to me.

I am equally grateful to my secondary supervisor, Professor Peter Cistulli, for his thoughtful guidance and constructive feedback. His insightful advice during project presentations and his reminder to balance technical detail with the broader purpose of research have greatly influenced how I approach my work. I would also like to thank the CPC Sleep Research Group for their collaboration, discussions, and encouragement throughout this project. The group's shared ideas and collegial spirit have enriched my research and fostered lasting connections.

My deepest appreciation goes to my mother and grandparents, whose unconditional love, financial support, and constant encouragement have sustained me through every challenge. Their belief in me has been an enduring source of courage and motivation. Despite the distance between us, their care and quiet understanding of my long absences and late-night work have been my greatest comfort. I also wish to express my deepest love and gratitude to my husband, whose patience, humour, and faith in me have carried me through the ups and downs of this journey. He listened to countless practice talks, helped me refine my thoughts, and reminded me to rest when I pushed too hard. His unwavering support and companionship have made this journey not only possible but truly meaningful and joyful.

Gen AI Attribution Statement

During the preparation of the thesis the author used ChatGPT for the purposes of grammar check and text revision. The use of this generative AI tool includes paraphrasing, clarity enhancement, and grammar correction. The author confirms that where text was modified by generative AI, the content was reviewed for possible errors, inaccuracies, and bias. The author takes full responsibility for the submitted thesis and ensures the work is their own and has used generative AI within the parameters of use.

Siying He

Author Attribution Statement

This thesis contains material that has been published, accepted for publication, or submitted for publication during my candidature. The details of these publications, the extent of their inclusion in the thesis, and the contributions of co-authors are outlined below.

Chapter 3 includes material published as:

- He, S., P. A. Cistulli and P. de Chazal, "A Review of Novel Oximetry Parameters for the Prediction of Cardiovascular Disease in Obstructive Sleep Apnoea," *Diagnostics*, vol. 13, no. 21, p. 3323, 2023, doi: https://doi.org/10.3390/diagnostics13213323.

This publication corresponds to Section 3.1 of Chapter 3. I conducted the literature review, structured the manuscript, and wrote the initial and final drafts. Professor Philip de Chazal and Professor Peter Cistulli contributed to conceptual guidance, topic refinement, comments, and editorial feedback.

Chapter 4 incorporates material that has been partially published and submitted for publication:

- He, S., P. A. Cistulli and P. de Chazal, "Comparison of Oxygen Desaturation Area-Based Methods in Predicting Cardiovascular Disease Mortality Outcomes," *Eur. Respir. J.*, (under review), 2025.
- S. He and P. de Chazal, "The Influence of Arousal Events on Desaturation Area-Based Parameters for Predicting 3-Year Cardiovascular Mortality," *Proc. 47th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Copenhagen, Demark, (accepted), 2025.
- He, S., P. A. Cistulli and P. de Chazal, "Comparison of Oximetry Event Desaturation Transient Area-Based Methods in Predicting Cardiovascular Disease Mortality Outcomes," *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Orlando, USA, 2024, doi: https://doi.org/10.1109/EMBC53108.2024.10782779.
- He, S., K. Cook, K. Sutherland, Y. S. Bin, P. A. Cistulli and P. de Chazal, "A Comparison of Hypoxic Burden Algorithms Using Three Different Methods for Calculating Baseline Oxygen Saturation for Predicting Cardiovascular Death in the Sleep Heart Health Study," *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sydney, Australia, 2023, doi: https://doi.org/10.1109/EMBC40787.2023.10340410.

I designed and conducted the analysis, developed the computational framework, and drafted the manuscripts. Professor Philip de Chazal contributed to algorithm development, supervision,

conceptual input, and manuscript editing. Professor Peter Cistulli provided conceptual guidance and comments on the manuscripts. Additional collaborators (Dr. Kristina Cook, Dr. Kate Sutherland, Dr. Yu Sun Bin) contributed to manuscript revision.

Chapter 5 (Phase 1) includes material published as:

- He, S. and P. de Chazal, "Can Oximetry-Derived Parameters Predict 3-Year Cardiovascular Mortality?" *Proc. 47th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Copenhagen, Denmark, (accepted), 2025.

I designed the experiments, performed the data analysis, and wrote the manuscript. Professor Philip de Chazal provided conceptual guidance, supervision, and manuscript editing.

In addition, this thesis includes supporting material from several conference abstracts and presentations, including *World Sleep 2025*, *Sleep Advances 2024 and 2023*, representing preliminary findings and related analyses contributing to Chapter 4. In all listed works, I am the lead author and responsible for the study design, data analysis and manuscript preparation.

In addition to the authorship attribution statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Siying He
23/10/2025

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Professor Philip de Chazal
24/10/2025

Abstract

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, while obstructive sleep apnoea (OSA) is recognised as an independent risk factor. Polysomnography (PSG), the gold standard for OSA assessment, records multiple physiological signals that can be transformed into parameters with prognostic value. The apnoea–hypopnoea index (AHI), although commonly used as a clinical diagnostic measure for OSA, is less informative when used as a predictor of CVD outcomes. Alternative oximetry-derived measures, including the oxygen desaturation index (ODI), the total sleep time spent below 90% oxygen saturation (T90), and desaturation area–based parameters, provide a more detailed characterisation of nocturnal hypoxaemia. Differences in the American Academy of Sleep Medicine (AASM) criteria, interpretations of parameter definitions, and computational approaches have resulted in methodological inconsistencies and limited comparability across studies. This thesis investigates whether PSG-derived parameters, with a particular focus on oximetry-derived parameters, can improve prediction of CVD mortality, and whether explainable machine learning frameworks can provide robust and clinically relevant outcome prediction at the individual level.

Two experiments were conducted using data from the Sleep Heart Health Study (SHHS). Experiment 1 systematically compared three major desaturation area–based algorithms within a unified computational framework. This comparison showed that algorithmic variations affect parameter value and predictive performance for CVD mortality. It clarified the reasons for inconsistent findings in prior studies and identified the robust and best-performing method, establishing the methodological foundation for subsequent work.

Experiment 2 was designed to test whether combining PSG-derived parameters could improve outcome prediction for CVD mortality beyond single-parameter approaches, while also addressing the current limitation that OSA–CVD analyses often focus on relative hazard rather than individual-level prediction. It evaluated the predictive value of ODI, T90, and desaturation area–based parameters in combination for short-term outcomes. Results showed that the combined use of multiple parameters improved predictive performance compared with single metrics. Based on these findings, an explainable machine learning framework was developed to integrate PSG-derived parameters with demographic, lifestyle, and medical information. The framework achieved strong predictive performance (Area under the curve (AUC) = 0.89± 0.05;

F1 score of 86.20%), generalised across 3-, 5-, and 10-year horizons, and remained interpretable through SHapley Additive exPlanations (SHAP). It also required fewer specialised clinical inputs than existing resource-intensive approaches, supporting its potential for use in large-scale screening, resource-limited healthcare settings, and home-based monitoring.

In summary, this thesis resolves methodological discrepancies in oximetry-derived parameter computation and develops an explainable machine learning framework for individual-level outcome prediction. These contributions demonstrate the potential of PSG-derived data to provide scalable and actionable prediction of CVD mortality.

Table of Contents

Glossary

AAA — Abdominal Aortic Aneurysm; localised dilation of the abdominal aorta that increases the risk of rupture.

AASM — American Academy of Sleep Medicine; provides standardised clinical guidelines for scoring sleep and respiratory events, with key respiratory event scoring standards updated in 1999, 2007, and 2012.

AHI — Apnoea–Hypopnoea Index; number of apnoea and hypopnoea events per hour of sleep, used to assess the severity of obstructive sleep apnoea.

AUC — Area Under the Receiver Operating Characteristic Curve; performance metric quantifies the overall performance of the classifier across all decision thresholds.

BMI — Body Mass Index; body mass divided by the square of height (kg m^{-2}).

CAD — Coronary Artery Disease; a subtype of cardiovascular disease caused by atherosclerotic narrowing of the coronary arteries.

CI — Confidence Interval; a statistical range within which the true population parameter is expected to lie with a specified probability.

COPD — Chronic Obstructive Pulmonary Disease; a progressive lung disorder that may coexist with sleep apnoea, contributing to nocturnal hypoxaemia and increased cardiovascular risk.

CPAP — Continuous Positive Airway Pressure; first-line therapy for moderate-to-severe OSA that maintains airway patency during sleep.

CSA — Central Sleep Apnoea; sleep-related breathing disorder characterised by cessation of airflow due to lack of respiratory effort, distinct from OSA.

CVD — Cardiovascular Disease; encompasses coronary artery disease, cerebrovascular disease, peripheral artery disease, and aortic atherosclerosis.

DesSev — Desaturation Severity; A desaturation area–based algorithms that automatically calculates the area of oxygen desaturation events.

ECG — Electrocardiography; records electrical activity of the heart to detect arrhythmia and evaluate heart rate variability.

EEG — Electroencephalography; measures cortical electrical activity for sleep staging and arousal detection.

EMG — Electromyography; measures skeletal muscle activity, used for REM stage identification and limb-movement detection.

EOG — Electro-oculography; records eye movements used to determine sleep stages, particularly REM.

HB — Hypoxic Burden; a desaturation area–based parameter that quantifies the cumulative area of desaturation events, associated with manually scored respiratory events.

HL — Hypoxic Load; a desaturation area-based parameter that is independent of any respiratory and desaturation events.

HR — Hazard Ratio; ratio of hazards between two groups in the Cox proportional hazards model, quantifying relative risk of an event.

HRV — Heart Rate Variability; beat-to-beat variation in heart rate reflecting autonomic nervous system balance.

LDA — Linear Discriminant Analysis; classical supervised learning algorithm that projects data to maximise class separability.

MACEs — Major Adverse Cardiovascular Events, typically including cardiovascular death, nonfatal myocardial infarction, and stroke.

MinSat — Minimum Oxygen Saturation during sleep; the lowest oxygen saturation value recorded, used as an indicator of nocturnal hypoxaemia severity.

MrOS — Osteoporotic Fractures in Men Study; a large-scale longitudinal cohort used in sleep and cardiovascular outcome research.

OAm — Oral Appliance Mandibular Advancement Device; mechanical treatment that repositions the jaw forward to reduce airway collapse.

ODI — Oxygen Desaturation Index; number of desaturation events (≥ 3 % or 4 %) per hour of sleep, reflecting intermittent hypoxia.

OSA — Obstructive Sleep Apnoea; disorder characterised by repetitive upper-airway collapse during sleep, leading to intermittent oxygen desaturation and arousals.

PAD — Peripheral Artery Disease; atherosclerotic obstruction of peripheral arteries, particularly in the lower limbs.

PSG — Polysomnography; comprehensive overnight recording of multiple physiological signals (EEG, EOG, EMG, ECG, airflow, thoracoabdominal movements, and oximetry) to diagnose sleep disorders.

REDTA — Respiratory Event Desaturation Transient Area; a desaturation area–based algorithm that quantifies the area of manually scored respiratory events, expressed in %·hours.

REM — Rapid Eye Movement; sleep stage characterised by vivid dreaming, rapid eye movements, and muscle atonia.

RF — Random Forest; ensemble tree-based algorithm that aggregates multiple decision trees trained on bootstrapped samples to improve predictive accuracy and reduce overfitting.

ROC — Receiver Operating Characteristic; curve plotting the true-positive rate against the false-positive rate to assess a model's ability to discriminate between classes.

SDB — Sleep-Disordered Breathing; umbrella term encompassing OSA, CSA, and related sleep-related respiratory disturbances.

SHAP — SHapley Additive exPlanations; model-agnostic interpretability framework based on cooperative game theory that quantifies the contribution of each feature to individual predictions.

SHHS — Sleep Heart Health Study; multicentre cohort providing PSG and cardiovascular outcome data, used for all analysis in this thesis.

SpO2 — Peripheral Capillary Oxygen Saturation; proportion of oxygen-saturated haemoglobin measured by pulse oximetry during sleep.

SRC — Sleep Reading Centre; coordinating laboratory responsible for respiratory event scoring in the SHHS.

SVM — Support Vector Machine; supervised learning algorithm that separates classes using maximum-margin hyperplanes and kernel functions.

T90 — The total sleep time with oxygen saturation < 90 %; quantifies cumulative hypoxemic burden.

The Cox model — Cox Proportional Hazards Model; a semi-parametric regression framework that estimates the effect of covariates on event risk over time under the proportional-hazards assumption.

TST — Total Sleep Time; cumulative duration of all scored sleep epochs during a PSG recording.

XGBoost — Extreme Gradient Boosting; scalable, regularised gradient-boosted decision-tree algorithm used for outcome classification.

List of Figures

List of Tables

Chapter 1
Introduction

1 Introduction

This thesis focuses on predicting cardiovascular disease (CVD) mortality outcomes using polysomnogram (PSG)-derived parameters. Two experiments were conducted alongside a comprehensive literature review of published oximetry-derived parameters. The thesis aims to analyse the use of PSG-derived parameters in predicting CVD mortality outcomes, refine existing algorithms, and propose a robust and flexible model for individual-level CVD outcome prediction suited to population screening. This chapter introduces the background of the thesis and outlines the motivation for the research, objectives, and its contributions and significance. The thesis structure and list of publications are also presented in this chapter.

1.1 Motivation

CVD remains the leading cause of mortality worldwide, accounting for nearly one-third of all deaths despite significant advances in prevention and treatment [1, 2]. In Australia alone, CVD accounted for more than 45,000 deaths in 2022 [1]. Parallel to this, obstructive sleep apnoea (OSA) is highly prevalent, affecting an estimated one billion adults globally and 3 million adults in Australia, yet remains substantially underdiagnosed and undertreated [3]. Mounting evidence suggests that OSA is not merely a comorbidity but an independent risk factor for adverse cardiovascular outcomes, mediated by mechanisms such as intermittent hypoxia, systemic inflammation, and vascular dysfunction [4-6].

OSA and CVD manifest very differently across the lifespan. OSA is often symptomatic and can be recognised through signs such as loud snoring, nocturnal arousals, and excessive daytime sleepiness, which frequently prompt medical attention. In contrast, CVD typically remains clinically silent for many years, with individuals often asymptomatic until experiencing a severe or fatal event later in life [7]. This asymmetry highlights a clinical opportunity: if OSA is diagnosed earlier and leveraged as a window into cardiovascular risk, then predicting CVD outcomes becomes feasible at a stage when preventive strategies can still alter long-term trajectories. Ideally, the early recognition of OSA combined with accurate prediction of future CVD mortality outcome could enable timely intervention, improve quality of life, and ultimately save lives.

PSG, the gold standard for OSA assessment, records a comprehensive set of physiological signals, including respiratory airflow, oxygen saturation (SpO2), electrocardiography (ECG),

electroencephalography (EEG), electromyography (EMG), and electro-oculography (EOG) [8-10]. These signals provide a comprehensive characterisation of sleep physiology and offer a valuable resource for understanding the pathways linking OSA to CVD. However, clinical practice continues to rely predominantly on the apnoea–hypopnoea index (AHI) for diagnosing and stratifying the severity of OSA. Although AHI quantifies the number of apnoea and hypopnoea events per hour, it has limited prognostic ability because it fails to capture the duration of respiratory events, the burden of hypoxia, or sleep fragmentation, and it has repeatedly been shown to be a poor independent predictor of cardiovascular outcomes [11-13]. This limitation underscores the need for alternative PSG-derived parameters that more comprehensively reflect nocturnal hypoxaemia and its cardiovascular consequences.

Over the last decade, oximetry-derived measures such as the oxygen desaturation index (ODI), the total sleep time spent below 90% oxygen saturation (T90), and, more recently, desaturation area–based parameters have been proposed as complementary or alternative predictors [6, 14]. These parameters better quantify both the depth and duration of desaturation events, thereby capturing the overall hypoxic burden more accurately. Emerging evidence indicates that they outperform AHI in predicting cardiovascular outcomes [12, 15]. However, computational discrepancies by varying American Academy of Sleep Medicine (AASM) criteria, inconsistent interpretations of parameter definitions, and computational approaches, have fragmented the field and limited cross-study comparability [16, 17]. To date, no unified evaluation has systematically compared these algorithms in the same population and computational framework. Experiment 1 of this thesis specifically addresses this gap by focusing on desaturation area–based measures, with the aim of identifying the most robust and clinically useful method for predicting CVD mortality in large population settings.

While PSG provides a rich source of multidimensional information, its potential for CVD outcome prediction remains underutilised. Existing studies have largely relied on single-parameter evaluations within conventional proportional hazards models, reporting relative effects such as hazard ratios [18]. While valuable for time-to-event analysis, this focus may overlook the benefits of multiparameter prediction, as cardiovascular risk arises from the combined influence of multiple interacting factors [19]. In addition, hazard ratios reflect relative rather than absolute risk, which can be difficult to interpret for non-statistical end-users, including patients [20]. These limitations restrict their application to personalised cardiovascular risk stratification and form the rationale for Experiment 2 in this thesis.

Machine learning approaches, by contrast, can accommodate multidimensional PSG data, capture complex nonlinear feature interactions, and generate individual-level predictions of CVD outcomes [21-26]. Importantly, explainable machine learning methods offer interpretable insights into feature contributions, addressing the "black box" concerns associated with conventional machine learning and supporting clinical translation [18]. Despite this promise, most prior studies either restricted their focus to features that demand substantial clinical inputs and specialist annotations or employed complex models such as deep learning. Such requirements hinder their applicability in medically underserved settings, where access to specialist equipment and expertise is limited. Consequently, the benefits of predictive modelling are often restricted to well-resourced populations, limiting broader translational impact.

Against this background, the motivation of this thesis is threefold: first, to clarify and compare desaturation area–based algorithms and identify the best-performing method for CVD prediction; second, to investigate whether PSG-derived parameters (particularly oximetry-derived parameters), when assessed collectively, can enhance predictive performance; and third, to establish explainable machine learning methods that can provide robust, individualised CVD outcome predictions, with the ultimate goal of enabling scalable and clinically actionable risk stratification for both population-level screening and patient-level decision-making. The detailed rationale for these aims is presented in Chapters 4 and 5.

1.2 Objective

The overarching objective of this thesis is to investigate the role of PSG-derived parameters, with a particular focus on oximetry-derived parameters, in predicting CVD mortality, and to establish explainable machine learning approaches for robust, individual-level outcome prediction. This objective is pursued through four interlinked aims:

1.2.1 Objective 1: Review the current oximetry-derived parameters

To critically review published oximetry-derived parameters for predicting CVD outcomes, with the aim of demonstrating the limitations of relying on AHI alone. This objective also seeks to highlight methodological inconsistencies in the computation of desaturation area–based metrics.

1.2.2 Objective 2: Systematically evaluate desaturation area–based algorithms

To conduct the first direct comparison of three major desaturation area–based methods (Hypoxic Burden (HB), respiratory event desaturation transient area (REDTA), and desaturation severity (DesSev)) within the same patient population (the Sleep Heart Health Study). This objective aims to examine how differences in event definitions, sampling windows, and baseline choices influence computational behaviour and predictive performance, and to identify the most robust approach for population-based CVD prediction.

1.2.3 Objective 3: Investigate the predictive value of oximetry-derived parameters

To assess whether established measures such as ODI3 and T90, together with desaturation area–based parameters, can predict short-term CVD mortality individually and in combination. This objective further aims to determine whether multiparameter integration improves predictive accuracy beyond single metrics, and to inform feature selection for subsequent model development.

1.2.4 Objective 4: Develop an explainable machine learning framework.

To design and evaluate interpretable predictive models that integrate PSG-derived parameters, demographics, lifestyle factors, and medical history for predicting CVD mortality at the individual level. The framework aims to achieve a balance between predictive accuracy and interpretability, with minimal reliance on specialised clinical inputs, thereby improving scalability for broader populations. In addition, this objective explores the capacity of models to generalise across different time horizons, enabling flexible CVD outcome prediction beyond the short-term window.

Through these objectives, the thesis aims to make both methodological and applied contributions. Methodologically, it seeks to clarify the computational foundations of desaturation area–based metrics, address key discrepancies in their implementation, and identify the most robust approach for predicting CVD mortality in large population datasets. Practically, it aims to propose a pathway for integrating sleep study data into CVD outcome prediction that is scalable, clinically interpretable, and capable of producing reliable individual-

level estimates for broader populations, while maintaining flexibility across different time horizons.

1.3 Thesis contribution

This thesis contributes to both methodological development and applied prediction of CVD mortality using PSG-derived parameters. The contributions are distributed across three main chapters. Chapter 3 reviews and consolidates existing evidence on PSG-derived parameters for predicting CVD mortality, with particular emphasis on oximetry-derived measures, and identifies methodological discrepancies, thereby providing the rationale and methodological foundation for subsequent investigations. Chapter 4 systematically compares desaturation area–based algorithms, clarifies the impact of varying computational approaches on predictive performance, and identifies the most robust method for CVD mortality prediction in a large population setting. Chapter 5 contributes in two ways: first, by evaluating the combined predictive utility of oximetry-derived parameters, and second, by developing an explainable machine learning framework that achieves strong predictive performance while minimising reliance on specialised medical expertise. The detailed contributions are summarised below.

1.3.1 Chapter 3: Clarification of current limitations and methodological discrepancies.

This chapter consolidates prior evidence on PSG-derived parameters and demonstrates that conventional diagnostic metrics such as the AHI are insufficient for predicting CVD outcomes [11, 12, 27]. It highlights methodological discrepancies in the computation of oximetry-derived measures, particularly desaturation area–based parameters, where differences in event definition, baseline determination, and sampling windows have led to inconsistencies across studies [12, 15, 28, 29]. By identifying these limitations, the thesis establishes the rationale for systematic comparison.

1.3.2 Chapter 4: Systematic comparison of desaturation area–based algorithms

A central methodological contribution is the first direct comparison of three widely used desaturation area–based algorithms (HB, REDTA, and DesSev) implemented under a unified computational framework within the Sleep Heart Health Study. This analysis demonstrates how computational differences influence both parameter values and predictive performance, thereby clarifying why results have varied across studies. The chapter refines algorithm

implementation, identifies the best-performing and robust method for predicting CVD mortality, and shows that predictive outcomes are affected by variations in computational choices.

1.3.3 Chapter 5: Development of machine learning model for predicting CVD mortality outcome at individual-level

1.3.3.1 Phase 1: Evaluation of the predictive value of oximetry-derived parameters

The thesis evaluates whether established metrics such as ODI3 and T90, together with desaturation area–based parameters, can predict CVD mortality outcomes. It demonstrates that while individual parameters have predictive value, integrating multiple oximetry-derived parameters achieves superior performance compared with single metrics. This provides empirical evidence supporting the combined use of these measures and establishes the basis for feature selection in subsequent modelling.

1.3.3.2 Phase 2: Development of an explainable and scalable machine learning framework with minimal clinical reliance

Building on these findings, this phase of experiment develops and evaluates explainable machine learning models that integrate PSG-derived parameters with demographic, lifestyle, and medical information. Unlike traditional Cox regression, which is restricted by linear assumptions and relative hazard ratios, the framework accommodates nonlinear interactions and generates individual-level predictions while maintaining interpretability [18, 21-26].

Moreover, this experiment demonstrates that strong predictive performance can be achieved with reduced reliance on specialised clinical inputs. By focusing on PSG-derived parameters and general patient information, the framework is implemented in a way that is scalable to large population cohorts and adaptable across diverse healthcare settings. It also shows that the proposed predictive model generalises across different time horizons, maintaining stable performance rather than being limited to a single follow-up period.

1.4 Significance of research

This thesis makes significant contributions to both the methodological development and the applied use of PSG-derived parameters in predicting CVD mortality.

From a methodological perspective, the thesis delivers the first systematic comparison of desaturation area–based algorithms within the same cohort. Previous studies have relied on heterogeneous implementations, leading to inconsistent results and limited comparability. By unifying the computational framework and directly comparing HB, REDTA, and DesSev, this work demonstrates how discrepancies in event definition, baseline determination, and sampling windows influence parameter values and predictive performance. This systematic evaluation resolves a long-standing gap in the literature, clarifies the consequences of methodological variation, and provides insights for future algorithm development. It also identifies a reliable and best-performing desaturation area–based method for predicting CVD mortality, thereby advancing the methodological rigour of OSA–CVD analysis.

From an applied perspective, the thesis demonstrates the predictive value of combining multiple PSG-derived parameters, moving beyond single-metric approaches such as the AHI, ODI3, or T90. It shows that integrated measures deliver more accurate and stable predictions, offering a more comprehensive characterisation of nocturnal hypoxaemia and its cardiovascular consequences. These findings support the adoption of multivariable approaches in OSA-CVD analysis, with direct implications for both research design and clinical practice.

In addition, the thesis develops an explainable machine learning framework that integrates PSG-derived parameters with demographic, lifestyle, and medical information. The framework addresses the limitations of traditional survival models by accommodating nonlinear feature interactions, generating individual-level predictions, and maintaining interpretability through SHapley Additive exPlanations (SHAP) analysis. It achieves predictive performance comparable to more resource-intensive approaches while requiring substantially fewer specialised medical inputs. This reduced dependence on expert annotation and complex equipment enhances scalability to large populations and adaptability across healthcare settings. Furthermore, the framework demonstrates generalisability across multiple prediction horizons (3-, 5-, and 10-year), in contrast to most existing models that are restricted to a single follow-up period. By relying on accessible inputs such as oximetry and general health information, the framework also supports potential integration into home-based monitoring or online platforms, extending predictive modelling beyond specialist clinical environments.

Overall, the significance of this research lies in its dual impact: advancing methodological clarity in the computation and application of oximetry-derived parameters, and proposing a

clinically relevant, explainable machine learning framework that delivers robust, individual-level predictions of CVD mortality with minimal medical resources required. Together, these advances provide a pathway for scalable and actionable cardiovascular risk stratification based on PSG data, highlighting the potential of sleep studies to improve early CVD outcome prediction and reduce the long-term health burden through timely intervention.

1.5 Thesis structure

This thesis is organised into six chapters, each addressing a distinct component of the research while contributing to the overarching objective of investigating the role of PSG-derived parameters, particularly oximetry-derived measures, in predicting CVD mortality through explainable machine learning frameworks.

Chapter 1 outlines the motivation, objectives, contributions, and significance of the thesis. It serves as an overview of the thesis and provides a guide for the reader. The structure of the thesis and the list of publications are also presented.

Chapter 2 reviews the epidemiological and pathophysiological associations between OSA and CVD, outlining the mechanisms that link the two conditions. It introduces PSG as the gold standard for OSA diagnosis and describes the physiological signals it records. The chapter highlights the potential role of sleep measurements in strengthening CVD prediction and improving risk stratification. It also summarises prior evidence on the limitations of the AHI for CVD outcome prediction and introduces alternative PSG-derived parameters that may provide greater prognostic value. By emphasising both the clinical importance of OSA physiology and the inadequacy of existing diagnostic metrics, Chapter 2 establishes the clinical background and provides the rationale for the methodological investigations pursued in the subsequent chapters.

Chapter 3 addresses Objective 1 by reviewing published PSG-derived parameters, primarily oximetry-derived, in relation to CVD outcome prediction. It discusses their methodologies and predictive performance, identifies methodological discrepancies, and demonstrates how such inconsistencies limit comparability across studies. These findings establish the methodological rationale and background for the systematic evaluation presented in Chapter 4. In addition, Chapter 3 introduces the statistical tool (the Cox proportional hazards analysis) used in Chapter

4 and the machine learning techniques and performance evaluation methods applied in Chapter 5.

Chapter 4 addresses Objective 2 by conducting the first systematic comparison of three major desaturation area–based methods within the same patient population. Each algorithm is implemented under a unified computational framework to ensure reproducibility and comparability. The chapter analyses how variations in algorithm design affect both computational behaviour and predictive performance for CVD mortality, and identifies the most robust and best-performing method for CVD prediction.

Chapter 5 addresses Objectives 3 and 4. In Phase 1, it evaluates the predictive value of established oximetry-derived parameters (ODI3, T90, and desaturation area–based measures) for short-term (3-year) CVD mortality, and investigates whether combining parameters improves predictive performance compared with individual metrics. The results from this stage inform feature selection for model development. In Phase 2 and its extension, the chapter develops and evaluates explainable machine learning models that integrate PSG-derived parameters, demographics, lifestyle factors, and medical history to provide individual-level outcome predictions. The framework emphasises predictive accuracy, interpretability, scalability, minimal reliance on specialised clinical inputs, and the ability to generalise across different time horizons (3-, 5-, and 10-year).

Chapter 6 concludes the thesis. It synthesises the findings from the preceding chapters, highlights the contributions, and discusses directions for future development in addressing the limitations identified in this work.

1.6 List of refereed publications

1.6.1 Journal publications

1. S. He, P. A. Cistulli and P. de Chazal, “Comparison of Oxygen Desaturation Area-Based Methods in Predicting Cardiovascular Disease Mortality Outcomes,” *Eur. Respir. J.*, (under review), 2025.
2. S. He, P. A. Cistulli and P. de Chazal, “Using PSG-Derived Parameters and Explainable Machine Learning Approaches to Predict CVD Mortality,” (in preparation), 2025.

3. S. He, P. A. Cistulli and P. de Chazal, "A Review of Novel Oximetry Parameters for the Prediction of Cardiovascular Disease in Obstructive Sleep Apnoea," *Diagn.*, vol. 13, no. 21, p. 3323, 2023, doi: https://doi.org/10.3390/diagnostics13213323.

1.6.2 Conference publications

4. S. He and P. de Chazal, "Can Oximetry-Derived Parameters Predict 3-Year Cardiovascular Mortality?" *Proc. 47th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Copenhagen, Demark, (accepted), 2025.
5. S. He and P. de Chazal, "The Influence of Arousal Events on Desaturation Area-Based Parameters for Predicting 3-Year Cardiovascular Mortality," *Proc. 47th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Copenhagen, Demark, (accepted), 2025.
6. S. He, P. A. Cistulli and P. de Chazal, "Comparison of Oximetry Event Desaturation Transient Area-Based Methods in Predicting Cardiovascular Disease Mortality Outcomes," *Proc. 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Orlando, FL, USA, pp. 1–4, 2024, doi: https://doi.org/10.1109/EMBC53108.2024.10782779.
7. S. He, K. Cook, K. Sutherland, Y. S. Bin, P. A. Cistulli and P. de Chazal, "A Comparison of Hypoxic Burden Algorithms Using Three Different Methods for Calculating Baseline Oxygen Saturation for Predicting Cardiovascular Death in the Sleep Heart Health Study," *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sydney, Australia, pp. 1–4, 2023, doi: https://doi.org/10.1109/EMBC40787.2023.10340410.
8. P. de Chazal, K. Sutherland, K. Cook, Y. S. Bin, S. He and P. A. Cistulli, "A Comparison of Cardiovascular Disease Associations of Time-Domain Oximetry Parameters in Sleep Apnoea Cases from the Sleep Heart Health Study," *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sydney, Australia, pp. 1–4, 2023, doi: https://doi.org/10.1109/EMBC40787.2023.10340541.

1.6.3 Conference abstract and presentations

9. S. He, P. A. Cistulli and P. de Chazal, "Comparison of cardiovascular disease mortality outcomes using oximetry event desaturation transient area-based methods," *J Clin Sleep Medicine*, World Sleep 2025, (abstract accepted for publication), 2025.
10. S. He, P. de Chazal and P. A. Cistulli, "O035 Refining the Desaturation Area-Based Parameters by Investigating Different Baseline Methods," *Sleep Adv.*, vol. 5, suppl. 1, p. A13, 2024.

11. S. He, Y. Bin, P. A. Cistulli and P. de Chazal, “O008 Refining the Hypoxic Burden Algorithm by Investigating Different Methods for Calculating the SpO2 Baseline,” *Sleep Adv.*, vol. 4, suppl. 1, p. A3, 2023.
12. B. K. Y. Tong, G. M. Stewart, S. He, P. de Chazal and P. A. Cistulli, “Pulse Wave Amplitude Drops Characteristics During Sleep in Acute Coronary Syndrome Patients with Obstructive Sleep Apnoea,” *Proc. Am. Thorac. Soc. Conf.*, A110. New Frontiers in Sleep Apnoea, 2024.
13. B. Tong, S. McClintock, S. He, P. de Chazal, B. Yee and P. A. Cistulli, “P038 Characterising Pulse Wave Amplitude Drops in Patients with Acute Coronary Syndrome,” *Sleep Adv.*, vol. 4, suppl. 1, p. A48, 2023.

Chapter 2
Clinical Background

2 Clinical background

This chapter provides key clinical information on CVD and OSA as relevant to this thesis, detailing their physiological and epidemiological characteristics. It further discusses the risk factors associated with CVD, along with the diagnosis and treatment of OSA, to elucidate the potential relationship between two diseases and highlight the challenges in current clinical practice. By establishing a strong clinical foundation, this chapter aims to enhance the reader's understanding of the context in which this research is situated and to underscore the critical need for the research presented in this thesis.

2.1 Cardiovascular disease

2.1.1 Epidemiology and Pathophysiology

CVD, also referred to as heart disease, remains the leading cause of mortality worldwide. According to the World Health Organization, CVD accounts for approximately 17.9 million deaths annually, representing 32% of global mortality [2]. In Australia, although the CVD-related mortality rate has declined over the years, reaching an all-time low, the condition still accounted for 45,000 deaths in 2022, corresponding to a rate of 173 deaths per 100,000 individuals. Furthermore, the 2022 National Health Survey conducted by the Australian Bureau of Statistics estimated that 1.3 million Australians aged 18 and over were living with one or more CVD-related conditions [1].

CVD encompasses four major diseases/conditions: coronary artery disease (CAD), cerebrovascular disease, peripheral artery disease (PAD), and aortic atherosclerosis [30]. CAD, which accounts for approximately one-third to one-half of all CVD cases, arises from reduced myocardial perfusion, leading to angina, myocardial infarction, and heart failure. Traditionally regarded as a cholesterol storage disease, CAD is now increasingly recognised as an inflammatory disorder [31]. Inflammation plays a critical role in all stages of atherogenesis, contributing to the formation of atherosclerotic plaques that obstruct the coronary artery lumen [32].

Cerebrovascular disease, commonly referred to as stroke, is a condition characterised by an abnormality in the brain's blood supply. Stroke can be classified into two main types: ischemic stroke, which results from a vascular blockage and accounts for approximately 85% of cases,

and haemorrhagic stroke, which occurs due to bleeding and comprises the remaining 15% [33]. Both types can lead to significant neurological deficits. While hypertension is often regarded as the primary cause of ischemic stroke, other factors such as clotting disorders, carotid dissection, and illicit drug use are more prevalent among younger populations [34]. Over the past decades, stroke has emerged as a leading cause of adult disability [35].

PAD is an arterial disorder affecting the limbs, predominantly observed in elderly men. Atherosclerosis is the primary cause of approximately 90% of PAD cases, with additional contributions from inflammation and lipid accumulation [7, 36]. The gradual build-up of atherosclerotic plaques in the arteries of the legs and arms often remains asymptomatic for years, typically manifesting in later life. Intermittent claudication, characterised by muscle pain in the legs during exertion, is the most common and earliest symptom of PAD and is often considered a key diagnostic indicator [7]. The severity of PAD is determined by the extent of atherosclerotic plaque formation and the diameter of the affected blood vessels, resulting in varying degrees of symptomatic presentation among individuals [36].

Aortic atherosclerosis, which includes thoracic and abdominal aortic aneurysms (AAA), is characterised by the dilation of the aorta, the major artery responsible for carrying blood from the heart to the abdomen. AAA has become a significant health concern, with its incidence increasing over the past decades and being more prevalent in men than in women [37]. Similar to PAD, AAA typically remains asymptomatic until a critical rupture occurs. Although the precise underlying mechanisms are not fully understood, AAA is associated with the degradation of the elastic media of the atheromatous aorta. Notably, smoking has been identified as a major contributing factor in the development of aneurysms [38].

In summary, CVD is the leading cause of death worldwide and imposes a significant health burden on the ageing population. The underlying pathophysiology of CVD events is predominantly driven by atherosclerosis. However, other mechanisms, including thrombosis, hypertension, arrhythmias, inflammation, and structural abnormalities, also play critical roles in disease progression [32, 39-42].

2.1.2 Traditional risk factors

Despite differences in pathophysiology, all cardiovascular diseases share common risk factors, which can be categorised into two groups: non-modifiable and modifiable. Non-modifiable risk factors cannot be controlled or altered by external factors, whereas modifiable risk factors can be mitigated through lifestyle modifications or behavioural changes. In general, possessing one or more risk factors increases the likelihood of experiencing CVD events; however, managing these risk factors does not eliminate the possibility of developing CVD [43, 44].

Non-modifiable risk factors refer to those that individuals are born with or cannot be controlled through lifestyle changes. Age, for example, is widely recognised as an independent risk factor for CVD events. Studies show a strong association between increasing age (particularly over 65 years), and stroke, with risks increasing incrementally per decade [45]. Gender is another factor that exhibits distinct differences between groups. Historically, CVD has been considered a predominantly male disease; however, the risk of CVD in women has often been underestimated [46]. While both men and women share a similar overall risk of developing CVD, studies have shown that women are more likely to have diabetes and tend to develop CVD at older ages than men [46, 47]. In contrast, PAD and AAA are generally more prevalent in men than in women [7, 38]. Similar differences are also observed across ethnicities, with studies suggesting that individuals of South Asian and African descent have a higher risk of developing CVD compared to other populations [48, 49]. Another critical non-modifiable risk factor is family medical history, which plays a significant role in determining an individual's susceptibility to CVD. Studies suggested that a family history of CVD increases risk among first-degree relatives, with a 40% increased risk among siblings and 60%-75% increased risks among offspring [50].

Modifiable risk factors, on the other hand, can be controlled and altered through lifestyle changes. These include body mass index (BMI), hypertension, diabetes, and various lifestyle habits. BMI, which measures weight in relation to height, is strongly correlated with CVD events. Notably, higher BMI has the strongest association with incident heart failure among all CVD subtypes [51]. Overweight and obese women exhibit progressively higher risk of CVD mortality, with hazard ratios reported as high as 4.71 in severe obesity categories [52]. Hypertension and diabetes are established predictors of CVD and are closely interrelated. Hypertension is found to be twice as prevalent in patients with diabetes compared to those

without, while diabetes develops more rapidly in hypertensive individuals than in those with normal blood pressure [53]. Fundamental lifestyle habits such as dietary choices, smoking status, alcohol consumption, and physical inactivity are key modifiable contributors to CVD risk. Epidemiological evidence indicates that even low-intensity smoking is associated with approximately a 50% increased risk of CVD relative to never smoking, and the combination of smoking with physical inactivity confers greater than a two-fold increased risk of CVD mortality. Conversely, adherence to a constellation of healthy lifestyle behaviours, including a healthy diet, regular physical activity, non-smoking, and moderate alcohol intake, has been associated with up to a ~66% lower risk of cardiovascular disease in prospective cohort analyses. [54-57].

2.2 Obstructive sleep apnoea

OSA is a prevalent sleep disorder caused by recurrent upper airway collapse during sleep. This obstruction leads to apnoea (complete airway blockage), hypopnoea (partial airway obstruction), or a combination of both, resulting in intermittent hypoxia, hypercapnia, cortical arousal, and sleep fragmentation. Sleep fragmentation induces daytime sleepiness, negatively impacting quality of life and increasing the risk of motor vehicle accidents. Beyond its immediate effects, OSA significantly heightens the risk of CVD events due to OSA-induced endothelial dysfunction, coagulation-fibrinolysis imbalance, oxidative stress, elevated sympathetic activity, and systemic inflammation [4, 5]. The cumulative hypoxic events caused by airway collapse contribute to nocturnal hypoxia, which may lead to serious systemic impairments, including high cholesterol levels, hypertension, insulin resistance, and glucose dysregulation [6, 58]. These conditions are well-established CVD risk factors and present a significant health burden, particularly among older populations.

2.2.1 Pathophysiology

The respiratory system comprises the organs responsible for air conduction, filtration, and gas exchange and is divided into the upper and lower respiratory systems, as shown in **Figure 2.1**A. The upper respiratory system, which includes the nasal cavity, pharynx, and larynx, facilitates airflow during ventilation [59]. Among these structures, the pharynx plays a critical role in maintaining airway patency, particularly in the context of OSA. OSA-induced airway obstruction occurs in the pharynx, where the balance between pharyngeal muscle function and transmural pressure determines whether the airway remains open or collapses. In healthy

individuals, pharyngeal muscle activity decreases during sleep, yet the airway remains unobstructed. However, in those with compromised neural activation or abnormal pharyngeal anatomy, the muscles are unable to maintain airway patency, leading to recurrent airway collapse and manifesting as apnoea or hypopnoea [60].

As a result, airway obstruction in OSA occurs recurrently throughout sleep and may involve collapse at multiple anatomical sites. The velopharynx is the primary site of collapse in most patients but narrowing can also occur in the anterior (nasopharynx) or lower posterior regions (oropharynx and hypopharynx), as illustrated in **Figure 2.1**B [61, 62]. The pattern and severity of upper-airway obstruction influence airflow dynamics and oxygenation. Both apnoea and hypopnoea are characterised by airflow limitation and increased airway resistance, which can lead to intermittent oxygen desaturation and sleep fragmentation. The extent of desaturation varies across events depending on their duration, depth of airflow reduction, and baseline oxygen reserve [10]. In the short term, these disturbances trigger nocturnal arousals and fragmented sleep architecture, often resulting in next-day symptoms such as excessive daytime sleepiness. With chronic and repeated exposure over months to years, persistent sleep disruption and intermittent hypoxemic stress may contribute to longer-term clinical consequences, including sustained fatigue, impaired cognitive function, and increased cardiometabolic risk [63].

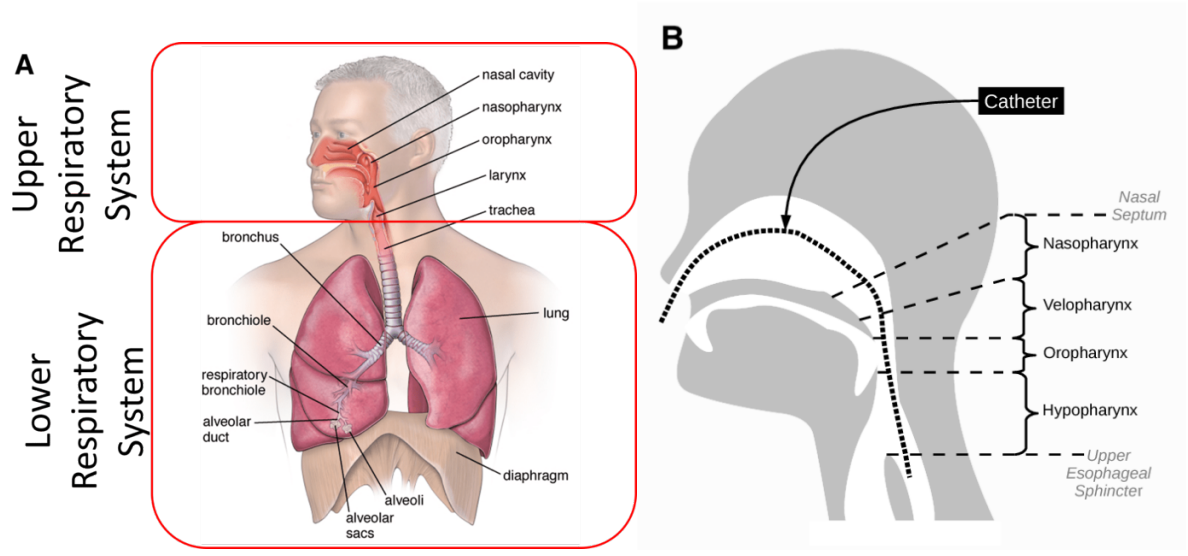


Figure 2.1 An overview of respiratory system. Figure 2.1A provides a summary of the upper and lower respiratory systems, while Figure 2.1B offers a detailed close-up view of the pharynx. (Figure 2.1A modified from [64], and Figure 2.1B modified from [63]).

Repetitive apnoea and hypopnoea events have significant impacts on ventilation, sleep stability, and overall health. With sleep onset, activity of the pharyngeal dilator muscles decreases, reducing upper-airway stability and increasing susceptibility to airway collapse. During inspiration, negative intraluminal pressure can then promote airway narrowing, thereby restricting airflow. Partial or complete obstruction results in reduced ventilation, leading to transient oxygen desaturation (hypoxaemia) and carbon dioxide retention (hypercapnia), which typically terminate with arousal and restoration of airflow. In response to these disturbances, the body triggers arousals, activating the pharyngeal dilator muscles and temporarily restoring airway patency. However, as the airway reopens, the carbon dioxide level decline and the oxygen level increase, allowing the patient to fall asleep again. This, in turn, leads to recurrent airway obstruction, perpetuating the cycle [60]. This process, as summarised in **Figure 2.2**, can occur hundreds of times per night without the patient's awareness, leading to intermittent hypoxia, cortical arousals, and fragmented sleep—hallmarks of OSA.

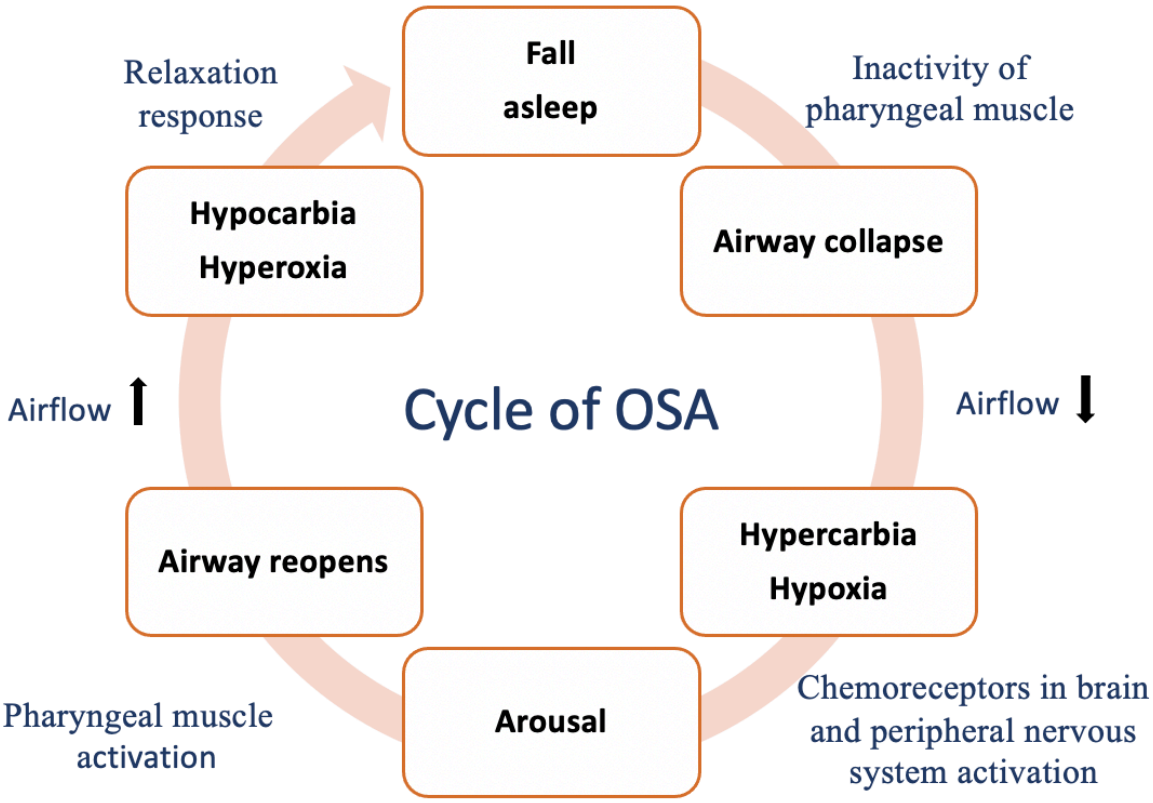


Figure 2.2 Schematic representation of the recurrent cycle of OSA. The factors outside the circle indicate the physiological causes that trigger each stage of the cycle.

2.2.2 Epidemiology

OSA is considered a significant contributor to healthcare burden. In general, OSA is more common in men than in women, particularly in individuals aged 50 years and older. Patients

with OSA typically have a higher average BMI compared to healthy populations, and studies have reported a positive, approximately linear association between BMI and OSA severity [65, 66]. In 2007, the World Health Organization estimated that OSA affected more than 100 million individuals; however, this figure likely underestimates the true prevalence, as it is based primarily on diagnosed and treated cases [67]. Studies have suggested that a substantial proportion of OSA cases remain undiagnosed or untreated, with variations between countries due to societal and economic factors. The diagnosis of OSA requires significant medical resources, and populations must have an adequate level of awareness and education about OSA before seeking consultation in sleep clinics. In developing countries, the diagnosis and treatment of OSA are often inaccessible to most of the population due to inadequate sleep laboratory infrastructure and limited investment in sleep research. Even in developed countries, high costs restrict widespread screening, resulting in many potential cases remaining undiagnosed [68].

Benjafield et al. conducted a comprehensive review of OSA prevalence across 16 countries, providing a reliable estimate of the total affected population worldwide. According to their study, approximately 1 billion individuals globally (29.16%) aged between 30 and 69 years have OSA, with 425 million (13.24%) experiencing moderate to severe forms of this disorder [3]. An estimated 24.5% of Australians aged between 30 and 69 years are affected by OSA, with approximately 4.8% meeting criteria for moderate-to-severe OSA [3]. The ten countries with the highest estimated OSA populations are presented in **Table 2.1**. In some countries, more than 50% of the population aged 30–69 years is affected by OSA, with nearly 40% having moderate to severe OSA. Notably, in France, an estimated 75% of adults in this age group are considered to have OSA, with approximately half experiencing moderate to severe symptoms. However, only 18.1% of French adults self-reported being at high risk for OSA, and merely 3.5% had received treatment for this disorder [69]. A similar pattern is observed worldwide. In the United States, approximately 8% of the population has been diagnosed with OSA, while the estimated prevalence exceeds 33% [70, 71]. According to Ip et al., 6.2% of the Chinese population has been diagnosed with OSA in Hong Kong, whereas nearly 24% are estimated to have the condition [3, 72, 73]. This disparity indicates a substantial number of undiagnosed OSA cases, highlighting significant potential health risks posed by untreated OSA patients.

Despite its high prevalence, OSA is frequently under-recognised and undertreated in clinical practice. Studies have shown that a large proportion of individuals with OSA remain

undiagnosed, which may lead to underestimation of disease burden and delay in management, especially in patients with mild disease where symptoms are less overt or classical presentations are absent [74, 75]. This diagnostic gap has important implications because even mild forms of OSA may be associated with adverse health outcomes, yet are often overlooked in both screening and routine care.

Table 2.1 Summary of ten countries with the highest estimated OSA populations aged 30-69 years, ranked from the highest to the lowest [3].

Country*	Total population aged 30-69 years, million	OSA population (OSA Prevalence), million (%)	Mild OSA** (OSA Prevalence), million (%)	Moderate to severe OSA **(OSA Prevalence), million (%)
Worldwide	3210	936 (29.16)	511 (15.92)	425 (13.24)
China	744	176 (23.66)	110 (14.78)	66 (8.87)
USA	163	54 (33.13)	30 (18.40)	24 (14.72)
Brazil	98	49 (50.00)	24 (24.49)	25 (25.51)
India	534	52 (9.74)	23 (4.31)	29 (5.43)
Pakistan	63	42 (66.67)	25 (39.68)	17 (26.98)
Russia	78	40 (51.28)	20 (25.64)	20 (25.64)
Nigeria	51	31 (60.78)	19 (37.25)	12 (23.53)
Germany	43	26 (60.47)	12 (27.91)	14 (32.56)
France	32	24 (75.00)	12 (37.50)	12 (37.50)
Japan	67	22 (32.84)	13 (19.40)	9 (13.43)

** The estimation model was developed using available data from 16 countries and was subsequently applied to estimate OSA prevalence in 193 countries. A detailed description of the estimation model can be found in the supplementary material provided by Benjafield et al. [3].

* The severity of OSA is categorised using the AHI, where an AHI ≥5 and <15 is classified as mild OSA, and an AHI ≥15 indicates moderate to severe OSA. The measurement methods and criteria for AHI will be explained in the following sections.

2.2.3 Diagnosis

Overnight PSG is considered the gold standard for diagnosing OSA. It measures a range of physiological signals, including EEG, ECG, EOG, pulse oximetry, EMG, oronasal airflow, body position, and respiratory effort (thoracic and abdominal movements), with sensor placements illustrated in **Figure 2.3** [9, 10, 76, 77]. PSG is conducted in accordance with the guidelines of the AASM to determine the AHI, the principal diagnostic metric for OSA.

The traditional type I PSG study requires overnight monitoring in a sleep laboratory, where patients sleep in an unfamiliar environment, such as a sleep clinic or hospital, with multiple sensors attached to their body. Consequently, PSG has several drawbacks: it is expensive, resource-intensive, and often inaccessible due to the limited availability of sleep laboratories. To address these challenges, current diagnostic strategies involve pre-screening patients with sleep disorder symptoms and referring only those with a high suspicion of OSA for PSG testing. The screening process typically includes validated questionnaires and self-reported surveys to identify at-risk individuals.

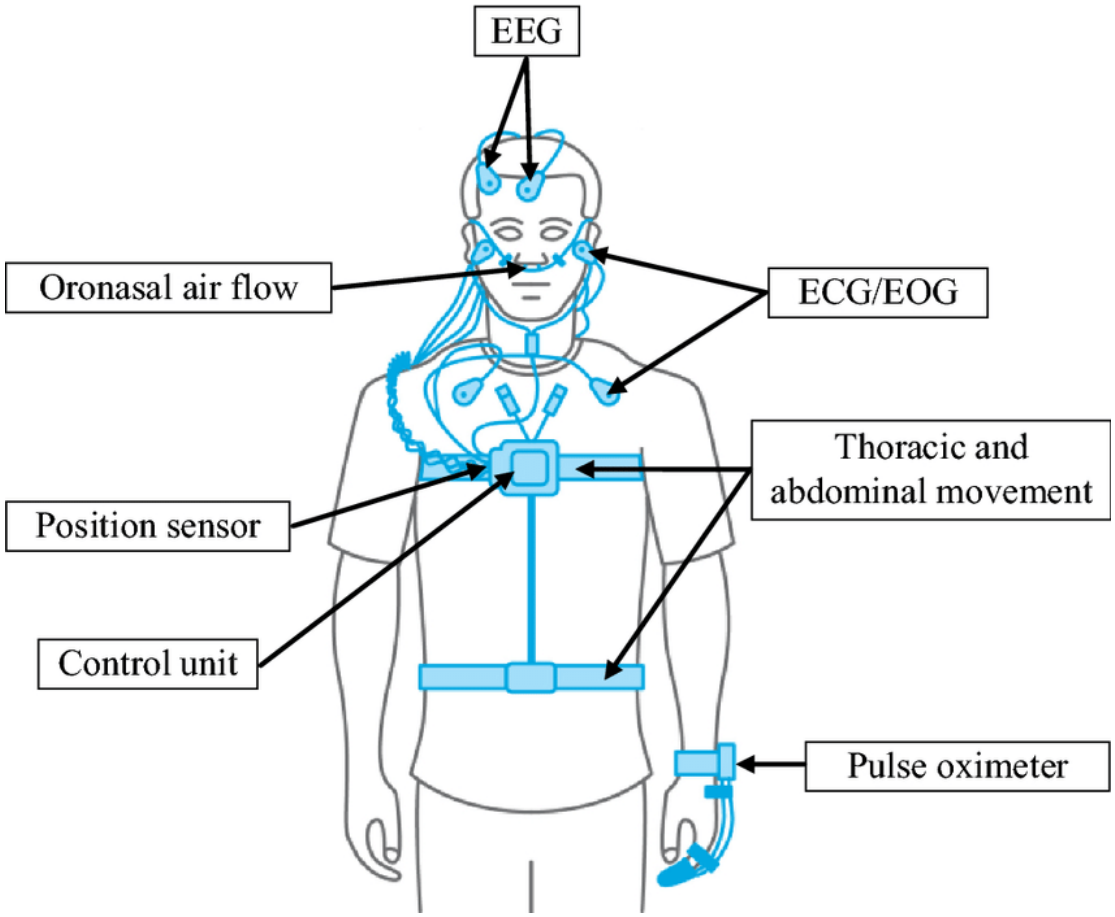


Figure 2.3 A graphic presentation of PSG wire connection. Figure adapted from [78] Fig 2.

2.2.3.1 Polysomnogram (PSG)

The standard diagnostic criterion for OSA is laboratory-based PSG, which monitors overnight sleep and respiratory patterns. PSG records several key physiological signals, including sleep stages, arousal events, apnoea and hypopnoea episodes, oxygen desaturation, and cardiac rhythm. A full Type I PSG study may also incorporate limb electromyography to detect periodic limb movements during sleep and, when clinically indicated, transcutaneous carbon dioxide monitoring [8]. To assess sleep stages, multiple physiological signals are involved. In a sleep study, at least three channels of EEG with ear references are required for sleep stage classification. Most sleep laboratories use four EEG channels, comprising two central and two occipital channels, referenced to the ears, to additionally measure sleep duration and detect arousal events. For patients with epilepsy, an extra EEG channel may be used to monitor abnormal epileptiform activity. EOG signals track horizontal and vertical eye movements, aiding in sleep stage classification. Typically, two EOG channels with four electrodes are used in OSA studies to capture the onset of rapid eye movement (REM) sleep and the presence of slow-rolling eye movements. EMG is recorded using at least three chin electrodes, with an additional two electrodes placed on the anterior tibialis muscles. EMG assists in sleep staging, identifying REM sleep without atonia, and detecting periodic limb movements [79].

ECG, recorded using a single modified lead II configuration, evaluates heart rate variability and detects arrhythmias. Apnoea and hypopnoea events are identified through oronasal airflow measurements, which are recorded using thermal airflow sensors and nasal pressure transducers. Different apnoea subtypes (obstructive, central, and mixed) are distinguished based on airflow patterns. Respiratory effort, assessed via thoracic and abdominal movement, helps differentiate between obstructive and central apnoea events. Pulse oximetry is another essential PSG measurement, focusing on detecting oxygen desaturation associated with respiratory events [79].

These physiological signals operate in concert to construct a comprehensive evaluation of a patient's sleep condition. Each recorded measurement offers unique insights into different aspects of sleep, yet they interact dynamically to reveal the intricate mechanisms underpinning sleep physiology [8]. **Figure 2.4**, as an example, illustrates the interplay between EEG, pulse oximetry, airflow, and respiratory effort in identifying obstructive apnoea events. In contemporary diagnostic practices and OSA analysis, the integration of data from multiple

signals enables a more holistic understanding of sleep disturbances. These measurements hold significant clinical value beyond the conventional reliance on the AHI as the key diagnostic metric, enriching range and depth of analysis.

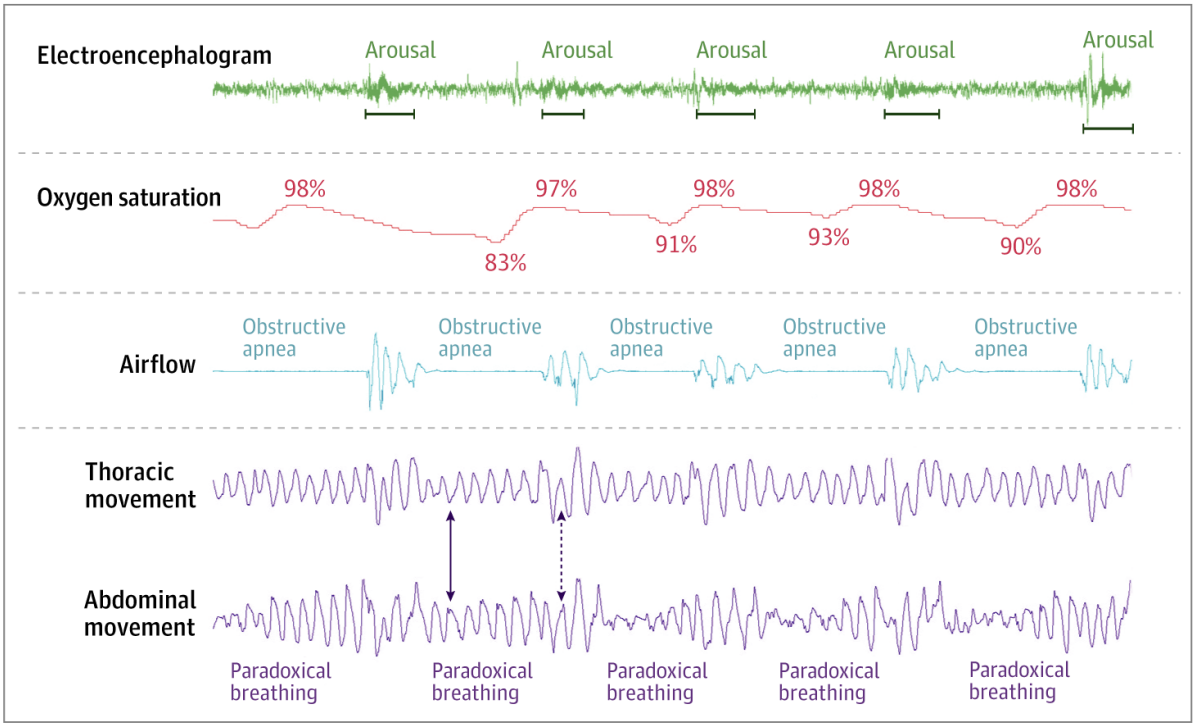


Figure 2.4 Example of a 5-minute PSG recording, demonstrating the correlation of each recorded signal with obstructive apnoea events. The absence of airflow is a key indicator of obstructive apnoea, with corresponding real-time thoracic and abdominal movements. This movement pattern, known as paradoxical breathing, occurs as a compensatory response to oxygen desaturation. The oxygen levels measured by pulse oximeter exhibit a slight delay due to the circulation time between lungs and fingertip. EEG is used to detect arousal events. The occurrence of arousal indicating airway reopening and may signal the transition into the next cycle of OSA. Figure adapted from [8], Figure 2.

2.2.3.2 American Academy of Sleep Medicine (AASM) Scoring Guideline

The AASM provides standardised clinical guidelines for scoring sleep and respiratory events, ensuring the consistency in sleep analysis [10]. The manual outlines the recommended sleep measurements, including sleep scoring data (such as sleep latency and total recording time), arousals, respiratory events (including different subtypes of apnoea and hypopnoea, as well as the oxygen saturation levels), cardiac responses, arrhythmias, and EEG/ECG abnormalities. These are considered essential measures, as they commonly occur during sleep and contribute to clinical decision-making. Additionally, the manual includes optional choices such as respiratory effort-related events, sleep hypnograms, and leg muscle movements, which, while insightful, are not always needed for routine sleep assessments.

Over time, the AASM manual has undergone multiple updates, refining criteria for scoring key sleep and respiratory events. A prominent example is the evolving definitions of apnoea and hypopnoea, as summarised in **Table 2.2**. In 1999, the initial guidelines lacked differentiation between apnoea and hypopnoea. The criteria at the time included an airway obstruction exceeding 50%, oxygen desaturation of ≥3%, and an event duration of at least 10 seconds [10]. The subsequent 2007 updates to the AASM manual redefined the apnoea as a complete airway blockage but introduced two versions (recommendation vs alternative) of the hypopnoea definition, differing in airflow reduction and oxygen desaturation thresholds [80]. These variations led to inconsistencies in scoring across studies, complicating cross-study comparisons and hindering the reproducibility of PSG-derived parameter algorithms, as different studies adopted different scoring criteria. Since 2012, a standardised definition of apnoea and hypopnoea has been established and has remained unchanged [16].

Table 2.2 Summary of updates of AASM criteria for apnoea and hypopnea from 1999 to 2012.

Year	Apnoea definition	Hypopnea definition
1999 Chicago criteria [10]	- Airflow drop > 50% from baseline* - Or a clear amplitude reduction of breathing with an oxygen desaturation ≥ 3% or an arousal - Or event duration ≥ 10 seconds	- Airflow drop > 50% from baseline* - Or a clear amplitude reduction of breathing with an oxygen desaturation ≥ 3% or an arousal-Or event duration ≥ 10 seconds
2007 Recommendation [80]	- Peak signal excursion** drops ≥ 90% of pre-event baseline for ≥ 10 seconds.	- Nasal pressure peak signal excursions drop **≥ 30%** of pre-event baseline for ≥ 10 seconds, and - Oxygen desaturation **≥ 4%** or event associated with an arousal
2007 Alternative [80]		- Nasal pressure signal drops **≥ 50%** of baseline for ≥ 10 seconds, and - Oxygen desaturation **≥ 3%** or event associated with an arousal
2012 Version 2.0 [16]	- Drop in peak thermal sensor excursion ≥ 90% of pre-event baseline for ≥ 10 seconds, and - If a shorter portion of a hypopnoea event meets apnoea criteria, the entire event is scored as apnoea	- Nasal pressure signal excursion** drops ≥ 30% of pre-event baseline, for ≥ 10 seconds, and - Oxygen desaturation ≥ 3% or event associated with an arousal

*Baseline is defined as the mean amplitude of stable breathing or the three largest breaths in the two minutes preceding event onset for patients without stable breathing.

**Peak signal excursion is measured using an oronasal thermal sensor, a positive airway pressure titration device, or alternative apnoea sensors.

2.2.3.3 Apnoea-Hypopnoea Index (AHI) and Obstructive Sleep Apnoea (OSA)

The AHI is considered the gold standard metric for OSA diagnosis and is commonly used in clinical practice to classify OSA severity. Based on the AASM criteria outlined in **Table 2.2**, AHI quantifies the frequency of apnoea and hypopnoea events per hour of sleep. An AHI of less than 5 is considered normal, indicating a healthy population. Mild OSA is defined as an AHI between 5 and 15, where sleepiness may occur during tasks requiring minimal attention, often making it difficult to realise at this stage. Moderate OSA, with an AHI between 15 and 30, is associated with worsening sleep quality, and sleepiness may become apparent during meetings or presentations. Severe OSA is diagnosed when AHI exceeds 30, requiring immediate medical attention. At this stage, symptoms are more pronounced, and excessive daytime sleepiness can occur even during active tasks such as talking or driving [81]. Although OSA is commonly categorised by severity using the AHI, exceptions exist whereby individuals with mild OSA may experience substantial daytime sleepiness or impaired sleep quality, whereas some patients with moderate or severe OSA may report minimal or no symptoms despite a high AHI [82, 83]

Notably, the event scoring process accounts for not only obstructive events but also central and mixed events. Central sleep apnoea (CSA), unlike OSA, is not caused by airway obstruction but rather by disrupted neural regulation of breathing cycles. CSA and OSA are distinguished using nasal airflow, respiratory effort, and pulse oximetry signals. OSA events are typically accompanied by thoracic movements, whereas CSA events lack this respiratory effort. Additionally, airflow patterns differ between the two conditions: OSA is characterised by a flattened inspiratory peak in nasal pressure recordings, whereas CSA displays rounded inspiratory peaks. Oxygen desaturation patterns also vary, with OSA exhibiting an asymmetrical decline and recovery, whereas CSA follows a clear sinusoidal curve. Furthermore, OSA cycles occur unpredictably, while CSA follows a consistent periodic pattern [84]. When a patient exhibits both OSA and CSA features during sleep, the condition is classified as mixed sleep apnoea. This condition begins with obstructive respiratory failure, followed by a brief period of CSA occurring seconds prior to the obstructive event, with thoracic and abdominal movements resuming thereafter [85].

2.2.3.4 Questionnaires and home-based monitoring

Despite the advantages of PSG-based diagnostic procedures, several limitations remain significant. Traditional PSG is expensive, resource-intensive, and often inaccessible due to the limited availability of sleep laboratories. The diagnostic process requires trained technicians and sleep specialists for both overnight monitoring and event scoring. Additionally, patients must sleep in a laboratory setting for a single night, an unfamiliar environment that may lead to atypical sleep patterns and potentially unrepresentative measurements.

To address these limitations, pre-procedure screening is incorporated into standard OSA diagnostic pathways to ensure that only patients with a high likelihood of OSA are referred for PSG. The diagnostic pathway and pre-procedure screening is detailed in section 2.2.4.5 and **Figure 2.7**. Common pre-screening tools for OSA include self-reported questionnaires such as the STOP-Bang questionnaire, the Sleep Apnoea Clinical Score, the Berlin Questionnaire, and the NoSAS score. While these assessments do not provide detailed insights into sleep physiology, they are sufficiently sensitive to exclude individuals who are unlikely to require PSG testing [86].

Home-based monitoring has emerged as an alternative for the diagnosis of OSA and became routine clinical practice for patients with high suspicion of OSA , enabling the measurement of key physiological parameters such as airflow, respiratory effort, and oxygen saturation. Unlike traditional PSG, home-based monitoring employs self-applied sensors without the supervision of sleep experts. Various studies have demonstrated that current at-home monitoring methods exhibit high sensitivity and specificity, with an area under the curve (AUC) exceeding 0.85 [87-89]. While home-based testing presents a viable alternative when PSG is unavailable, further improvements and validation are necessary, particularly for patients with a high pre-test probability of disease [90, 91].

2.2.4 Treatment

The treatment strategies for OSA include continuous positive airway pressure (CPAP), lifestyle modifications, oral appliances, and surgical interventions. These strategies can be used in combination to improve outcomes, and some serve as alternatives to others.

2.2.4.1 Continuous positive airway pressure

CPAP is the most common and effective treatment for OSA and was first introduced by Dr Sullivan at the University of Sydney in 1981 [92]. It is the first-line therapy for severe OSA (AHI ≥15) and is also beneficial for treating mild to moderate OSA (AHI ≥5 and <15) [9, 93-97]. CPAP delivers a continuous stream of positive pressured air through a mask connected to a bedside air-pumping machine via a tube. This creates a pneumatic splint in the pharynx, keeping the airway open and maintaining oxygen levels. Studies have validated the effectiveness of CPAP therapy in reducing ventilatory drive in response to upper airway collapse and increasing the arousal threshold [95]. Additionally, improvements in memory, recognition skills, and increased grey matter volume following CPAP treatment highlight its potential to reverse OSA-induced cognitive and neurological impairments [98].

However, the side effects of CPAP play a critical role in influencing patient compliance. Common side effects include rhinorrhoea (runny nose) (35% of CPAP users), dry skin (65% of CPAP users), nasal congestion (25% of CPAP users), mask intolerance, air leakage, difficulty exhaling, and excessive device noise [99]. When CPAP was first introduced, the compliance rate was as low as 46% due to these side effects. However, with advancements in patient education and machine design, compliance rates have significantly improved, now exceeding 80% [100, 101]. These improvements have been specifically aimed at mitigating side effects to enhance patient adherence.

To address mask-related discomfort, various CPAP mask designs have been developed to improve fit and adherence. These include oronasal masks, nasal masks, and nasal pillow masks, settings as shown in **Figure 2.5**. Oronasal masks cover both the nose and mouth, whereas nasal masks fit over the nose only. Nasal pillow masks, a lightweight and user-friendly alternative, provide a modified version of nasal masks. Other less commonly used designs include total face masks, which cover the entire face, oral masks, which cover only the mouth, and hybrid masks, which combine the previously mentioned designs. Studies have shown that switching to a different mask type can effectively reduce air leakage, improve patient adherence, and lower residual AHI [102].

Additionally, unlike CPAP, which provides constant air pressure and may cause breathing discomfort, auto-titrating CPAP (APAP) and bi-level positive airway pressure (BiPAP) deliver

variable airway pressure to improve tolerance and compliance. APAP, also referred to as a "smart machine," continuously adjusts air pressure based on detected sleep stages and body position. It has been shown to be as effective as CPAP in normalising the AHI and improving daytime sleepiness while also enhancing adherence in both adults and children during the initial stages of treatment [103, 104]. BiPAP, similar to APAP, delivers two levels of air pressure, higher during inhalation and lower during exhalation. It specifically addresses a common CPAP-related side effect, where users may experience resistance during exhalation [105]. However, some studies have reported that neither APAP nor BiPAP is necessarily more effective than CPAP in improving compliance rates or OSA symptoms, and some patients prefer CPAP over these alternatives [73, 105].

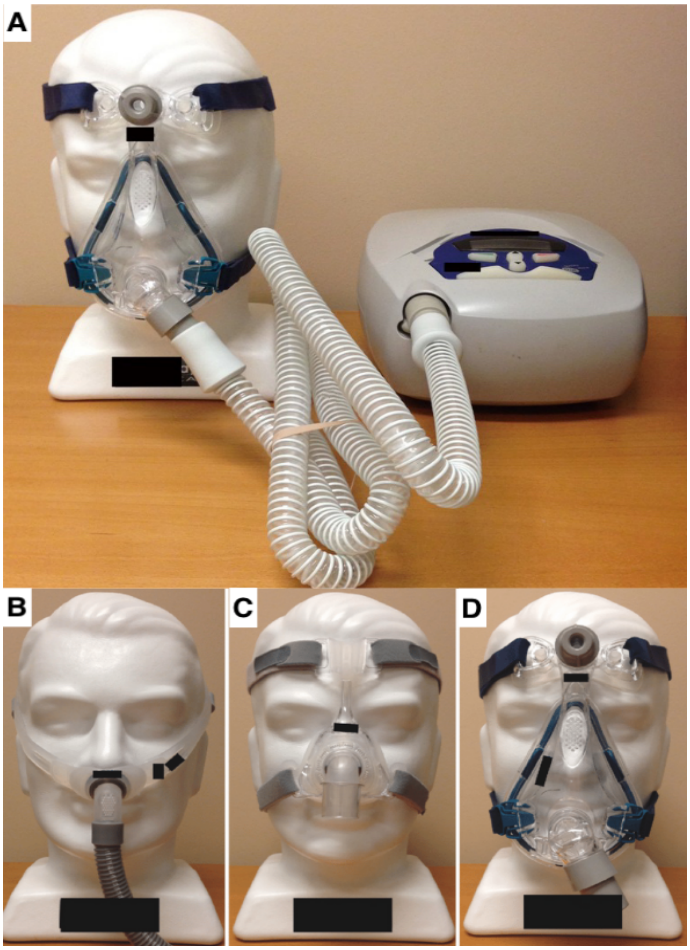


Figure 2.5 Example of CPAP setup (A) and face masks for CPAP treatment (B–D). (A) A standard CPAP setup: a tube connects the bedside CPAP blower unit to an oronasal face mask. The CPAP blower generates constant airway pressure, which is transmitted through the tube to the mask and then to the pharynx. (B) Nasal pillow masks, (C) Nasal masks, and (D) Oronasal masks. Figure modified from [106], Figures 1 and 4.

2.2.4.2 Lifestyle modification

Lifestyle modification aims to manage the risk factors associated with OSA. Rather than directly treating airway obstruction, this approach focuses on altering lifestyle habits to reduce the risk of developing OSA. Recommended strategies include weight loss, positional therapy, reducing smoking frequency, limiting alcohol intake, and preventing sleep deprivation, as these factors are commonly associated with OSA risk. Notably, lifestyle modifications are typically used in conjunction with other treatments, such as CPAP or oral devices, rather than as standalone therapies. Positional therapy, used as a supplementary treatment to CPAP, involves encouraging patients to sleep in a lateral position with their heads propped at a 30- to 60-degree angle. This technique helps stabilise the pharynx and consequently reduces the frequency of apnoea and hypopnoea events. However, positional therapy alone is insufficient for alleviating OSA symptoms and must be combined with CPAP for optimal effectiveness. Additionally, studies have shown that positional changes have limited impact during REM sleep [60]. Furthermore, improvements in OSA symptoms do not always correspond linearly with lifestyle modifications. For instance, weight loss has a curvilinear relationship with OSA improvement, meaning that reduction of AHI is observed only with substantial weight loss [107]. GLP-1 receptor agonist weight-loss medications, such as semaglutide and tirzepatide, have shown potential to reduce OSA severity by lowering body weight and improving AHI in individuals with obesity [108].

2.2.4.3 Oral appliance

Oral appliance mandibular advancement devices (OAm) function as mouthguards that reposition the tongue and jaw forward, increasing the space in the posterior pharyngeal area and consequently keeping the airway open for breathing, as shown in **Figure 2.6** [60, 94]. Unlike CPAP, which delivers steady positive airway pressure, OAms rely purely on mechanical displacement. As a result, their ideal use is limited to patients with healthy teeth to anchor the device, flexible temporomandibular joints, and unobstructed nasal airways. OAms are commonly prescribed as an alternative for patients who cannot tolerate CPAP. Most oral appliances are designed primarily to address snoring, with only a subset specifically developed for OSA treatment. Studies have shown that OAms are effective for patients with mild to moderate OSA but are not recommended for those with severe OSA [109]. Even among patients with mild to moderate OSA, CPAP has been demonstrated to provide greater improvements in the AHI and daytime sleepiness compared to OAms [110, 111].

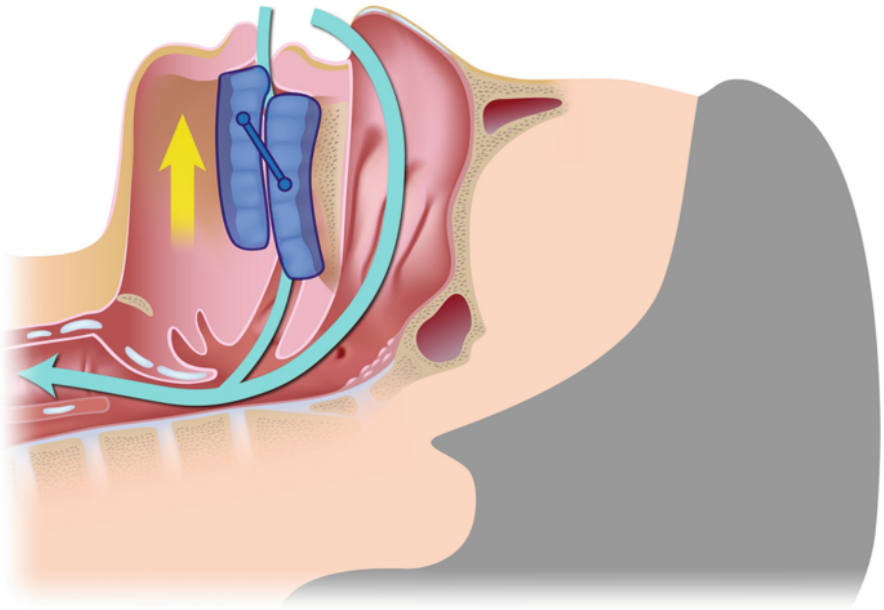


Figure 2.6 Example of an oral appliance mandibular advancement device. The mouthguard positionally advances the jaw and tongue (yellow arrow), increasing the pharyngeal space and reducing airway collapse (blue arrow). Figure adapted from [109], Figure 3 (Copyright Alila Medical Media).

2.2.4.4 Surgical interventions

Surgical intervention is often offered to patients with moderate to severe OSA, who are unable or unwilling to undergo CPAP or oral appliance therapy. A wide range of surgical options are available, most of which target anatomical obstructions [60]. Surgical treatments can be categorised into three main types: procedures to improve nasal patency and breathing, procedures to address retropalatal obstruction, and procedures to treat tongue-base and hypopharyngeal obstruction [112].

Procedures targeting nasal patency and breathing aim to clear nasal obstructions, thereby improving airflow and allowing patients to better tolerate CPAP or oral appliances. Retropalatal procedures, such as uvulopalatopharyngoplasty and the uvulopalatal flap technique, widen the airway by addressing obstructions in the soft palate, lateral pharyngeal walls, and tonsils. However, these procedures are not designed to enhance CPAP adherence and may result in side effects such as oral air leakage, which can reduce CPAP compliance. Even when performed by experienced otolaryngologists, the success rate for procedures targeting retropalatal obstruction is relatively low, at approximately 40%.

Surgical interventions for tongue-base and hypopharyngeal obstruction vary in technique but generally demonstrate higher success rates compared to other methods. Among these, maxillomandibular advancement is currently the most effective surgical option, showing significant reductions in AHI postoperatively. Initially considered a secondary option after soft tissue surgery, maxillomandibular advancement is increasingly being used as a first-line treatment for patients with significant craniofacial anomalies or multiple obstruction sites. The success rate of maxillomandibular advancement ranges from 75% to 100%, making it a highly effective alternative to CPAP. For patients who have failed CPAP or oral appliance therapy, maxillomandibular advancement provides a promising opportunity to alleviate OSA severity [113, 114].

2.2.4.5 Standard referral algorithm for OSA treatment

For patients presenting with sleep disorder symptoms, the standardised diagnostic and treatment procedure is outlined in **Figure 2.7** [112]. A thorough medical history assessment is conducted to identify risk factors associated with OSA, including BMI, blood pressure, and cholesterol levels. Additionally, an examination of the tonsil size, nasal passages, oral cavity, and neck is performed to assess anatomical factors relevant to treatment planning. If OSA is suspected, patients undergo a sleep study for diagnosis and are referred to a sleep physician. Based on the diagnosis, patients may be advised to undertake non-invasive treatments such as lifestyle modifications, CPAP therapy, or using oral appliances. For individuals who decline or are unable to tolerate CPAP or oral appliance therapy, surgical intervention may be considered. At this stage, patients are referred to an otolaryngologist for a more detailed pre-surgical evaluation and preparation for surgery.

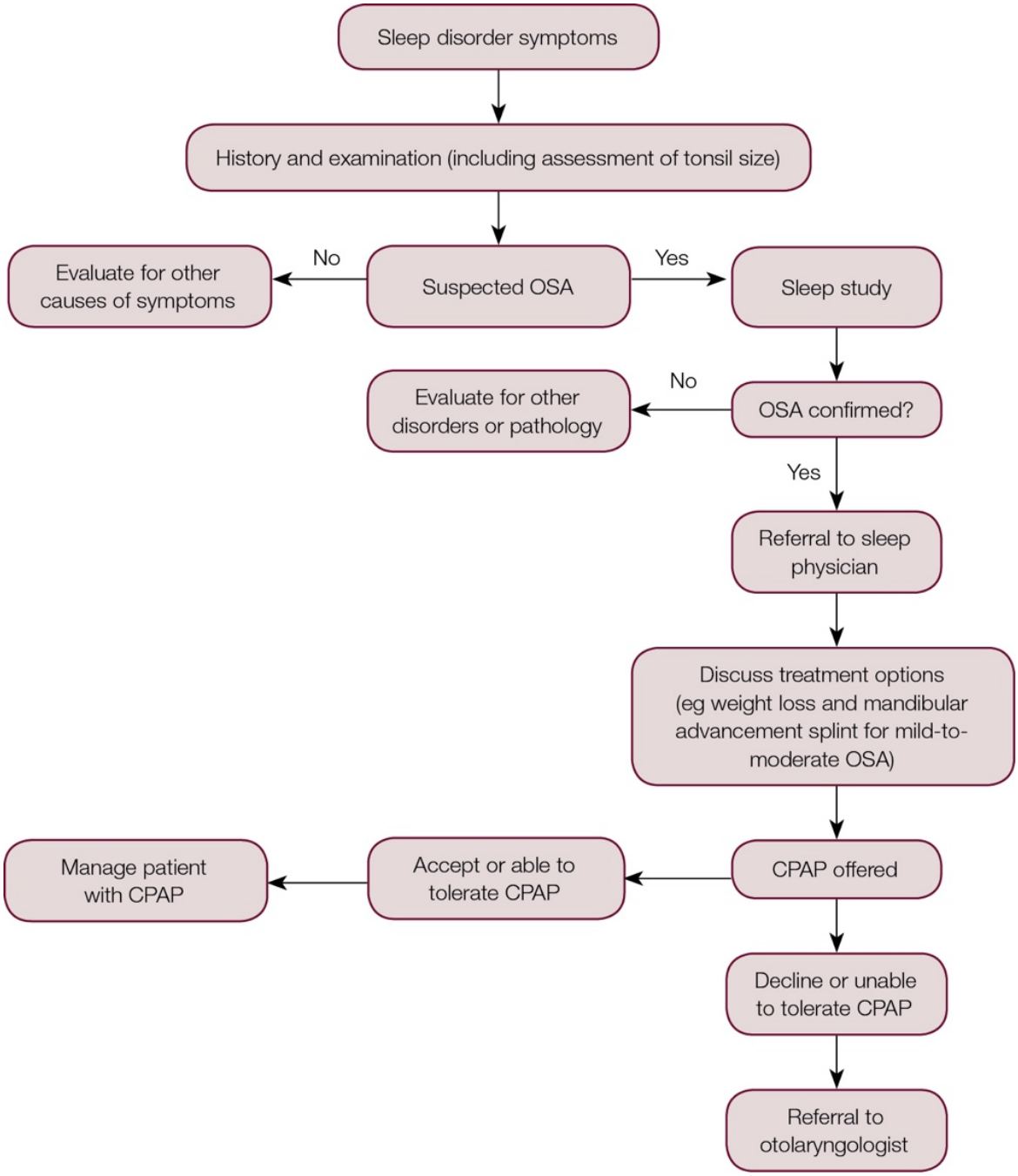


Figure 2.7 Referral algorithm for OSA treatment. Figure is adapted from [112] Figure 2.

2.3 Associations between Cardiovascular disease (CVD) and Obstructive Sleep Apnoea (OSA)

OSA is increasingly recognised as a significant risk factor for CVD, with recurrent hypoxemic cycles during sleep contributing to adverse cardiovascular effects through inflammation, sympathetic activation, and oxidative stress. Untreated OSA has been linked to neurocognitive impairments (such as daytime sleepiness and reduced attention) and metabolic complications (including diabetes, hypertension, and stroke), ultimately driving the development and progression of CVD. Studies have reported a co-aggregated relationship between CVD and OSA, with a high prevalence of OSA observed among patients with CVD.

Given the global health burden of CVD and its association with OSA, improving CVD outcomes prediction through OSA measurements could help mitigate its impact, particularly in older populations. In sleep research, AHI is traditionally used to quantify OSA severity and assess its relationship with CVD. However, evidence suggests that AHI is a poor predictor of CVD outcomes, as it fails to capture critical physiological factors involved in cardiovascular pathophysiology [11, 27]. To address this limitation, novel PSG-derived parameters are being explored as alternative predictors to enhance risk stratification.

2.3.1 Pathophysiology and Epidemiology

Previous studies have not established a direct link between OSA and CVD, as OSA is frequently associated with comorbid conditions such as diabetes, hypertension, and obesity. Each of which is an independent risk factor for CVD [86]. The theoretical pathway linking OSA to CVD is summarised in **Figure 2.8** [115]. OSA contributes to intermittent hypoxia, sleep fragmentation, negative intrathoracic pressure, arousals, and, in some cases, hypercapnia. These physiological disturbances stimulate inflammatory response, trigger sympathetic nervous system activation, and increase oxidative stress levels. These changes, in turn, contribute to metabolic dysregulation and vascular endothelial dysfunction. Such pathophysiological responses may directly drive the development of CVD events or lead to intermediate clinical conditions, such as hypertension, diabetes, and dyslipidaemia, that are widely recognised as primary risk factors for CVD.

A key component of the pathophysiological pathway is oxygen disturbance during sleep, with OSA generating recurrent episodes of hypoxaemia, which is defined as reduced arterial oxygen

saturation and quantifiable using pulse oximetry metrics [16]. Hypoxia, in contrast, refers to inadequate oxygen availability at the tissue level and may arise downstream of hypoxaemia or from impaired oxygen delivery or utilisation independent of arterial saturation [116]. Although hypoxaemia can contribute to tissue hypoxia, the two states are not synonymous, and tissue hypoxia may occur even when arterial oxygen saturation remains within the normal range [116]. For example, in anaemia, oxygen saturation may be preserved despite reduced haemoglobin concentration and diminished total oxygen-carrying capacity [117]. Similarly, low cardiac output states can impair tissue oxygen delivery despite adequate arterial oxygenation [118]. Histotoxic hypoxia, in which tissues are unable to utilise delivered oxygen (e.g., due to cyanide toxicity), provides an additional example of dissociation between blood oxygenation and cellular oxygen use [116]. These distinctions highlight that oximetry-derived measures capture an important but incomplete aspect of oxygen-related mechanisms linking OSA to cardiovascular risk.

OSA has been shown to have a high prevalence among individuals with CVD risk factors and CVD outcomes. Epidemiological studies and clinical trials suggest that OSA is an independent risk factor for hypertension, a well-established risk factor of CVD [119-121]. Peppard et al. demonstrated that individuals with mild to moderate OSA have twice the odds of developing hypertension (odds ratio = 2.03), while those with severe OSA face a threefold increase in risk [119]. A prospective analysis of a Spanish cohort further indicated that long-term treatment of OSA with CPAP significantly reduces the risk of developing hypertension [122]. Additionally, OSA is frequently comorbid with CAD [123]. Studies have reported that 30.5% to over 50% of CVD cases consist of OSA, while an estimated 20% to 25% of OSA patients have CAD [124, 125]. This prevalence suggests that OSA may serve as an independent risk factor for CAD, despite the lack of a clearly established link between two diseases [86]. Similarly, research has shown that 47% to 76% of heart failure cases present with OSA symptoms. The presence of OSA exacerbates heart failure morbidity via increased sympathetic activation, cardiac afterload, and myocardial oxygen desaturation [126, 127]. Furthermore, OSA has been strongly associated with atrial fibrillation, with evidence of a co-aggregated relationship. Studies indicate that 43% of OSA patients experience persistent atrial fibrillation, while OSA prevalence reaches 49% in atrial fibrillation patients beyond cardioversion [128, 129].

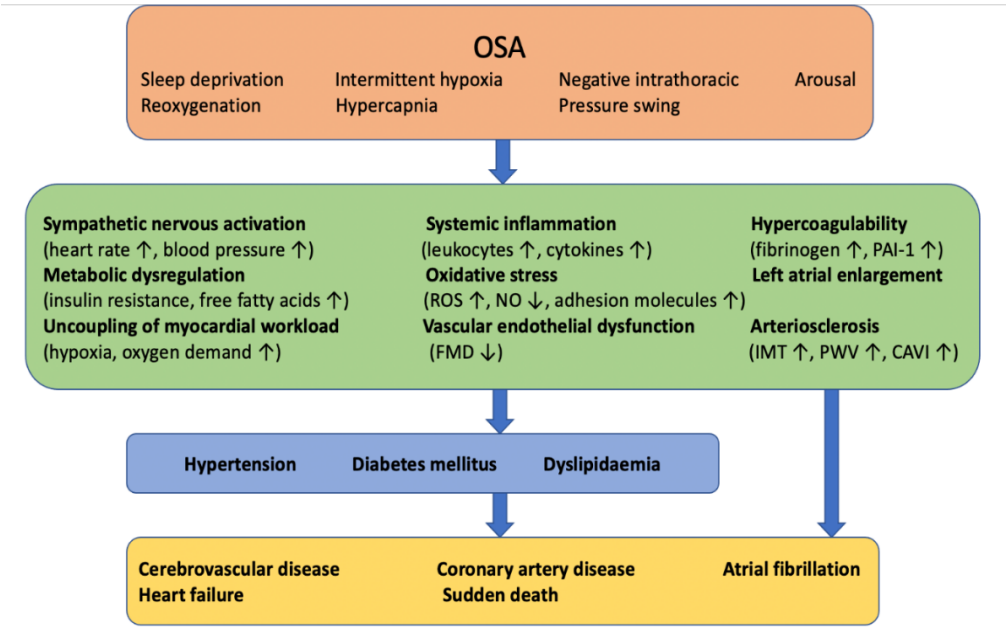


Figure 2.8 The pathophysiological link between OSA and CVD. The top block highlights the key mechanisms of OSA, followed by its pathophysiological consequences. These processes contribute to intermediate clinical conditions (blue block) and ultimately lead to major CVD events (yellow block). Figure is modified from [115], Fig 1.PAI-1: Plasminogen activator inhibitor-1; ROS: Reactive oxygen species; NO: Nitric oxide; FMD: Flow-mediated dilatation; IMT: Intima-media thickness; PWV: Pulse wave velocity; CAVI: Cardio-ankle vascular index.

2.3.2 Why predicting cardiovascular disease (CVD) outcomes matters

CVD is a leading cause of death worldwide, particularly among older populations, and often progresses asymptomatically for years, potentially leading to irreversible or life-threatening damage. Traditional CVD risk factors are classified into modifiable and non-modifiable categories, with effective management of modifiable factors helping to reduce overall risk. Accurately predicting an individual's future CVD risk could enable early intervention and targeted risk managing strategies. OSA, increasingly recognised as an independent risk factor for CVD, presents a valuable opportunity in this regard. Given that OSA symptoms, such as snoring, daytime sleepiness, and impaired attention, are often more detectable than early-stage CVD, leveraging sleep study data could enhance CVD prediction and facilitate timely and effective risk management. Integrating OSA assessment into CVD management strategies may ultimately help mitigate the global health burden of CVD.

2.3.3 Limitation of Apnoea-Hypopnoea Index (AHI)

While the AHI remains the gold standard for assessing OSA severity in clinical practice, its predictive value for future CVD events in OSA patients has been shown to be limited [130]. AHI quantifies only the frequency of apnoea and hypopnoea events, assuming that all respiratory disturbances have the same impact on OSA severity. However, it fails to account for inter-individual variability in the pathological consequences of these events and lacks the ability to capture detailed characteristics of associated oxygen desaturation and cortical arousals [131]. Muraja-Murro et al. demonstrated that an adjusted AHI, incorporating the duration of airway obstruction, was a stronger predictor of CVD mortality and morbidity than AHI alone. Similarly, Azarbarzin et al. found that AHI was insufficient for predicting heart failure in men, whereas T90% (the percentage of total sleep time with oxygen saturation below 90%) and desaturation area outperformed AHI in predicting all-cause mortality [12, 132]. Furthermore, AHI has not only been shown to be inferior to T90% in predicting CVD mortality but also exhibits limited effectiveness in forecasting hypertension risk [120]. Butler et al. further highlighted that additional PSG-derived parameters, such as respiratory event duration, provide predictive value for CVD outcomes beyond what AHI alone can offer [27]. These findings underscore the necessity of incorporating alternative PSG-derived metrics beyond AHI to enhance the analysis of the relationship between OSA and CVD [133].

2.3.4 Novel PSG-derived parameters

To address the limitations of AHI and maximise the utility of recorded PSG signals in sleep studies, alternative PSG-derived parameters have been investigated for predicting CVD events. This thesis primarily focuses on oximetry-derived parameters, which can be categorised into four main groups, as summarised in **Figure 2.9** [13]. Oximetry was selected from among the various PSG signals as an initial focus due to its direct reflection of hypoxaemia and its widespread accessibility. A detailed exploration of each parameter and its predictive performance for CVD outcomes will be presented in the following chapters. This section serves as an overview of the key parameters used in this thesis.

Desaturation-related parameters are typically associated with scored respiratory events and play a crucial role in describing the oxygen desaturation profile of patients. The ODI, despite variations in desaturation threshold criteria, quantifies the frequency of desaturation events. Desaturation depth and duration provide further insights into desaturation severity and can be

manifested as desaturation area. Time-based parameters quantify the overall distribution of oximetry data and are widely used in the prediction of CVD outcome. A notable example is T90, which calculates the cumulative duration of oximetry recordings below 90% and has demonstrated strong predictive performance for multiple CVD outcomes [134-136]. Frequency-domain metrics derived from the power spectral density and non-linear metrics, such as, regularity, complexity, and non-linear variability, are less commonly explored compared to desaturation-related and time-based parameters in current studies [13]. However, they provide valuable insights into sleep characteristics and can serve as important predictive markers in future research.

Moreover, recent studies have also incorporated ECG- and EEG-derived parameters in the analysis of OSA and its association with CVD. ECG-derived metrices, including PR interval, QRS interval, corrected QT interval, left ventricular hypertrophy, and heart rate, have been identified as strong predictors of CVD mortality, as demonstrated by Deo et al. [137]. Additionally, quantitative EEG measures in both the time and frequency domains have shown significant associations with all-cause mortality. As summarised in **Figure 2.10**, common frequency-domain metrics include the mean power of delta, theta, alpha, sigma, and beta frequency bands derived from the power spectral density. Commonly referenced time-domain parameters include the percentage of time spent in non-REM and REM sleep stages, total time of sleep, and the arousal index [138]. This thesis integrates selected EEG time-domain metrics both directly and indirectly to enhance the prediction of CVD risk.

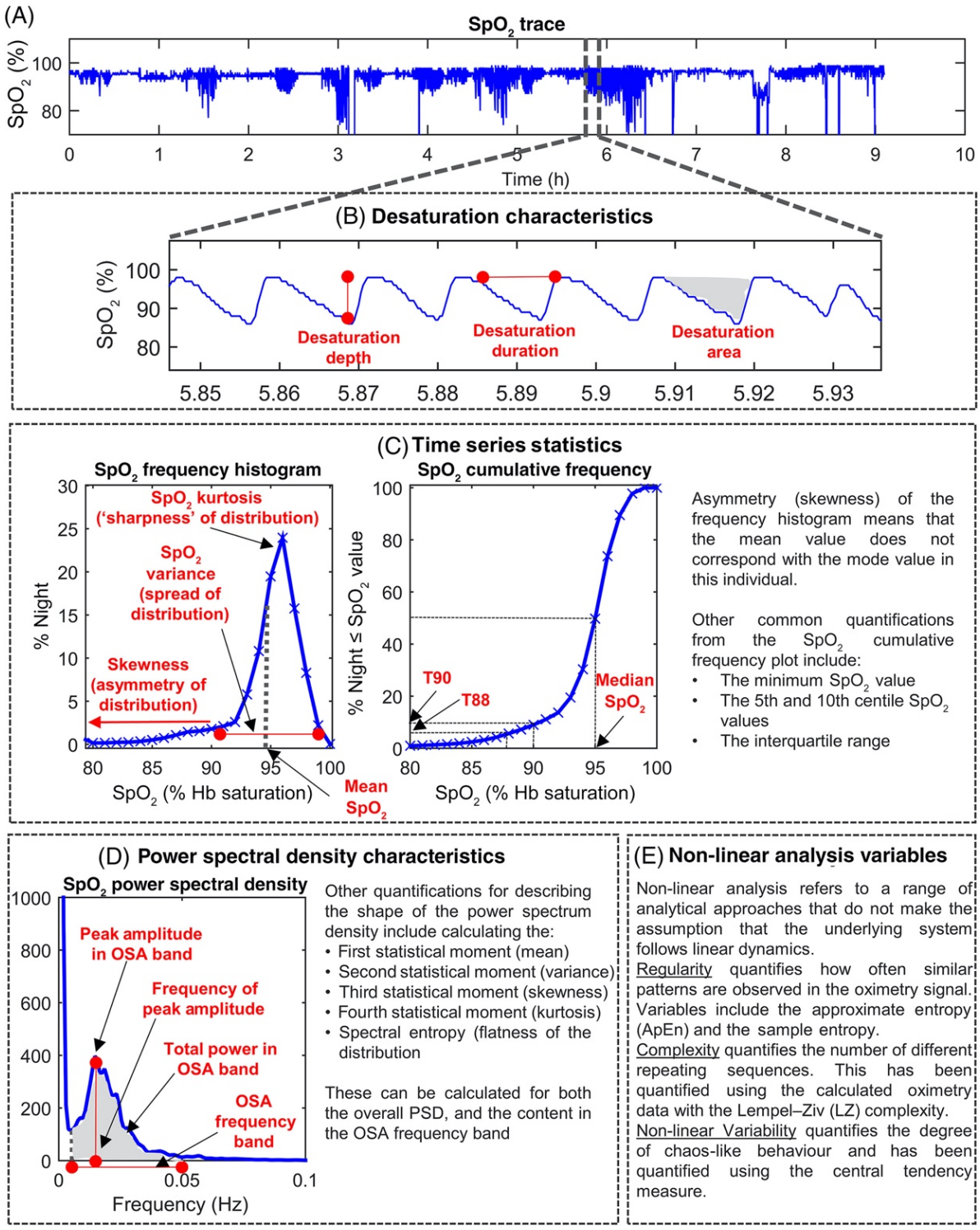


Figure 2.9 Summary of Oximetry-Derived Parameters for Predicting CVD outcomes. (A) Example trace of an oximetry recording. (B) Zoomed-in segment of oximetry data highlighting desaturation-related metrics, which collectively characterise oxygen desaturation and indicate OSA severity. (C) Overview of time-domain parameters. (D) Frequency-domain representation of the oximetry recording. (E) Summary of non-linear variables, including measures of regularity, complexity, and variability. Figure is adapted from [13], Fig 1.

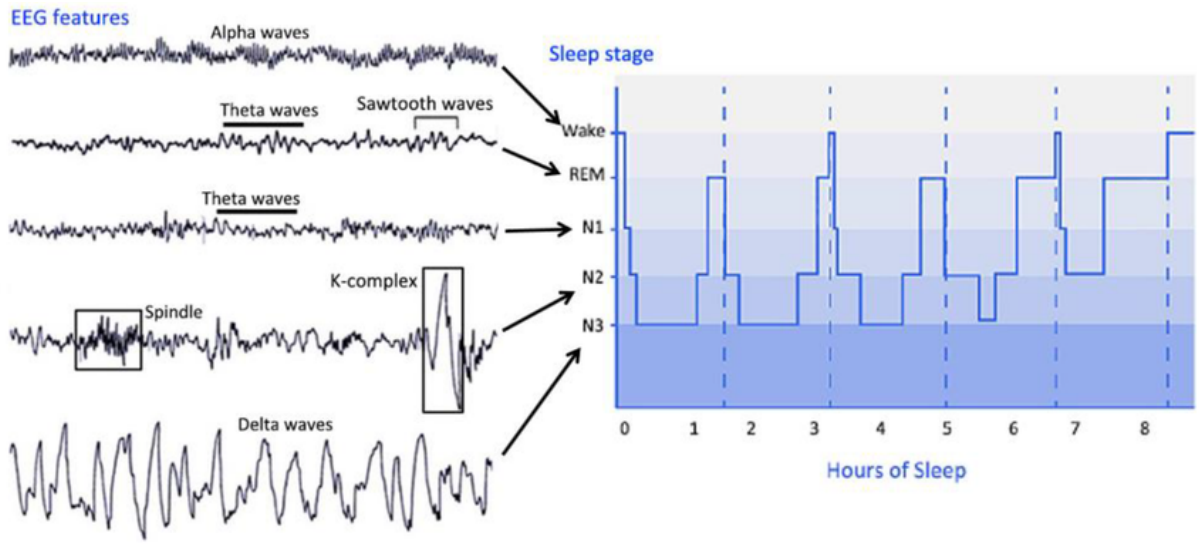


Figure 2.10 Common EEG measures across different sleep stages (REM and non-REM sleep) paired with corresponding hypnogram on the right. Figure is adapted from [139], Figure 2.

2.4 Summary

OSA is a sleep disorder caused by repeated collapse of the upper airway during sleep. It is more commonly observed in patients over 40 years old, with a larger BMI, or who have a narrow airway and a unique facial structure. Patients with OSA have a higher chance of developing depression, CVD, and diabetes, and of having car accidents. The OSA-induced repetitive upper airway obstruction leads to intermittent hypoxic events overnight and sleep fragmentation, resulting in adverse neurocognitive, daytime sleepiness, and metabolic complications. The nocturnal hypoxemic burden caused by cumulative hypoxic events can increase vascular inflammation, blood pressure, and sympathetic nervous system action, and ultimately may increase the risk of CVD, which is the leading cause of death worldwide. Studies have shown a clear association between OSA and CVD events, including but not limited to CAD, heart failure, and atrial fibrillation. Considering the global health burden of CVD and its association with OSA, the risk of CVD can be mitigated by measuring OSA condition, forecasting CVD outcomes, and managing relevant risk factors.

PSG is commonly used for OSA diagnosis with the AHI being the standard measure for determining the presence and severity of OSA. PSG signals record blood oxygen level (measured with finger-based pulse oximetry), respiratory pressure/flow and effort, brain activity, skeletal muscle activity, heart rate, and eye movements. The AHI measures the number of apnoea and hypopnoea events per hour of sleep. However, studies show that AHI is not a

good predictor of CVD mortality as AHI fails to capture factors that have crucial impacts on the cardiovascular system, namely, blood oxygen levels, high sympathetic activity, respiratory event duration, sleep fragmentation, and arousal events.

As the understanding of the links between CVD and OSA has grown, new PSG-based parameters have been proposed that may reveal more information about the impact of sleep apnoea on hypoxemia that may be predictive of future CVD events. Furthermore, PSG is primarily conducted for diagnostic purposes and is both expensive and resource intensive. Given that multiple physiological signals are recorded during PSG, it is inefficient to use these data solely to calculate the frequency of respiratory events.

This thesis aims to maximise the utility of PSG-derived signals to improve the understanding of the relationship between OSA and CVD, ultimately advancing the application of OSA measures in cardiovascular clinical practice. The thesis primarily focuses on oximetry signal, aiming to address current limitations of widely used oximetry-derived parameters. It examines how variations in computational approaches influence CVD outcome prediction and identifies suitable algorithms for handling large-scale datasets. Furthermore, the thesis develops an explainable machine learning model using PSG-derived parameters to predict CVD mortality outcomes at the individual level and across specific time horizons. The model is designed to deliver reliable predictions with minimal clinical input, making it accessible to the general population. It also provides insights into how sleep measurements contribute to CVD outcome prediction.

Chapter 3
Methodological background

3 Methodological background

In Chapter 2, the clinical background for this thesis was presented. The review examined the pathophysiological and epidemiological association between OSA and CVD, highlighted the need for using OSA measurements to predict future CVD outcomes, identified the limitations of the standard OSA diagnostic metric, AHI, in predictive studies, and briefly summarised the novel parameters explored as alternative solutions.

This chapter introduces the novel parameters employed in Experiments 1 and 2, along with the predictive performance evaluation methods used in this thesis. Section 3.1 provides a detailed explanation of the selected oximetry-derived parameters, including their mathematical definitions, variations in computational approaches, predictive performance in CVD analysis, and current limitations. Section 3.2 expands on additional parameters used in the study, though these are discussed briefly, as the primary focus of this thesis is on oximetry-derived parameters. Section 3.3 describes the Cox proportional hazards model, the principal statistical method used to evaluate the predictive performance of these parameters. As the predictive ability of oximetry-derived parameters is evaluated using hazard analysis, this thesis further explores the potential of machine learning models to address the central research question: *Can PSG-derived parameters effectively predict CVD outcomes at the individual level?* Section 3.4 introduces the machine learning techniques applied alongside the features described in Sections 3.1 and 3.2, aiming to enhance predictive accuracy and risk stratification.

This chapter serves solely as a literature review, providing the technical background necessary for a better understanding of the two experiments conducted in this thesis. It highlights the potential limitations of existing novel parameters and established the rationale for the experiments. The developed methods and algorithms for both experiments will be detailed in the following chapters.

3.1 Oximetry-derived parameters

The primary focus of this thesis is on oximetry-derived parameters that quantify nocturnal hypoxaemia from pulse oximetry signals. Hypoxaemia during sleep may manifest as either sustained reductions in oxygen saturation or intermittent, event-related desaturations [140]. Measures based on cumulative duration, such as T90, are commonly used to characterise prolonged or sustained nocturnal hypoxaemia, which may reflect underlying cardiopulmonary impairment or sleep-related hypoventilation [140]. In contrast, event-based metrics such as the ODI and desaturation area-based parameters are designed to capture intermittent desaturation patterns that are typical of OSA [15, 16]. Although these oximetry-derived parameters are often conceptually linked to distinct hypoxaemia patterns (sustained versus intermittent), they are not mutually exclusive and are widely applied in the assessment of OSA [15].

This section provides a detailed description of oximetry-derived parameters. The computational approaches discussed here are based on published literature and serve as baseline methods for experiments. This section covers T90, ODI, and variations of desaturation area-based parameters (described in Sections 3.1.1–3.1.3). The predictive performance and limitations of each parameter in relation to CVD outcomes are discussed in detail to establish the motivation for later experiments (presented alongside each parameter and summarised in Section 3.1.4).

3.1.1 Time below 90% Saturation

T90 is widely recognised as an independent predictor of all-cause CVD mortality and is extensively used in research [6, 141-145]. The calculation methods for T90 vary across studies and can be broadly classified into time-based and percentage-based approaches, as summarised in **Table 3.1** [6, 12, 14, 145-147].

The time-based TST90 quantifies the total duration during sleep in which oxygen saturation falls below 90%, thereby measuring the cumulative hypoxemic burden. In contrast, the percentage-based T90% calculates the proportion of total sleep time spent below 90%, reflecting the rate of hypoxemic insult. Both methods have demonstrated strong predictive performance for CVD events. Xu et al. concluded that TST90 is a robust predictor of major adverse cardiovascular events (MACEs) and significantly outperforms AHI in this role [145]. Baumert et al. further categorised time-based TST90 according to the proximity of the dips

below 90% to desaturation events. T90desaturation represents the cumulative duration during which SpO2 experiences at least a 4% oxygen desaturation associated with TST90. In contrast, T90non-specific represents the duration during which SpO2 fails to fully recover following a preceding desaturation event, defined as an incomplete return to at least two-thirds of the baseline saturation within 150 seconds of event onset. This incomplete recovery results in sustained baseline oxygen levels below 90%. Both metrics have been shown to be strong indicators of the association between OSA and CVD mortality. The researchers suggested that incorporating T90desaturation and T90non-specific as multivariate inputs could enhance the predictive performance of CVD mortality models [6].

The percentage-based T90% is also a reliable predictor of CVD events. Wang et al. found that T90% outperformed TST90 in predicting incident CVD in patients with non-sleepy sleep-disordered breathing (SDB) [14]. However, other studies have not consistently supported this conclusion. For instance, Sutherland et al. reported no significant association between T90% and incident CVD in OSA patients [148]. Currently, it remains unclear whether TST90 or T90% serves as the superior predictor of CVD outcomes. Further research comparing the performance of these metrics across different datasets may be required.

Table 3.1 Different calculation methods of T90. Examples of each method with the corresponding database, the aim of analysis, and results are provided.

Type of T90	Name of T90	Calculation Method	Population/Aims
Time-based parameter	TST90 [145]	The total sleep time below 90% oxygen saturation. The unit of this parameter is hours.	1860 Chinese participants from a clinic-based retrospective cohort study in Hong Kong, China. Participants were excluded from study if they had a sleep disorder other than OSA; received treatments other than CPAP; or had conditions with a known effect on OSA. Aims: Association between T90 and MACEs.
	T90desaturation [6]	The total sleep time with at least 4% oxygen desaturation while the oxygen level drops below 90% as shown in **Figure 3.1**A. The unit of this parameter is minutes.	3135 community-dwelling male participants aged 65 years old and above from the Osteoporotic Fractures in Men Study (MrOS). Aims: Association between T90desaturation and CVD mortality.
	T90non-specific [6]	The total sleep time associated with non-specific drifts in oxygen saturation. As shown in **Figure 3.1**B, due to the incomplete recovery of the previous desaturation event, the oxygen baseline is drifting, and the oxygen level is below 90% without experiencing oxygen desaturation. The unit of this parameter is minutes.	3135 community-dwelling male participants aged 65 years old and above from the MrOS sleep study. Aims: Association between T90desaturation and CVD mortality.
Percentage-based parameter	T90% [14]	The percentage of sleep time with oxygen saturation level below 90%. The unit of this parameter is %.	3626 randomly selected Chinese community-dwelling participants. A total of 30.7% of the participants suffer from SDB, of which 96.5% is non-sleepy SDB. Aims: Association between T90% and CVD incident in non-sleepy SDB patients.

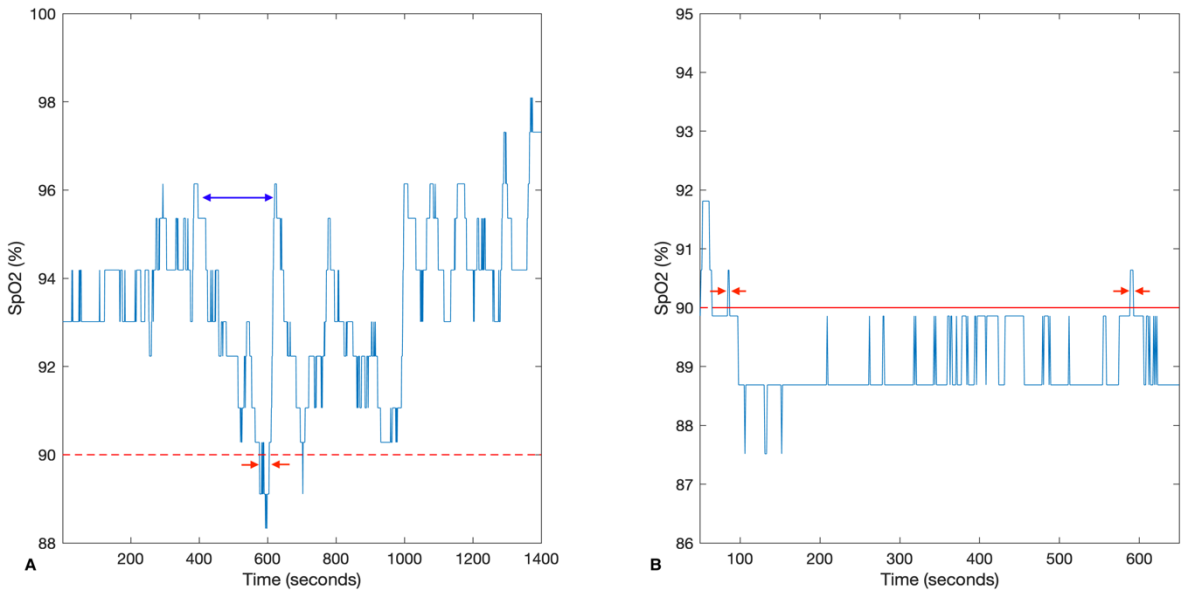


Figure 3.1 Pulse oximetry trace from the SHHS database. (A) T90desaturation represents the time when SpO2 is below 90% while experiencing at least 4% oxygen desaturation. It counts the duration of oxygen level below 90% (as indicated by red arrows) associated with acute desaturation events (as indicated by blue arrow). (B) T90non-specific represents the time when SpO2 fails to fully recover from previous desaturation event or SpO2 experience baseline drifts and results in the baseline oxygen level below 90%. It counts the duration of oxygen level below 90% (below red baseline) and excludes time of oxygen level above 90% (as indicated by red arrow).

3.1.2 Oxygen Desaturation Index

ODI is commonly used to indicate intermittent hypoxemia and is defined as the number of oxygen desaturation events per hour of sleep [12]. Although ODI and AHI both measure event rates, ODI performs better in predicting adverse CVD outcomes [149, 150]. ODI measures the number of transient desaturation events from a baseline value and divided by the hours of sleep. The AASM does not specify the criteria for scoring desaturation events [150-154], and hence a range of methods have been used to calculate ODI. Some studies define ODI as the rate of oxygen desaturation events occurring when SpO2 drops lower than the desaturation threshold from the average saturation in the previous 120 s and persists for at least 10 s [155, 156]. An issue with this definition occurs when events are separated by less than 120 s (as it happens on average for a severe sleep apnoea case), resulting in the baseline being influenced by previous events. Other studies chose the baseline as either the average SpO2 value of the whole recording or the mean SpO2 value in the first 3 min [157-160]. The desaturation thresholds of

3% (ODI3) or 4% (ODI4) are commonly chosen in the analysis of OSA and CVD [152, 161, 162]. Sutherland et al. provided a comparison of ODI2, ODI3, ODI4, and ODI5 for predicting prevalent CVD in OSA patients free of CVD at baseline, and found that 4% and 5% provide the best performance in predicting CVD events in women [148, 163]. Karhu et al. concluded that ODI4 is more reliable than ODI3 in determining the impact of OSA, since respiratory events with desaturation ≥ 4% are usually considered as hypopnoea [16, 152]. However, results from several studies showed that ODI3 as a CVD risk factor has a higher significant odds ratio than ODI4 [164, 165]. Further research undertaken by Punjabi et al. explored whether ODIs within a specific range (2–2.9%, 3–3.9%, and 4–4.9%) are associated with CVD events. The results showed that only ODI (4–4.9%) is statistically significant in the analysis,and supported Tuomas et al.'s findings on ODI4 [163]. The hardware and software used to measure ODI metrics also impacts the ODI parameter. There was a clinically significant difference between the ODI measurements from the same studies measured using the ResMed ApneaLink Plus device (ResMed, Sydney, Australia) and the Compumedics Grael Profusion PSG3 system (Compumedics Limited, Abbotsford, Victoria, Australia) [153]. Ng et al. suggested that this discrepancy may be caused by the noise cancellation process rather than the ODI scoring algorithm [153].

3.1.3 Desaturation Area-Based Parameters

Recent studies introduced novel parameters as potential indicators of future CVD events. These parameters quantify the area above the SpO2 curve associated with key sleep disorder breathing events per hour of sleep. They are distinguished by their different calculation methods. The units of these measures are oxygen saturation %, and they thus provide a weighted average representation of the SpO2 trace. This group of parameters can be categorised according to their dependence on the respiratory event scoring. Hypoxic Burden (HB) and Respiratory Event Desaturation Transient Area (REDTA) are derived from manually scored respiratory events, while hypoxic load (HL) and Desaturation Severity (DesSev) are independent of respiratory events [12, 28, 166, 167].

3.1.3.1 Respiratory event scoring

Respiratory event scoring is a systematic process used to identify sleep stages, arousals, and respiratory events. This scoring occurs after overnight PSG recording and serves as a foundation for subsequent sleep analysis. This section uses the Sleep Heart Health Study (SHHS) as an example to provide a detailed explanation of the PSG scoring process.

Respiratory event scoring in SHHS was conducted by the Sleep Reading Centre (SRC) in Boston, USA, following AASM criteria. All PSG recordings were initially pre-analysed using the Compumedics software and subsequently reviewed by sleep experts. The scoring process involved two review stages. In the first stage, sleep experts manually identified sleep stages and arousals on an epoch-by-epoch basis. In the second stage, oxygen desaturation and respiratory events (apnoea and hypopnoea) were manually marked using a 2- or 5-minute sampling period.

Sleep stages and arousals were identified based on EEG signals, divided into 30-second epochs, following the Rechtschaffen and Kales criteria [168, 169]. The classification of respiratory events (apnoea and hypopnoea) was based on airflow measurements. Events with airflow reduction >75% lasting ≥10 seconds were classified as obstructive apnoea, while events with airflow reduction >30% lasting ≥10 seconds were classified as hypopnoea. Central apnoea was labelled in the absence of chest or abdominal movement. Desaturation events associated with respiratory events were assessed based on amplitude attenuation. The desaturation sampling window was centred around the nadir point, typically within a 30-second timeframe. Desaturation magnitude was calculated as the amplitude difference between the nadir and the maximum oxygen level within the sampling window [170].

This thesis incorporates apnoea and hypopnoea events that involve either oxygen desaturation >3% or an arousal event occurring within 5 seconds after the respiratory event in the calculations presented in Sections 3.2.3.02 and 3.2.3.03.

3.1.3.2 Hypoxic Burden (HB)

Azerbarzin et al. proposed HB which is defined as the sum of the area between the SpO2 trace and the desaturation baseline associated with all apnoea and hypopnoea events divided by the total time of sleep, as shown in Equation 3.1 [12]:

$$\boldsymbol{HB} = \frac{\sum^{\boldsymbol{events}} \boldsymbol{area\ of\ an\ individual\ desaturation\ event}}{\sum \boldsymbol{time\ of\ sleep}} \tag{3.1}$$

The authors present the unit of HB as %minutes per hour of sleep which is equivalent to a measure with units of % scaled by a factor of 60.

The calculation for HB can be decomposed into three main steps: (1) All SpO2 segments associated with manually scored respiratory events for one individual recording are averaged

and processed to calculate the boundaries of a sampling window. The sampling window boundaries are determined by the two peaks of the averaged respiratory event as shown in **Figure 3.2**A. (2) The desaturation baseline for each respiratory event (**Figure 3.2**B) is calculated as the maximum SpO2 value within 100 s prior to the end of the event (**Figure 3.2**C). (3) The desaturation area for a single respiratory event is the area within the sampling window, desaturation baseline, and the SpO2 trace, as shown in **Figure 3.2**C. HB is then calculated using Equation 3.1 [12].

Other researchers have attempted to replicate HB with varying success. Trzepizur et al. [171] developed their own algorithms for HB but post hoc analysis by Mehra and Azarbarzin [12] suggested Trzepizur et al.'s method underestimated HB [172]. Based on the published material in [12], the algorithm was replicated by de Chazal et al., and the MATLAB code is publicly available in the online sharing platform GitHub (https://github.com/pdechazal/Hypoxic-Burden (14 March 2025)). Two other commercial software packages calculate HB, Respironics (Murrysville, PA, USA) and Cidelec (Sainte-Gemmes-sur-Loire, France) [173]. However, the lack of a full disclosure of the algorithmic details of HB by the original authors has led to some confusion in the reproducibility of the HB calculation.

Researchers suggest that HB has a better performance than AHI in predicting CVD mortality and morbidity as it measures more information about the depth and duration of desaturations associated with apnoea and hypopnoea events [132, 171, 173-175]. Azarbarzin et al. conducted analysis on two population groups and demonstrated that HB has uniformly good performance for predicting CVD mortality in the two groups [12]. Blanchard et al. explored the correlation between OSA and stroke incidences using the database of the Pays de la Loire Sleep Cohort and, concluded that HB was a significant predictor of CVD events [174]. Trzepizur et al. compared the performance of ODI, T90, and HB in predicting MACEs, and concluded that T90 performs the best, while HB also proved to be a promising predictor [28, 171]. However, HB outperformed T90 in predicting CVD mortality in patients from the SHHS [28]. The varied conclusions regarding the performance of T90 and HB may be caused by the differences in database and target CVD events. As T90 and HB present different information derived from the SpO2 trace, future research could consider T90 and HB as multivariate predictors of CVD events.

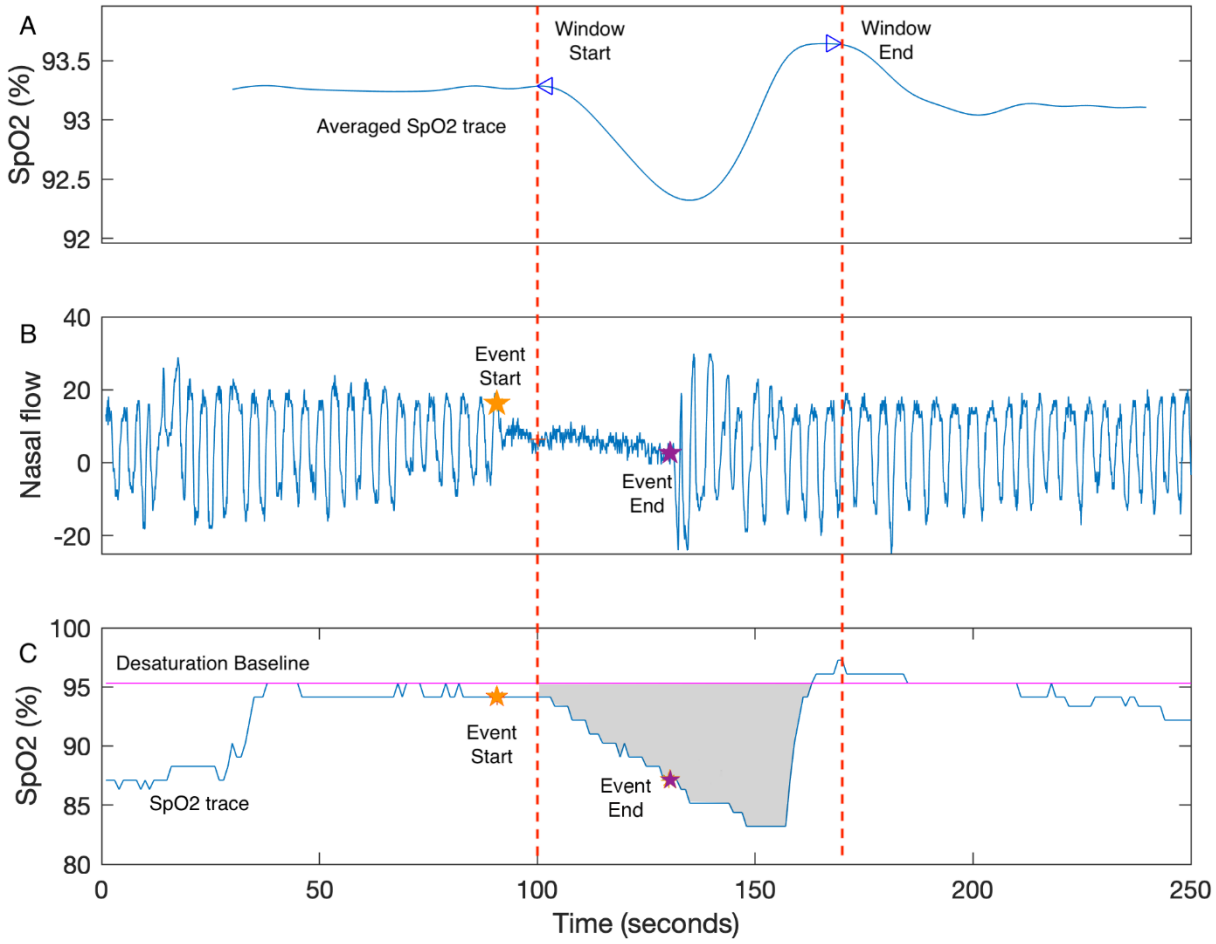


Figure 3.2 The example of HB calculation. (A) The sampling window is defined as the two peaks of the averaged SpO2 trace. (B) The nasal flow (blue) and the end points of a respiratory event (event start: yellow star; event end: purple star) are shown. (C) The SpO2 trace of the corresponding respiratory event is shown. The desaturation area for a single event (grey) is the area above the SpO2 trace (blue), below the desaturation baseline (magenta), and within the sampling window (between window start and window end). The desaturation baseline is the maximum SpO2 value within 100s prior to the event end (yellow star). HB is calculated as the sum of desaturation events divided by the total time of sleep [12].

3.1.3.3 Respiratory Event Desaturation Transient Area (REDTA)

Studies note that HB has limitations when used for some noisy recordings or recordings with few respiratory events. The accurate calculation of the desaturation area is challenging, as the desaturation baseline and the onset or offset of the average desaturation response are susceptible to noise. Moreover, the desaturation baseline is difficult to estimate when the interval between desaturation events is less than 100 s [28]. Our group proposed REDTA as a novel desaturation area-based parameter, which is less sensitive to noise than HB and has good predictivity of long-term CVD outcomes. REDTA is defined as the sum of the area between the SpO2 trace and the 100% desaturation baseline for all manually scored respiratory events divided by 3600, as shown in Equation 3.2:

$$\boldsymbol{REDTA} = \frac{\sum_{\boldsymbol{events}} \boldsymbol{area\ of\ an\ individual\ desaturation\ event}}{\boldsymbol{3600}}, \tag{3.2}$$

where the unit of each desaturation area is %seconds and the unit of REDTA is % hours [28].

REDTA is calculated using three main steps: (1) The sampling window is fixed and starts from the midway through the event and extends for 2.5 times the event duration. The sampling window is population-based (derived from the SHHS study) and is assumed to be appropriate for all respiratory events. (2) The desaturation area for a single respiratory event is the area between the 100% desaturation baseline and the SpO2 trace within the sampling window, as shown in **Figure 3.3**. (3) REDTA is the sum of the desaturation area divided by 3600. REDTA does not include the total time of sleep in the calculation. Its value increases with longer desaturation duration, more desaturation events, and greater depth of desaturation [28], and it thus is a measure of the hypoxemia insult per night. The software ABOSA (Version 1.1) implements REDTA [176]. The unit of REDTA is %hours.

REDTA was proposed to provide a simple, reproducible desaturation area-based SpO2 measure [28, 177-179]. Pahari et al. and de Chazal et al. compared the prognostic value of T90, ODI3, HB and REDTA for predicting CVD mortality. They concluded that REDTA performed comparably to HB (HR = 1.71 vs 1.62) and provided stronger predictive associations than conventional metrics such as T90 (HR = 1.48) [28, 180]. Further work investigating REDTA and other CVD events is required.

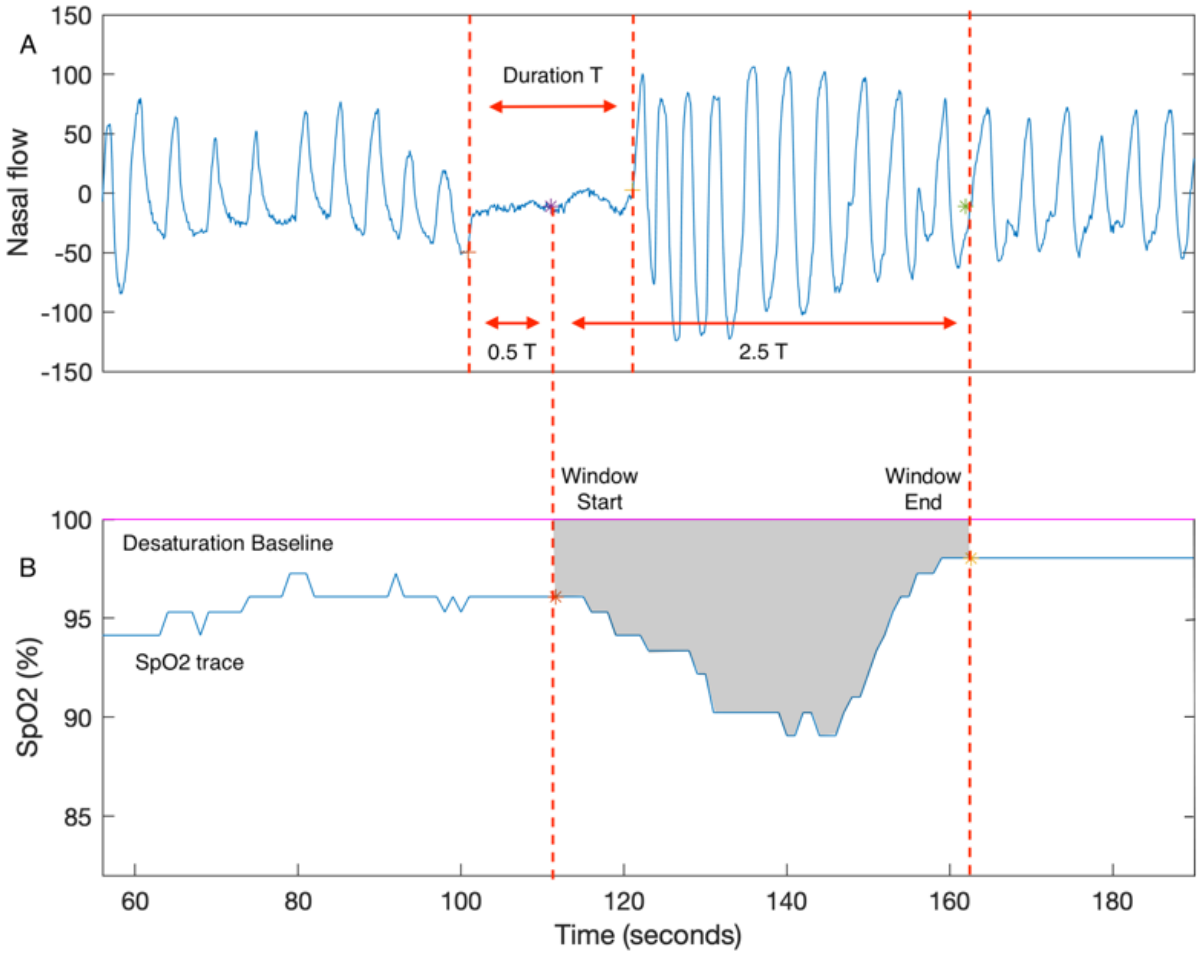


Figure 3.3 The example of REDTA calculation. (A) The nasal flow and a respiratory event are shown. The sampling window starts at the midway of the respiratory event and extends for 2.5 T, where T is the event duration. (B) The SpO2 trace of the corresponding respiratory event is shown. REDTA is calculated as the sum of the area (grey) within the sampling window, SpO2 trace, and the 100% desaturation baseline divided by 3600 [28].

3.1.3.4 Desaturation Severity (DesSev)

Unlike HB and REDTA, DesSev does not use the respiratory events to calculate the area, which is derived based on automatically detected desaturation events. The software package ABOSA implements DesSev and is freely available for other researchers to use [176]. DesSev is defined as the sum of the desaturation area associated with SpO2 events with a saturation drop greater than 3% divided by the total time of sleep [29, 176]. The calculation of DesSev can be divided into three steps, including an automated desaturation events detection algorithm: (1) The potential start and end points of desaturation events are approximately identified. The start point is the peak of the SpO2 signal, and the end point is located at the minimum of the SpO2 signal, with at least 5 s between start and end point. (2) The start and end points are matched

to form the candidate desaturation event list. The desaturation events are automatically selected from the candidate desaturation event list based on four criteria: the event duration does not exceed 180 s; desaturation events do not overlap; if the flat plateau is longer than 30 s, the corresponding end point moves up to the end of the plateau; and the transient drop of desaturation event is greater than 3%. (3) As shown in **Figure 3.4**, the desaturation area is calculated as the area between the desaturation baseline and the SpO2 trace within the sampling window, while the desaturation baseline is the value of the start point, and the sampling window is defined by the start and end points of each automatically detected desaturation events. DesSev is calculated using Equation 3.3:

$$\boldsymbol{DesSev} = \frac{\sum_{n=1}^{\boldsymbol{number\ of\ desaturation\ events}} \boldsymbol{Desaturation\ area_n}}{\sum \boldsymbol{time\ of\ sleep}}, \tag{3.3}$$

and the unit of DesSev is % [176]. The algorithm for calculating is DesSev is complex but conveniently the authors provide an online analysis package (ABOSA) that can process a supplied SpO2 signal [176].

Studies have used DesSev to explore the association between OSA, CVD events, and cardiac response, and have concluded that DesSev is an informative indicator of OSA and cardiac response [181-187]. Kainulainen et al. concluded that there is a stronger association between average daytime sleepiness latency and DesSev than AHI or ODI. They suggested that excessive daytime sleepiness is more related to the depth and duration of desaturation events rather than to the number of desaturation events [181, 182, 184]. Associations between DesSev and the short-term time- and frequency-domain heart rate variability (HRV) parameters have been explored, and authors concluded that there is a significant association between DesSev and HRV in OSA patients [185]. DesSev may have some key limitations when applied to OSA patients. Because DesSev is independent of respiratory events, the desaturation area associated with respiratory events may be overestimated due to incomplete recovery from prior desaturation or non-OSA-induced hypoxemia [173]. To improve the accuracy of respiratory event-related desaturation severity estimation, Kulkas et al. introduced obstruction severity (ObsSev), later renamed sleep breathing impairment index (SBII) by Cao et al., which links DesSev to hypopnoea and apnoea events, as shown in Equation 3.4:

$$ObsSev = SBII = \frac{\sum_{n=1}^{number\ of\ Hyps} HypDur \times Desaturation\ area_n + \sum_{n=1}^{number\ of\ Aps} ApDur \times Desaturation\ area_n}{\sum time\ of\ sleep}, \tag{3.4}$$

where Hyps is the number of hypopnoea events, Aps is the number of apnoea events, HypDur is the duration of a single hypopnoea event, and ApDur is the duration of a single apnoea event. The unit of ObsSev (SBII) is %seconds [29, 188].

It has been suggested that as ObsSev (SBII) captures more respiratory event information than other conventional SpO2 parameters, it may better predict OSA-related CVD outcomes [29, 188]. Investigators have also found that ObsSev (SBII) is more age-related than AHI, and therefore can be used to estimate long-term CVD progression [189].

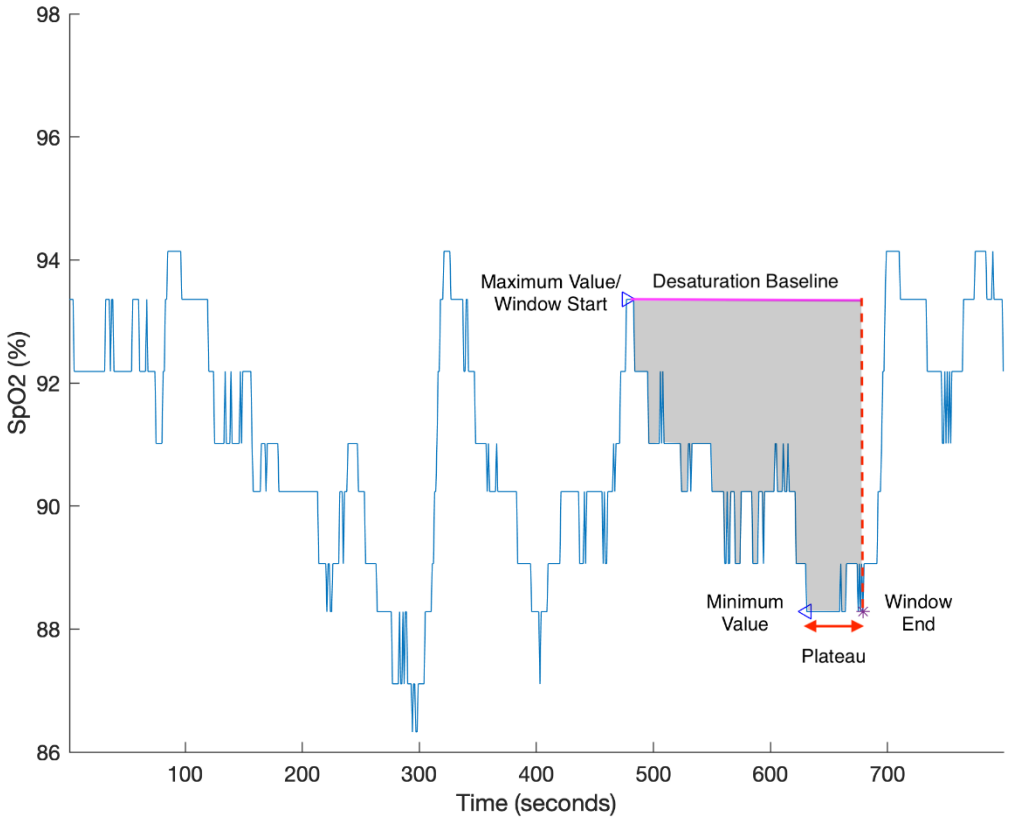


Figure 3.4 The example of DesSev calculation. DesSev is calculated as the sum of the desaturation area divided by the total time of sleep associated with automated desaturation event detection algorithm. The desaturation area (grey) is calculated as the area between the desaturation baseline (magenta) and the SpO2 trace (blue) within the sampling window. The sampling window is to be defined as the time between the maximum and minimum values (blue triangles). However, due to the presence of a plateau (red arrow), the end of the sampling window is shifted forward to the end of the plateau (purple star). The desaturation baseline is the maximum value [176].

3.1.3.5 Hypoxia Load (HL)

HL differs from other desaturation area-based parameters, as it is independent of any desaturation threshold or respiratory events. As shown in **Figure 3.5**, HL is defined as the desaturation area above the SpO2 trace divided by the total time of sleep. The calculation of HL can be divided into two steps: (1) The SpO2 saturation area is calculated by the numerical integration of the SpO2 trace using the trapezoidal rule, as shown in Equation 3.5:

$$\int_0^{total\ time\ of\ sleep} Area_{SpO2} \approx \sum_{n=1}^{N} \frac{(SpO2_n + SpO2_{n+1}) \times (t_{n+1} - t_n)}{2}, \quad (3.5)$$

where SpO2n and SpO2n+1 are successive samples of the SpO2 trace, and the unit of the integrated area is %seconds [167]. (2) The average saturation is then calculated by dividing this area by the sleep time. HL is calculated using Equation 3.6. It is worth noting that the average saturation values in HL are also most exactly equal to the average SpO2 value during sleep reported by most PSG analysis software packages, the only difference being that the trapezoidal rule is used to calculate the area for HL whereas the average value use the rectangular rule to calculate the area.

$$HL = 100\% - \frac{\text{saturation area during sleep}}{\sum time\ of\ sleep}$$

$$= 100\% - average\ SpO2\ value\ during\ sleep, \quad (3.6)$$

The unit of HL is % [167]. Because HL, unlike all previously discussed parameters, is not affected by any desaturation event criteria and thresholds, it is less likely to be miscalculated, which is ideal for cross-study analysis. However, HL has limitation as the information revealed by HL is not specific to transient changes in the SpO2 and may down-play the importance of OSA-related oxygen transients.

HL has not been widely used in the prediction of CVD events, but its association with other OSA-related symptoms has been explored. HL has been shown to be an independent predictor of fasting blood glucose and haemoglobin A1c levels [190]. Linz et al. demostrated that HL is also significantly correlated with CVD indicators in OSA patients after acute myocardial infarction, whereas AHI and other traditional metrics are not [167]. Similarly, Khoshkish et al. found a strong association between HL and blood pressure, while the conventional metrics of hypoxemia do not present such an association. Although the correlation between nocturnal systolic blood pressure and HL became insignificant after adjusting for BMI, HL was strongly associated with the pulse pressure before and after the adjustment for BMI. It was suggested that HL is a suitable marker of blood pressure patterns [191, 192]. Considering that

hypertension and diabetes are both associated with OSA, future studies may reveal associations between OSA and CVD [193-196].

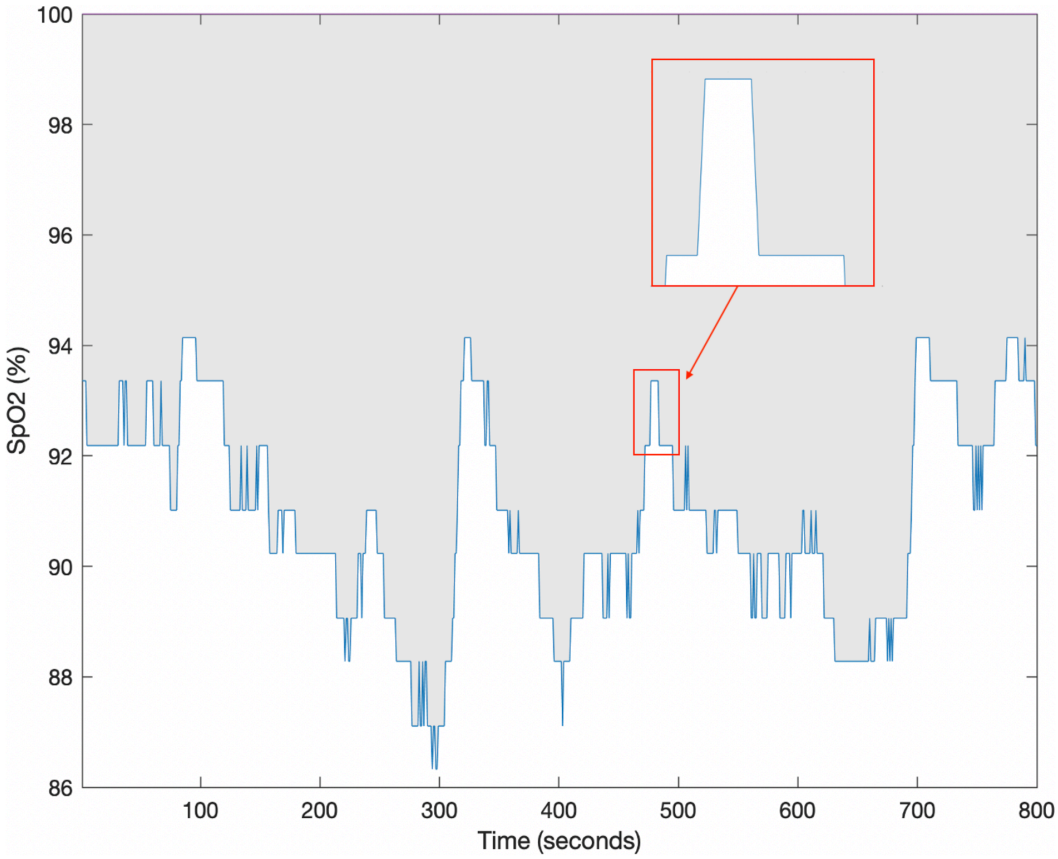


Figure 3.5 The example of HL calculation. HL is calculated as the integrated area (grey) between the 100% baseline and the SpO2 trace divided by the total time of sleep [167]. The inset demonstrates the use of the trapezoid rule for determining the area.

3.1.4 Summary of novel oximetry-derived parameters and their performance in predicting cardiovascular disease (CVD) events

Current studies have introduced various oximetry-derived parameters as potential predictors of CVD outcomes. However, no single parameter has demonstrated predictive capability across all event types. **Table 3.2** summarises the performance of these novel parameters, with statistically significant predictors of CVD outcomes listed in the second column and non-significant predictors in the third column.

As shown in **Table 3.2**, HB, T90desaturation, and REDTA have been identified as the most effective predictors of CVD mortality [6, 12, 28]. In contrast, T90non-specific and ODI4 exhibit non-significant hazard ratios in predictive analysis [6]. TST90 has shown strong

predictive performance for MACEs, hypertension, diabetes, and endothelial dysfunction, many of which are recognised as risk factors or precursors to CVD events [134, 145, 197]. T90% outperforms other parameters in predicting right ventricular stroke, metabolic syndrome in women, and incident CVD in patients with non-sleepy SDB, while T90non-specific is particularly useful in predicting pulmonary hypertension [14, 135, 136, 198]. Although Karhu et al. suggested that ODI4 is a more reliable metric than ODI3, more recent studies indicate that ODI3 is superior in predicting coronary plaque burden, OSA severity, and specific CVD risks [6, 158, 161, 199, 200]. HB demonstrates strong performance in predicting stroke incidence, incident heart failure, and MACEs compared to other desaturation area-based parameters [132, 171, 174]. Unlike HB and REDTA, DesSev and ObsSev/SBII are independent of scored respiratory events and have been used to predict excessive daytime sleepiness, HRV, mean daytime sleep latency, acute stroke, transient ischaemic attack, and CVD morbidity [131, 178, 180, 181, 201]. Few studies have examined HL, though it has been shown to predict HbA1c levels and blood pressure in patients with sleep disorders, which may assist in diagnosing diabetes and hypertension [190, 191].

Despite the promising results of these parameters in predicting CVD events, most fail to provide consistent predictive value in OSA patient populations. Sutherland et al. and Linz et al. reported that TST90, T90%, ODI3, ODI4, and HB had non-significant predictive value for incident CVD in OSA patients, and composite CVD outcomes in OSA patients with high CVD risks, respectively [148]. This inconsistency in predictive performance may be attributed to several factors. First, differences in patient populations can significantly impact parameter performance. The SHHS dataset, for example, is drawn from a community-based sample, which has a much greater representation on controls than clinical samples or OSA-only samples. It is perhaps unreasonable to expect a particular parameter to perform well across these distinct populations. Theoretically the most useful populations to study these parameters in are clinical populations, because these are patient populations (by definition) seen in clinical practice. Second, variations in baseline comorbidities and treatment regimens must be accurately accounted for in studies. Without comprehensive documentation of these factors, differences in patient health status may influence parameter performance. Third, the lack of standardised and reproducible definitions for these parameters affects their predictive reliability. Differences in computational implementations across studies may lead to inconsistent results.

Table 3.2 Summary of AHI and oximetry-derived parameters and their performance in the prediction of CVD outcomes.

Parameter	Is a Predictor of	Is Not a Predictor of
AHI	All-cause mortality [202]	Incident heart failure [132] Composite CVD outcomes * [141] Hypertension [122]
TST90	MACE [145] Hypertension [197] Diabetes [197] Endothelial dysfunction [134]	Composite CVD outcomes ** [199]
T90%	Incident CVD in patients with non-sleepy SDB [14] Right ventricular stroke [135] Metabolic syndrome in women [136]	Incident CVD in patients with OSA only [148]
T90desaturation	CVD mortality [6]	
T90non-specific	Pulmonary hypertension [198]	CVD mortality [6]
ODI3	Coronary plaque burden [161] Severity of OSA [158] CVD risks *** [200]	Incident CVD in patients with OSA only [148]
ODI4		CVD mortality [6] Composite CVD outcomes * [199]
HB	CVD mortality [12]; Stroke incidence [174] MACE **** [171]; Incident heart failure [132]	Incident CVD in patients with OSA only [148]

Parameter	Is a Predictor of	Is Not a Predictor of
REDTA	CVD mortality [28]	
DesSev	Excessive daytime sleepiness [179] HRV [178]	
ObsSev/SBII	Mean daytime sleep latency [181] Acute stroke and transient ischemic attack [201] CVD morbidity [131]	
HL	haemoglobin A1c level [190] Blood pressure [191]	

* Composite CVD outcomes include myocardial infarction, congestive heart failure, stroke, revascularization procedure, or death from any cause. ** Composite CVD outcomes (CVD death, stroke, myocardial infarction, heart failure, angina, transient ischemic event) in OSA patients with high CVD risks. *** CVD risks: the risks of CVD and CVD events including hypertensive disease, ischemic heart disease, cerebrovascular disease, diseases of arteries, arterioles, capillaries, and congestive heart failure. **** MACE: a composite outcome including all-cause mortality, acute myocardial infarction, stroke, and unplanned coronary revascularization.

3.1.5 Limitation of published algorithms and detailed motivation of Experiments

A major limitation in current research is the variation in computational algorithms and parameter definitions, which hinders cross-study comparisons. The absence of standardised computational approaches and discrepancies in AASM scoring criteria impact parameter performance in predicting CVD events. Some studies rely on commercial software for parameter calculations, which often lacks transparency and validation due to the unclear criteria for calculation and variations in data processing techniques [153].

To address these limitations, Experiment 1 of this thesis was designed to investigate how different computational approaches of parameters influence the predictive performance for CVD mortality outcomes. The primary focus of Experiment 1 is on desaturation area-based parameters, as these are the most affected by computational variability. Additionally, each published desaturation area-based parameter has specific limitations related to calculation complexity and application in large datasets, which will be discussed in detail in the following chapters. Experiment 1 aims to address these challenges and identify the most effective predictor for CVD mortality.

Furthermore, most of current studies focus on the predictive performance of individual parameters, without exploring their combined utility. Each parameter captures only a partial representation of oximetry information, limiting its ability to fully characterise sleep disturbances. For instance, T90 focus on the duration of exposure to hypoxemia, whereas desaturation area-based parameters measure the transient hypoxemia associated with desaturation events. Until the complete understanding of the role of hypoxia in impacting CVD outcomes, a multivariate approach to CVD outcome prediction, which includes a range of oximetry and other PSG-derived parameters, will likely be a more successful approach than focusing on one particular parameter [15]. This rationale forms the basis of Experiment 2, which will be described in detail in the subsequent chapters.

Motivated by the above considerations, Experiment 1 systematically selects and characterises three existing desaturation area-based algorithms, analysing how variations in event selection, sampling window, and baseline calculation influence the performance of parameters in predicting CVD mortality outcomes. It further identifies an algorithm that demonstrates

robustness and scalability for large-scale datasets. Experiment 2 proposes an explainable machine learning model to predict CVD mortality outcomes and investigates whether incorporating PSG-derived parameters as multivariate inputs enhances predictive performance. The goal is to enable accurate, individual-level prediction of CVD mortality outcomes using input features that require minimal clinical resources, thereby making the approach suitable for widespread application in the general population.

3.2 Additional Polysomnogram (PSG)-derived parameters and medical information

To further support the motivation behind both experiments, additional PSG-derived parameters and medical information were incorporated into the analysis, as summarised in **Table 3.3**.

Total sleep time (TST), derived from EEG recordings, has been shown to be associated with adverse CVD outcomes, including stroke, incident CAD, and total CVD risk. The relationship between TST and CVD outcomes follows a U-shaped pattern, suggesting that both short and long sleep durations are linked to different CVD events. P. Cappuccio et al. summarised that short TST is associated with a higher risk of CAD mortality and stroke, whereas long TST is linked to an increased risk of total CVD [203]. Minimum SpO2 during TST (MinSat), though rarely considered in CVD analysis, has shown potential associations with CVD development. Gunnarsson et al. proposed that MinSat is an independent predictor of future carotid plaque burden, a well-established indicator of subclinical arterial disease capable of predicting CVD events [204].

This thesis also incorporates demographic information in the predictive analysis, including age, BMI, race, gender, smoking status, and alcohol intake, as these are traditionally well-established predictors of CVD outcomes. According to Damen et al., the most common predictors in CVD models are age and smoking status, with many models being gender- and race-specific [205].

Additionally, participants' medical history was considered to ensure a less biased analysis. In Experiment 1, medical history and other parameters were included as covariates, which were adjusted during the performance evaluation process to eliminate their confounding influence in predicting CVD mortality. In Experiment 2, some medical history variables were used as

features to assist in predicting CVD mortality within a specific timeframe. Medical history was categorised into three types:

1. Diseases coexisting with CVD – Chronic obstructive pulmonary disease (COPD) and CVD frequently co-occur, with up to 70% of CVD patients also having COPD. COPD has been linked to an increased risk of CVD events [206, 207].
2. Doctor-reported CVD symptoms, events, and procedures – Angina (chest pain) is strongly associated with future CVD events and was included in the analysis as a predictor [208].
3. CVD risk factors – Conditions such as diabetes, hypertension, and abnormal cholesterol levels were incorporated into the analysis, as they are precursors of CVD [86].

Table 3.3 Summary of other PSG-derived parameters/ medical information used in this thesis and their association with CVD events.

Type	Parameters	Associated with
EEG-derived	TST	Stroke [203] CAD morbidity and mortality [203] CVD [203]
Oximetry-derived	MinSat	Future carotid plaque burden [204]
Demographic information	Age, BMI, Race, Gender, Smoking status, and Alcohol intake.	independent predictor of CVD [205]
Medical history	History of COPD, Angina, Stroke, Heart failure, Myocardial infarction, Coronary artery bypass graft, Coronary angioplasty, Diabetes, Hyperlipidaemia and Hypertension.	COPD: Co-exist with CVD [206] Symptoms of CVD: Angina [208] CVD events: Stroke, Heart failure, and Myocardial infarction CVD procedures: Coronary artery bypass graft and Coronary angioplasty CVD risk factor: Diabetes, Hyperlipidaemia and Hypertension [86]

3.3 Cox proportional hazards analysis

The evaluation tool for Experiment 1 is the Cox proportional hazards model (hereafter referred to as the Cox model), a subtype of survival analysis used to assess the relationship between target predictors and the risk of an endpoint event under a proportional assumption. Experiment 1 applies the Cox model using desaturation area-based parameters as predictors, with the model adjusted for covariates, including other PSG-derived parameters and medical information, to estimate the risk of CVD mortality outcomes over time.

To establish a strong statistical foundation for readers, this section begins with a brief introduction to survival analysis, covering some key concepts as a basis for understanding the Cox model. The discussion of the Cox model will include:

1. Definition of the model
2. Key assumptions for the model
3. Advantages compared to other inferential approaches
4. The necessity of incorporating covariates
5. Example applications of the model in OSA-CVD analysis
6. Potential risks of using the model

3.3.1 Survival analysis

Survival analysis is a family of statistical methods used to examine the relationship between a designated exposure and the occurrence of an outcome within a specified time period [209]. While commonly applied to mortality data, survival analysis is broadly used for all time-to-event outcomes. Two of the most frequent applications in medical research include the time-to-death and relapse-free survival time (or disease-free survival time), which measures the duration between a patient's response to treatment and the recurrence of symptoms [210].

Since survival analysis evaluates time-dependent risks, defining the starting point and endpoint is essential. The starting point is not pre-standardised and can vary, particularly in observational studies. The definition of "time zero" depends on study design. For instance, in survival analysis for CAD patients, the starting point could be the first day the patient self-reports angina, the date of an official CAD diagnosis by a physician, or the detection of significant coronary artery stenosis [209]. In contrast, the endpoint is typically well-defined, referring to the date of the target outcome, such as death, symptom relapse, or disease progression.

An important concept in survival analysis is censored observation, which occurs when the target outcome is not observed for some patients before the study ends. Censoring typically arises in the following conditions: the patient does not experience the target outcome (e.g., remains disease-free or alive) by the end of the study period; the patient is lost to follow-up due to withdrawal from the study; the patient experiences an unrelated event that prevents further follow-up or affects the ability to observe the target outcome. Censoring can be categorised into left censoring and right censoring, with right censoring being more common in the survival analysis. Right censoring occurs when the event of interest has not been observed by the end of the follow-up period. For example, in a 10-year mortality study, if a participant remains alive beyond the 10-year mark, the outcome is considered as right censoring. Left censoring occurs when the event of interest has already occurred before the start of follow-up period, but the exact time of occurrence is unknown. For instance, in tracking an infectious disease, if the study's starting point is the first positive test for infection, but the actual date of exposure to the infectious agent is unknown, this scenario qualifies as left censoring [211].

Compared to other time-related statistical models, such as simple proportion models, survival analysis is particularly advantageous due to its ability to handle censored observations. Simple proportion models fail to account for varying follow-up durations, making them less effective in longitudinal studies. In contrast, survival analysis accommodates censored data, allowing for partial observations while maintaining an unbiased time-to-event estimation. This capability makes survival analysis particularly useful in medical research involving cancer, CVD, and OSA.

3.3.2 Key concepts used in survival analysis

Survivor function represents the probability that an individual remains alive (or event-free) beyond a specified time point, as shown in Equation 3.7:

$$S(t) = \Pr(T > t) \quad \text{where } 0 \leq t < \infty \tag{3.7}$$

where *t* represents the time of interest, S(t) represents the probability of survival beyond time *t*, *T* denotes the patient's lifetime. Since *t* ranges from 0 to infinity, *S(t)* is always a positive value between 0 and 1. As time progresses, *S(t)* can either remain constant or decrease, but it can never increase [212]. The survivor function is a fundamental component in the interpretation of survival data and is commonly estimated using the Kaplan–Meier estimator.

Hazard function quantifies the instantaneous risk of a target event at a specific time point during the follow-up period, as shown in Equation 3.8:

$$h(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t+\Delta t)}{\Delta t} \tag{3.8}$$

where *t* is the specific time which the hazard rate is calculated and Δt represents a small increase in time *t* [213]. The hazard function reflects the rate of decline in the survivor function over an infinitesimal time interval, essentially capturing the instantaneous failure rate at time *t*. Unlike the survivor function with probability ranging from 0 to 1, *h(t)* is a rate, and its value depends on the unit of time used in the analysis. The relationship between the survivor function and the hazard function is defined through the cumulative hazard function, which is the integral of the hazard function over time. Mathematically, the survivor function *S(t)* is expressed as the exponential of the negative cumulative hazard, as detailed in Equation 3.9:

$$S(t) = \exp\big(-H(t)\big) = \exp\left(-\int_0^t h(u)du\right) \tag{3.9}$$

where *H(t)* is the cumulative hazard function and *h(t)* represents the instantaneous risk [214]. Equation 3.9 indicates that a higher hazard rate *h(t)* results in a faster decline of the survival curve *S(t)*. In other words, increased instantaneous risk leads to lower survival probabilities over time.

Given two covariate profiles x_0 and x_1, the hazard ratio (HR) measures the comparative instantaneous risk ratio between two hazard functions, as shown in Equation 3.10:

$$HR = \frac{h(t|x_1)}{h(t|x_0)} \tag{3.10}$$

Under the Cox proportional hazards model, the hazard function is specified as Equation 3.11:

$$h(t|x) = h_0(t) * \exp\left(\beta^T x\right) \tag{3.11}$$

where $h_0(t)$ is an unspecified baseline hazard, which all variables are 0 at baseline, and β^T denotes the vector of regression coefficients. Substituting Equation 3.11 into Equation 3.10 yields

$$HR = \exp\left(\beta^T (x_1 - x_0)\right) \tag{3.12}$$

which is constant over time when the proportional hazards assumption holds. The hazard function, *h(t),* describes the instantaneous risk of an event occurring at time *t* within a single group or individual, whereas the HR represents the relative risk of the event between two groups, typically a target group and a control group [211].

HR and survival curves are two of the most used tools in medical research, each serving distinct purposes. Survival curves, derived from the survivor function, provide a graphical overview of survival trends over time. In contrast, the HR quantifies the group differences on the risk of an event occurring and further reveals the relative effect of covariates. HR is typically estimated using the Cox model and is widely employed to assess the relative impact of variables on event risk in survival analysis. In this thesis, the predictive performance of desaturation area-based parameters will be evaluated by comparing the HRs between groups defined by parameter values. Greater differences in hazard between groups indicate that the parameter more effectively distinguishes patients with different survival risks, thereby implying stronger predictive ability of parameters. The detailed methodology for calculating HR, as well as a comparison between the Cox model and other estimating approaches, will be presented in the following sections.

3.3.3 Cox proportional hazards model

As the aim of Experiment 1 is inferential: comparing survival differences between groups defined by parameter values, methods focusing on comparative analysis are more appropriate. The Kaplan–Meier estimator, while a valuable descriptive tool in survival analysis, is not suited to this experiment because it visualises overall survival trends but does not formally test group differences or quantify effect sizes. In contrast, the Cox model is well-suited for this purpose. As a semi-parametric and inferential method, it not only evaluates for differences between groups but also estimates HRs, thereby quantifying the relative impact of parameters on survival. In addition, the model can incorporate multiple covariates without requiring strict distributional assumptions about the baseline hazard.

3.3.3.1 Definition of the model

The Cox model was introduced by Sir David Cox in 1972 as a method for estimating differences in survival attributable to independent variables. As a regression-based approach, it has become the most widely used method among inferential statistical techniques in survival analysis [215, 216]. The Cox model is based on hazard function, which can be expressed as Equation 3.13:

$$H(t) = H_0(t) \times \exp\left[b_1 x_1 + b_2 x_2 + \cdots b_k x_k\right] \tag{3.13}$$

where $H(t)$ represents the hazard at time *t*, $H_0(t)$ is the baseline hazard when all variables are zero, b_k indicates the regression coefficient and x_k denotes the predictor and covariate variables. By estimating the regression coefficients, one can compute the HR for a given

predictor as exp (b_k), along with its corresponding confidence interval. Mathematically, both predictors and covariates are represented as x_k in the model. However, they differ in interpretation depending on the study objective. Specifically, regression coefficients for predictors are interpreted as measures of their independent association with the event of interests, while covariates are included to adjust for potential confounding and help isolate the true impact of the target predictor on the event of interest.

3.3.3.2 Key assumptions for the model

There are four key assumptions underlying the Cox proportional hazards model [217, 218]:

1. Proportional hazards assumption: The model assumes that HRs associated with a given predictor remain constant over time. In other words, the effect of a covariate on the hazard is multiplicative and does not vary with time.
2. Semi-parametric structure: The Cox model is semi-parametric, meaning it does not require specification of the baseline hazard function. This flexibility allows the model to estimate hazard ratios without assuming a particular form for the hazard over time.
3. Independence of survival times: The model assumes that survival times are independent between groups or individuals. That is, the survival time of one subject or group does not influence the survival time of another.
4. Non-informative censoring: If censored data are included, censoring must be non-informative, meaning the reason for censoring is unrelated to the likelihood of experiencing the event of interest. More specifically, censored individuals are assumed to have the same risk of the event as those who remain in the study, had follow-up continued.

Before applying the Cox model in survival analysis, each variable should be evaluated for compliance with the proportional hazard assumption, and the event-to-time data should be examined to ensure that censoring is non-informative.

3.3.3.3 Advantages compared to other inferential approaches

In addition to the Cox model, other inferential strategies are available for survival analysis, such as the log-rank test. Also known as the Mantel–Cox test, the log-rank test is a non-parametric method used to evaluate the null hypothesis that there is no difference in survival between groups, that is, the hazard functions are equal over time. Although it does not estimate HRs like the Cox model, the log-rank test also performs under the proportional hazard

assumption, where the relative risk between groups is assumed to remain constant throughout the observation period [219].

Although the log-rank test is suitable for detecting survival differences between groups, it has important limitations compared to the Cox model, particularly for analysis such as Experiment 1. First, the log-rank test evaluates only one variable at a time, whereas the Cox model allows for the inclusion of multiple covariates, helping to control for confounding effects. Second, unlike the Cox model, the log-rank test does not accommodate multiple categorical or continuous variables (e.g. age, BMI), which are commonly used predictors in the CVD studies. Most importantly, the log-rank test is non-parametric and does not assume any mathematical form of survivor and hazard functions, simply comparing the observed events and the expected events. Thus, it can detect the present of differences in survival between groups, but cannot quantify the effects through hazard ratios, as the Cox model does. Additionally, the log-rank test is most effective when survival curves are clearly separated, a scenario that is uncommon in many CVD-related survival studies [219, 220].

For these reasons, the Cox model is the more appropriate evaluation method for Experiment 1, offering greater flexibility, interpretability, and control of confounders compared to other inferential approaches.

3.3.3.4 The necessity of incorporating covariates

As mentioned in the previous section, one of the key advantages of the Cox model is its ability to incorporate covariates. This concept is particularly important in comparative analyses that assess the predictive ability of specific parameters, such as Experiment 1 in this thesis. The inclusion of covariates allows the model to control for confounding effects, which is essential for producing valid and unbiased estimates.

In predictive analyses, confounders are variables that are associated with both the target predictor and the outcome of interest, potentially distorting the observed relationship between them. For example, when examining the association between the severity of OSA and CVD mortality, age acts as a key positive confounder, as it influences both the development of OSA and the risk of CVD mortality. As age increases, the severity of OSA tends to worsen, since OSA is more common in older populations, and the risk of CVD mortality also rises. If age is not accounted for in the model, the HR may be overestimated, falsely attributing the effect of

age to the predictor. Conversely, the HR could be underestimated if a confounder exerts a negative effect on the predictor but a positive effect on the outcome. In both cases, the confounder's influence leads to a biased estimation of the HR, affecting the reliability of the conclusions drawn about the predictor's value in forecasting outcomes.

By including relevant covariates in the analysis, the Cox model helps adjust for these confounding effects, enabling a more accurate and isolated assessment of the association between the target predictor and the outcome of interest.

3.3.3.5 Example applications of the model in OSA-CVD analysis

In addition to the theoretical justification for using the Cox model with covariates in Experiment 1, numerous published studies have employed the same modelling approach to evaluate the predictive ability of OSA-related parameters for CVD outcomes. For example, Baumert et al. used the Cox model to investigate the association between T90 and CVD mortality in community-dwelling older men. Their model was adjusted for anthropometric characteristics, lifestyle factors, and medical history, and the study concluded that T90 is an independent predictor of CVD mortality [6]. Similarly, Azarbarzin et al. applied the Cox model, with similar covariate adjustments, to assess the predictive value of HB for CVD mortality. Their analysis, conducted across both the SHHS and the MrOS cohorts, demonstrated that HB is a robust predictor of CVD mortality [12].

These examples, along with others, further support the appropriateness of using the Cox model in Experiment 1 of this thesis, particularly when aiming to isolate the independent predictive value of desaturation area-based parameters for CVD mortality outcomes while adjusting for confounding variables.

3.3.3.6 Potential risks of using the model

Despite the advantages of the Cox model and its widespread use in medical research, it has several limitations that require careful adjustments of variables in analysis. First, the Cox model assumes a proportional hazards framework, meaning it estimates a constant difference in HR between groups over time. However, if the impact of a predictor varies with time, this assumption may not hold, leading to inaccurate estimates [221]. Second, the Cox model assumes that predictors and covariates are independent. When variables are highly correlated, the model produces unstable estimates, reducing the reliability of the analysis. Moreover, the model assumes a log-linear relationship between predictors and the hazard function. If

predictors exhibit a nonlinear association with survival, the estimated hazard risks may be misleading [222]. To address these limitations, several strategies can be applied: considering stratified Cox model, applying principal component analysis to mitigate collinearity among highly correlated inputs, and performing variable transformations (e.g., log transformation) to improve the linearity between predictors and hazard risks. These adjustments help improve the robustness and accuracy of the Cox model estimations, all of which can be considered in the Experiment 1.

3.4 Machine learning

In Experiment 2, machine learning methods are employed to predict CVD mortality at a given time, in contrast to the traditional statistical approach, the Cox model used in Experiment 1. This methodological shift is driven by both the limitations of the Cox model and the broader scope of Experiment 2, which aims to predict individual-level outcomes using multivariate approaches for enhanced predictive accuracy.

While the Cox model is widely used in survival analysis, it has several constraints in the context of Experiment 2. First, it requires strict assumptions, such as the proportionality of hazard functions and the independence of survival times. In addition, the Cox model primarily provides relative effect estimates in the form of hazard ratios, rather than absolute risk probabilities that may be more directly interpretable for individual-level risk assessment. Finally, the model focuses on assessing the independent effect of each predictor. To capture the combined or interactive effects of multiple variables, explicit interaction terms must be included, which can substantially increase the complexity of the model and reduce interpretability [18].

Accordingly, while the continued value of traditional survival models for time-to-event analysis is fully acknowledged, machine-learning approaches are considered in Experiment 2 as a complementary methodological framework to support individual-level outcome prediction. Previous studies have demonstrated that machine learning approaches can perform at least as well as the Cox model in predictive analysis, particularly in complex, multivariate contexts [21-26]. Machine learning methods offer several advantages that align with the aims of Experiment 2. They can accommodate interactions among predictors, enabling the assessment of both independent and combined predictive contributions. While interactions among variables are more readily captured by flexible machine learning models, the models can be

interpreted by explainability tools as well (discussed in Section 3.4.6.5). In addition, these approaches can generate absolute risk estimates at defined time horizons, supporting more patient friendly individual-level risk assessment.

3.4.1 Supervised learning vs unsupervised learning

Machine learning techniques are generally categorised into supervised and unsupervised learning, with the primary distinction being the use of target outcome labels in the training data [223]. Unsupervised learning operates labelled outcomes-free and is typically used for exploratory or descriptive analysis. It focuses on identifying underlying patterns, groupings, or structures within the input data, without prior knowledge of the target labels. In contrast, supervised learning relies on labelled datasets, where the model learns to predict the outcome of interest based on input features. In other words, supervised learning establishes a functional relationship between input variables and known outcomes. Once trained, the model can use this relationship to predict or forecast outcomes for new, unseen data, making it particularly suitable for the predictive objectives of Experiment 2 in this thesis [224].

Supervised machine learning is designed to address classification and regression problems. In classification tasks, the model maps input features to predefined class labels, while in regression tasks, the model maps inputs to a continuous value domain [225]. Experiment 2 in this thesis is framed as a classification problem, where the CVD mortality outcome is defined as a binary label (e.g., event vs no event). The model uses PSG-derived parameters, demographic information, and medical history as input features. The classification process consists of two key phases: training and testing [225]. During the training phase, the model is built by feeding both features and corresponding outcome labels into a learning algorithm, which seeks to establish a functional relationship between inputs and the target outcome [226]. In the testing phase, the trained model is applied to unseen data to predict the individual probability of the outcome occurring for each participant. Based on the estimated probabilities, the performance of the learning model and the combined predictive ability of input features can be assessed using the confusion matrix. Additionally, the independent contribution of each feature to the predictions can be interpreted using SHAP.

Commonly used supervised learning methods in medical research are summarised in **Figure 3.6**, encompassing both single models and ensemble learning techniques. Single models include approaches such as linear models, K-Nearest Neighbours, Decision Trees, Support

Vector Machines (SVM), and Neural Networks. These models operate independently and apply distinct strategies for classification. In contrast, ensemble learning combines multiple models to improve predictive performance and robustness. This category includes methods based on bagging and boosting, with widely used examples being Random Forest (RF) and Extreme Gradient Boosting (XGBoost).

A selected subset of machine learning models is applied in this thesis, with the rationale for their inclusion explained in detail in the following sections. The primary principles guiding model selection are effectiveness and explainability, meaning that the chosen models are both well-suited for data in Experiment 2 and interpretable by end-users, including clinicians and patients. Although previous studies have demonstrated the strong predictive performance of machine learning models in medical applications, many treat the models as uninterpretable "black boxes", particularly in the case of neural networks [18]. This lack of transparency can pose a significant barrier to clinical trust and adoption, as clinicians are often cautious about relying on predictions when the decision-making process is unclear [18]. Considering the ultimate goal of this thesis is to contribute to clinical practice and to early risk stratification in OSA and CVD patients, it is essential to prioritise explainability in models.

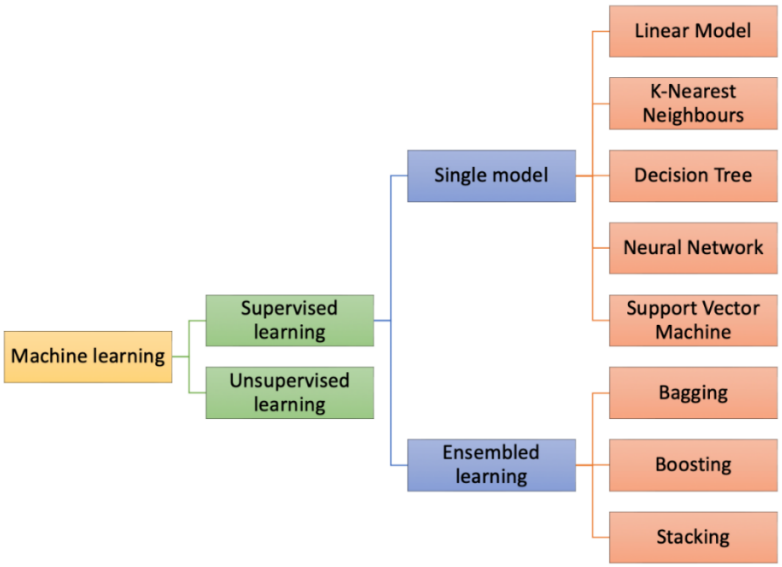


Figure 3.6 Summary of commonly used machine learning models and highlights their key differences. Not all methods are covered in this thesis, as some are not well suited to the scope of Experiment 2.

3.4.2 Linear Discriminate Analysis (LDA)

LDA is a well-established linear dimensionality reduction technique that seeks to identify a linear combination of input features that best separates distinct classes [227]. In the context of Experiment 2, LDA aims to distinguish between CVD mortality outcomes using linear combinations of PSG-derived and clinical features. LDA has been widely applied in fields such as face recognition, bioinformatics, and medical signal processing, due to its simplicity, interpretability, and effectiveness in reducing dimensions and extracting discriminative features [228, 229]. For a given set of features X and class labels C, the goal of LDA is to find the projection vector ω that maximises the Fisher criterion:

$$J(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega} \tag{3.14}$$

where ω is the direction onto which the data is projected for maximum class separability, S_W is the within-class scatter matrix and S_B is the between-class covariance of X. The equations of S_B and S_W are calculated as follows:

$$S_B = \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{3.15}$$

$$S_W = \sum_{i=1}^{C} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T \tag{3.16}$$

where μ_i is the averaged vector of class i, μ is the global mean vector across all samples, N_i is the number of samples in class i, and C_i is the set of samples in class i [227]. Intuitively, S_W captures the compactness: how tightly clustered the samples are within each class, while S_B measures separability: how far apart the class means are. To achieve better class discrimination, LDA seeks to maximise between-class variance (S_B), while minimising within-class variance (S_W). In scenarios where the input data is imbalanced, one way to improve the performance of LDA is by introducing class weighting. This technique adjusts the influence of each sample or class when calculating the scatter matrix, thereby reducing bias toward the majority class. The weighted within-class and between-class scatter matrices are defined as:

$$S_W^{weighted} = \sum_{i=1}^{c} \sum_{x \in C_i} \omega(x)(x - \mu_i)(x - \mu_i)^T \tag{3.17}$$

$$S_B^{weighted} = \sum_{i=1}^{c} \omega(C_i)(\mu_i - \mu)(\mu_i - \mu)^T \tag{3.18}$$

where $\omega(x)$ is the weight to sample x and $\omega(C_i)$ is the weight to class C_i.

The use of LDA relies on one essential condition and three key assumptions to ensure reliable class separation. The essential condition is the number of samples in the smallest class should be greater than the number of input features. If this condition is not met, the S_W may become

singular (non-invertible), leading to instability or overfitting during model training [230]. Key assumptions for optimal LDA performance are listed as follows [230, 231]:

1. Multivariate normality: The input features within each class are assumed to follow a multivariate Gaussian distribution. This supports the generative modelling approach of LDA.
2. Homoscedasticity (equal covariance): The variance–covariance structure of the features is assumed to be the same across all classes. That is, LDA assumes a common within-class covariance matrix.
3. Independence of observations: Each observation (row of features) is assumed to be independently drawn. This assumption ensures that samples do not introduce dependency biases into the model.

In practice, although LDA is sensitive to outliers and violations of its underlying assumptions, it often performs robustly under mild deviations from ideal conditions [232]. However, to further enhance its performance and broaden its applicability, several variants of LDA have been proposed. For instance, Kumar and Andreou introduced Heteroscedastic Discriminant Analysis, a maximum likelihood approach addresses cases where the Gaussian-distributed classes are not proportional or equal in their variances [233]. To accommodate clustered data, Mixture Discriminant Analysis was developed, incorporating a Gaussian mixture model to better capture complex data distributions [234]. In addition, regularisation techniques have been applied to LDA to reduce overfitting and improve generalisability, particularly in high-dimensional settings [235, 236]. Furthermore, combining LDA with Principal Component Analysis has been shown to enhance predictive performance, especially when dealing with highly correlated or redundant features [237].

In this thesis, LDA with class weighting is employed as a benchmark model for the initial evaluation of feature combinations and serves as a baseline for performance comparisons across various machine learning models.

The key difference between LDA and the other machine learning approaches in this thesis lies in their modelling philosophy: LDA is a generative model, whereas the others are discriminative. Generative models, such as LDA, attempt to model the distribution of input features using Bayes' Rule. They estimate the feature distributions within each class and then compute the probability of a given sample belonging to a particular class. In contrast,

discriminative models focus solely on learning the decision boundary between classes. They model the conditional probability of the class label given the input features and predict the class label directly, without assuming any specific distribution of the features [238]. The shift toward discriminative models in this thesis is motivated by their greater robustness to assumption violations, such as skewed data distributions or non-linear relationships, and their higher tolerance to outliers, making them particularly well-suited to the predictive objectives of Experiment 2.

3.4.3 Support Vector Machine (SVM)

SVM is a well-established discriminative machine learning model that aims to maximise the margin between classes [239]. Traditional statistical learning methods often rely on empirical risk minimisation, which focuses on reducing error on the training data. However, this approach may fail to accurately estimate the expected risk (i.e. the error on unseen data), resulting in poor generalisability [240]. To improve the model performance, SVM is grounded in statistical learning theory and follows the principle of structural risk minimisation, which explicitly balances empirical risk and model complexity. This foundation gives SVM a strong ability to generalise, particularly in high-dimensional or limited-sample datasets [241]. Owing to these advantages, SVM has become a widely used and powerful tool for classification tasks in medical research, including applications in diagnosis, prognosis, predictive modelling, and risk assessment. These characteristics make SVM a suitable choice for Experiment 2 in this thesis [242].

SVM employs a hyperplane (decision boundary) to separate training data into distinct classes. The complexity of the classifier is constrained by the margin, which is the minimum distance between the hyperplane and the nearest training examples, referred to as the support vectors, as shown in **Figure 3.7**. SVM adheres to the principle of structural risk minimisation, which posits that classification performance improves as the margin increases. Accordingly, an ideal SVM model identifies a maximum-margin hyperplane that both maximises the separation between classes and ensures correct classification of the training data. The optimal hyperplane is found by solving a quadratic optimisation problem [241].

The classic approach of SVM for classification problem, which used in this thesis, is based on a linear hyperplane function [243]:

$$y = \omega^T x + b \tag{3.19}$$

where x is the sets of input features, ω is the weight vector, and b is the estimated bias. The optimal hyperplane equation is achieved solving the quadratic programming model under soft margin optimisation [242]:

$$min\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \quad (3.20)$$

where ξ_i is the slack variable that allow margin violation and C is the user-defined regulation term that controls the trade-off between maximising the margin and minimising errors. The geometric margin is given by:

$$margin = \frac{2}{\|\omega\|} \quad (3.21)$$

The constraints for the optimisation are given by:

$$y_i(\omega^T x_i + b) - 1 + \xi_i \geq 0 \quad (3.22)$$

where $\xi_i \geq 0$ [242]. The loss function employed is hinge loss, which is regulated by $\frac{1}{2}\|\omega\|^2$. In the scenario where one class is significantly smaller size wise than the other, class weighting can be introduced to the soft margin optimization, as shown in Equation 3.23:

$$min\frac{1}{2}\|\omega\|^2 + \sum_{i=1}^{l} C_i\xi_i \quad (3.23)$$

where C_i is the class weighting regulation term, with a higher value for the minority class to increase its influence during training. The constraint for the optimization remains the same.

Over the years, various extensions of SVM have been developed, including models that employ non-linear hyperplane functions using kernel methods. However, for the purposes of Experiment 2 in this thesis, a linear SVM is considered more suitable. The linear SVM offers several key advantages: it provides greater interpretability, is computationally efficient (particularly important for high-dimensional data), and is generally faster to train. Additionally, linear SVMs are less prone to overfitting, especially with imbalanced datasets, due to the presence of a regularisation term that constrains model complexity and improves generalisation [244]. These characteristics make linear SVM a practical and effective choice for the predictive task in this thesis.

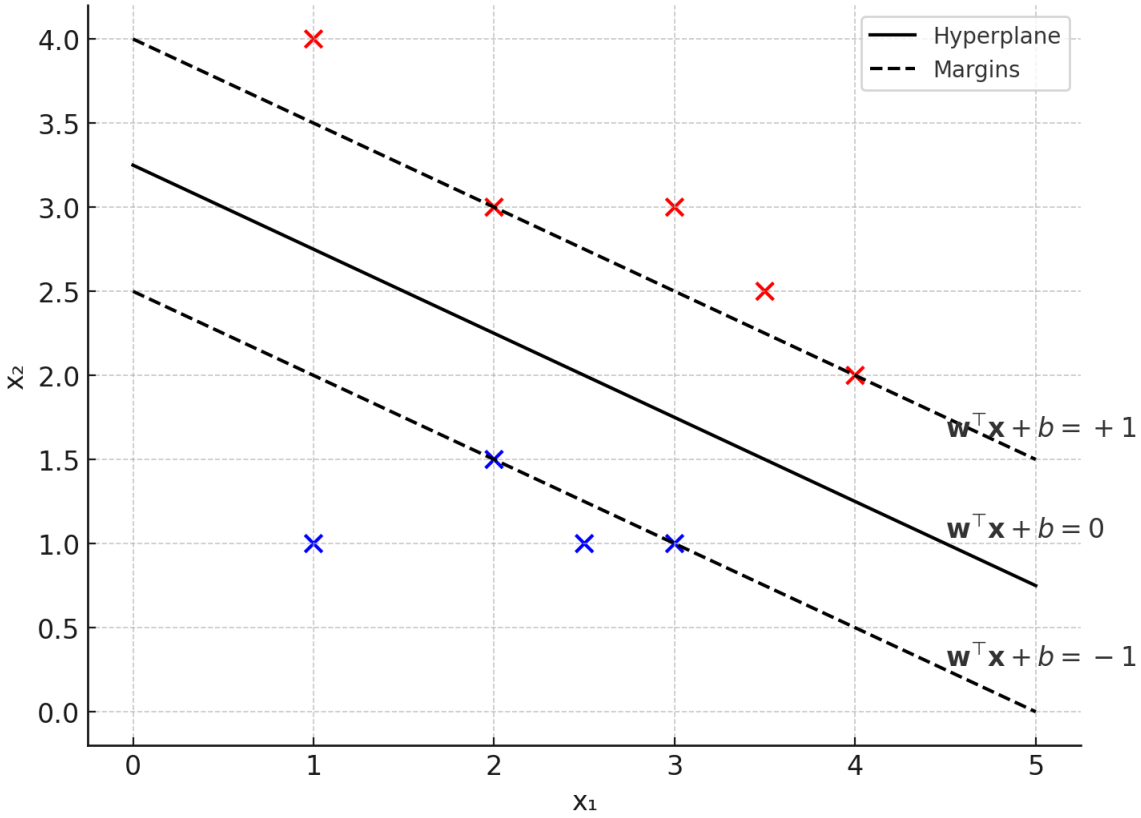


Figure 3.7 An example of linear SVM classification. Two classes, represented in red and blue, are separated by a hyperplane with margins clearly shown on either side.

3.4.4 Ensemble learning

Beyond single-model classification approaches, this thesis also employs ensemble learning methods. The primary motivation for adopting ensemble learning stems from the understanding that individual models often have limitations and may introduce errors that hinder predictive accuracy, particularly in real-world medical applications. Ensemble learning addresses these limitations by combining the predictions of multiple classifiers through a predefined aggregation strategy, forming a "committee" of decision-makers. By leveraging the complementary strengths of individual models, typically through weighted or unweighted voting, ensemble methods often achieve greater performance and robustness than single models alone [245]. For ensemble learning to be effective, two fundamental conditions must be met: the base classifiers must be both accurate and diverse [243]. Accuracy implies that each classifier performs better than random guessing, while diversity requires that the classifiers be statistically uncorrelated and make independent predictions [246].

Ensemble models have been widely applied in the medical field, including tasks such as disease diagnosis, disease progression prediction, and cell type classification [247-249]. Given that ensemble methods frequently outperform individual models that makes them up, this thesis adopts two widely used decision tree-based ensemble algorithms: RF and XGBoost, in place of standalone decision trees. Both RF and XGBoost use decision trees as base learners but differ in their training philosophies, as shown in **Figure 3.8**. RF is a bagging algorithm that builds multiple fully grown trees from randomly divided training subsets and aggregates predictions via majority voting for classification tasks [250]. In contrast, XGBoost is a boosting algorithm that constructs an ensemble of shallow trees sequentially, with each tree aiming to correct the errors of its predecessors [251]. This fundamental difference, independent learners in bagging versus dependent learners in boosting, also influences how each method handles imbalanced data. RF relies heavily on class weights to guide decision-making, whereas XGBoost often performs well without the need for explicit weighting.

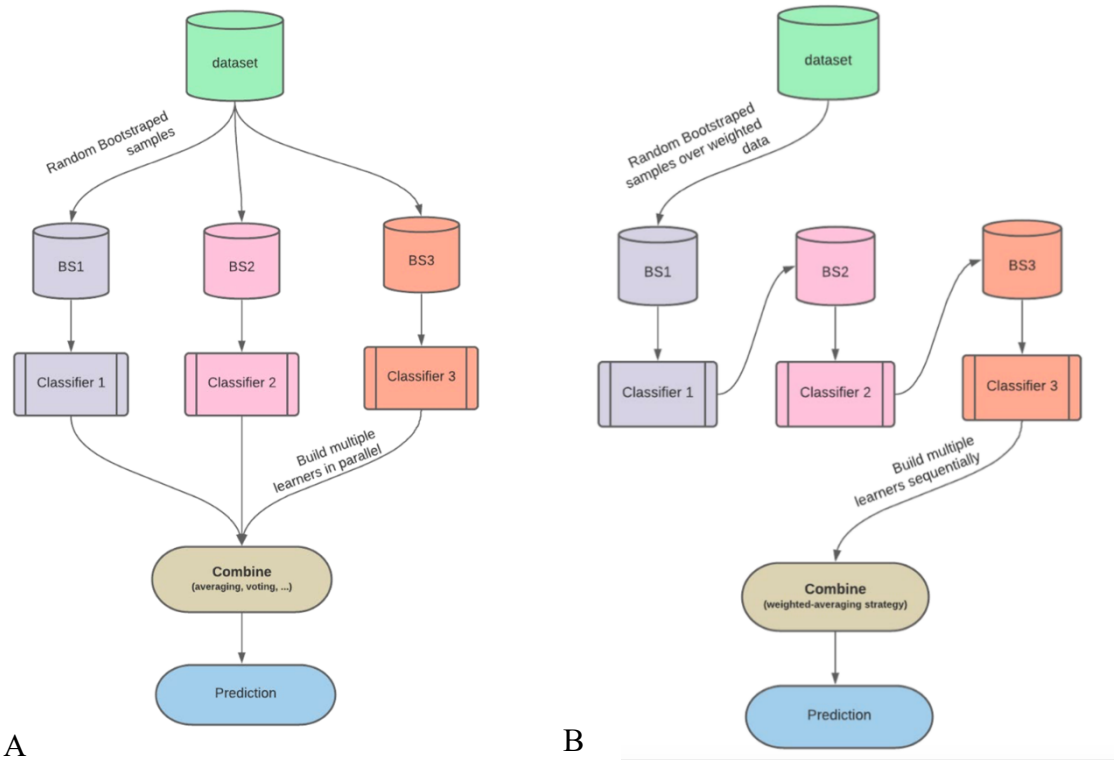


Figure 3.8 Overview of bagging (A) and boosting (B) approaches in ensemble learning. The bagging algorithm constructs multiple independent trees in parallel and combines their outputs through aggregation or majority voting. The boosting algorithm builds trees sequentially, with each new tree focusing on correcting the errors of the previous ones.

3.4.4.1 Random Forest (RF)

RF is an ensemble learning method composed of multiple tree-structured classifiers, where each decision tree is trained using bootstrap aggregating and random feature selection. Each tree acts as base classifier and independently casts a vote for the predicted class. The final decision is made by majority voting across all trees, as shown in **Figure 3.8**A [252]. As an ensemble of decision trees, RF inherits several advantages of individual decision trees, including invariance to feature transformation, robustness to irrelevant features, the ability to capture complex feature interactions, and resilience to noisy data. Moreover, RF addresses some key limitations of decision trees, it reduces overfitting and provides reliable estimates of feature importance [253].

This algorithm relies on two key sources of randomness. Bootstrap aggregating (bagging) trains each tree on a randomly drawn subsets with replacement (the bootstrap sample) from the training dataset. On average, each bootstrap sample includes approximately 64% of instance, while the remaining 36% out-of-bag instance can be used for internal validation. Bagging helps to reduce the variance of the model by averaging predictions over multiple slightly different training sets [249]. Random feature selection refers to the process in which, at each split within a tree, a random subset of features is selected, and the best split is determined within that subset. This technique introduces additional diversity among the trees, reduces inter-tree correlation, and significantly enhances generalisation performance compared to a single decision tree [254-256].

Based on the principle of bagging and random feature selection, the general process of the RF algorithm can be summarised as Algorithm 1, where *N* is the number of trees in the forest, *S* is the number of features randomly selected at each split, and *F* is the full set of input features.

Algorithm 1: Summarily of the RF algorithm [252].

Create an empty vector $\overrightarrow{RF}$
for $i = 1 \rightarrow N$ do
Create an empty tree T_i
Repeat
    Sample *S* features from *F* using bootstrap sampling
    Create a vector of the *S* features $\overrightarrow{F_S}$
    Find the best split feature $B(\overrightarrow{F_S})$
    Create a new node using $B(\overrightarrow{F_S})$ in T_i
Until no more instances to split
Add T_i to the $\overrightarrow{RF}$
end for
Output: A vector of trees $\overrightarrow{RF}$

In addition to bagging and random feature selection, the splitting criterion used at internal nodes plays a critical role in the performance of RF. During tree construction, the algorithm evaluates all possible feature–threshold combinations at each node to identify the split that most effectively separates the classes. The splitting criterion is designed to minimise node impurity, thereby selecting the most informative feature for partitioning the data. This process determines the optimal split feature $B(\overrightarrow{F_S})$ in Algorithm 1, ensuring that each decision tree contributes to improved class discrimination within the ensemble. For classification problem, as in the case in Experiment 2, RF uses Gini impurity as the splitting criterion during tree construction [257]. The standard Gini impurity is defined as:

$$Gini = 1 - \sum_{i=1}^{C} {p_i}^2 \tag{3.24}$$

where C is the class, p_i represents the probability of a sample belonging to class i, which is calculated as:

$$p_i = \frac{n_i}{n} \tag{3.25}$$

here n denotes the number of samples in the node, and n_i refers to the number of samples belonging to class i. Gini impurity is widely used in classification problems, which is easier in

computational complexity and is effective in node splitting [257]. To address class imbalance, a weighted version of Gini impurity can be applied by incorporating class weights w, which penalise the misclassification of minority classes. The weighted class probability is defined as:

$$p_i = \frac{w_i * n_i}{\sum_{j=1}^{C} w_j * n_j} \tag{3.26}$$

where w_i is the class weight of class i, j is the loop variable to sum across all class, and $\sum_{j=1}^{C} w_j * n_j$ counts for the total weighted sample in the node.

To enhance the performance of RF, this thesis employs Grid Search to optimise key hyperparameters, controlling the structure of individual trees and the level of randomness in the model. Grid Search systematically evaluates all possible combinations of predefined values within the hyperparameter space to identify the optimal configuration [258]. **Table 3.4** Overview of commonly adjusted RF hyperparameters and their typical values. Table adapted from [258]. summarises the commonly adjusted RF hyperparameters, along with their typical values. One of the central hyperparameters is *mtry*, which defines the number of features randomly selected at each split during tree construction. A smaller *mtry* encourages the development of more diverse and uncorrelated trees, leading to more stable aggregate predictions. However, this may come at the cost of reduced individual model accuracy, requiring a trade-off between stability and predictive power [258]. A commonly used rule of thumb for classification problems is $\sqrt{F}$, where F is the total number of features [259]. The sample size determines the number of observations drawn (with or without replacement) for training each tree, which is often experiment dependent. Like *mtry*, it involves a trade-off between stability and accuracy and can be tuned based on out-of-bag prediction performance. When the optimal sample size is chosen, the replacement strategy (sampling with or without replacement) typically has limited influence on overall performance [260]. However, some studies have raised concerns that sampling with replacement may introduce a slight variable selection bias [261, 262]. The node size controls the minimum number of observations required in a leaf node. Smaller values result in deeper trees and finer splits. For classification tasks, a node size of 1 is standard, though performance can be further improved by fine-tuning this parameter [258, 263]. Finally, the number of trees in the forest is also an important consideration and data dependent. While more trees generally lead to improved performance, the greatest gains are often achieved in the first 100 trees, especially in larger datasets [264, 265].

Table 3.4 Overview of commonly adjusted RF hyperparameters and their typical values. Table adapted from [258].

Hyperparameter	Description	Typical values
mtry	Number of features randomly selected in each split	$\sqrt{F}$ (classification)
Sample size	Number of observations that are drawn for a tree	Number of observations
Replacement	Draw observation with or without replacement	With/ Without replacement
Node size	Minimum number of observations in a leaf node	[1, 5, 10, 20]
Number of trees	Number of trees in the forest	[50, 100, 200, 500]
Splitting criterion	How split is chosen	Gini impurity (classification & used), MSE (regression)
Class weight	Particularly used for imbalanced data	Custom value via class weighting equation

3.4.4.2 Extreme Gradient Boosting (XGBoost)

XGBoost is a scalable, loss-driven, decision tree–based boosting ensemble model that is widely used for tabular data [266]. Boosting refers to a strategy in which the model begins with a weak learner (shallow tree) and sequentially builds new models that learn from the errors of previous iterations, ultimately improving overall performance, as shown in **Figure 3.8**B. Unlike bagging, which is used by RF and builds trees independently, boosting constructs trees sequentially, with each tree trained to reduce the residual errors of the previous ones [267].

XGBoost builds the model as combination of T additive functions to predict the final output, as shown in Equation 3.27 [266]:

$$f(x) = \sum_{t=1}^{T} f_t(x_i) \tag{3.27}$$

where $f(x)$ is the final prediction for given input feature x, and $f_t(x_i)$ indicates each decision tree trained at iteration t. The input feature subset x_i for each iteration can be optimised by hyperparameter tuning, allowing each tree to use either a randomly selected subset of features

or the full feature set. Each decision tree is associated with a mapping that assigns an input instance x_i to a corresponding leaf node, defined as Equation 3.28:

$$f_t(x_i) = \omega_{q(x_i)} \tag{3.28}$$

where ω is a set of leaf weights, representing the predicted score for all instances falling into that leaf. The structure function $q(x_i)$ defines trees' architecture by feature-based split conditions, determining the traversal path from the root to a specific leaf node [266]. Details of how these splits are selected will be discussed in subsequent sections.

The fundamental goal of XGBoost is to minimise the regularised objective function, as shown in Equation 3.29, for better predictive performance [266]:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{3.29}$$

where $\hat{y_i}^{(t-1)}$ is the current prediction of instance i at previous iteration $t-1$, $f_t(x_i)$ indicates the current decision tree fitted to correct the current residual error, $\Omega(f_t)$ is used as a regularisation term penalising model complexity, and $l(y_i, \hat{y_i})$ denotes the logistic loss used for binary classification problem, which is suited for Experiment 2 in this thesis. The loss function is approximated by the second-order Taylor expansion, as illustrated in Equation 3.30 to 3.32:

$$l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) \approx l\left(y_i, \hat{y_i}^{(t-1)}\right) + g_i f_i(x_i) + \frac{1}{2} h_i f_i(x_i)^2 \tag{3.30}$$

$$g_i = \frac{\partial l}{\partial \hat{y_i}} = \hat{p_i} - y_i \tag{3.31}$$

$$h_i = \frac{\partial^2 l}{\partial \hat{y_i}^2} = \hat{p_i}(1 - \hat{p_i}) \tag{3.32}$$

where g_i is the first derivative, h_i is the second derivative, and $\hat{p_i}$ represents the sigmoid predicted probability, which $\hat{p_i} = \sigma\hat{y_i}$. Thus, the objective Equation 3.29 can be approximated as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_i(x_i) + \frac{1}{2} h_i {f_i}^2(x_i)\right] + \Omega(f_t) \tag{3.33}$$

with regularisation term defined as:

$$\Omega(f_t) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_j^T {\omega_j}^2 \tag{3.34}$$

If I_j is defined as a set of instances in leaf j. Combined Equation 3.33 and 3.34, the objective equation can be expanded as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_i(x_i) + \frac{1}{2} h_i {f_i}^2(x_i)\right] + \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_j^T {\omega_j}^2 \tag{3.35}$$

Then substituting Equation 3.28 into 3.35, the objective function becomes:

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)\omega_j{}^2] + \gamma \mathrm{T} \tag{3.36}$$

For the feature-based splitting approach mentioned above, this thesis employs the approximate framework, as shown in Algorithm 2 [268-270].

Algorithm 2: approximate algorithm for split finding, adapted from Algorithm 2 [266].

for k = 1 to m do
Propose $S_k = \{s_{k1}, s_{k2}, s_{k3}, \dots, s_{kl}\}$ by percentiles on feature k
Propose can be done per tree or per split
end
for k = 1 to m do
$G_{kv} \leftarrow= \sum_{j\in\{j\mid s_{k,v}\geq x_{jk}>s_{k,v-1}\}} g_j$
$H_{kv} \leftarrow= \sum_{j\in\{j\mid s_{k,v}\geq x_{jk}>s_{k,v-1}\}} h_j$
end
Split with maximised gain

In summary, XGBoost optimises the splitting threshold by discretising continuous features into percentile-based bins (e.g., 10th, 20th percentiles, etc.). The binning process can be applied once per tree across all nodes for faster computation, or independently at each split for greater precision. The aggregated first (G_{kv}) and second (H_{kv}) derivative for each candidate split s_{kv} are computed for all instances. Then, based on the computed G_{kv} and H_{kv}, the split gain is determined as Equation 3.37, assuming a binary partition of tree nodes (left and right):

$$\mathcal{L}_{split\ gain} = \frac{1}{2}\left(\frac{G^2{}_L}{H_L+\lambda} + \frac{G^2{}_R}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \tag{3.37}$$

where H_L or G_L is the aggregated statistic to the left of split, H_R or G_R is the aggregated statistic to the right of split, and λ or γ is the regularisation term preventing overfitting [266]. The maximised $\mathcal{L}_{split\ gain}$ helps determine the optimal feature splitting and the corresponding split thresholds. This approximation improves the scalability of model and is as robust as the mostly used exact greedy split finding approach.

Unlike the previously discussed models, XGBoost does not strictly require the explicit specification of class weights when handling imbalanced data. This is attributed to its boosting framework and gradient-based optimisation, which inherently focus on learning from

misclassified instances. By sequentially constructing trees that correct errors made by previous iterations, XGBoost gradually enhances its ability to identify minority class samples.

Similar to RF, the performance of XGBoost can be significantly improved through hyperparameter tuning. In Experiment 2 of this thesis, XGBoost was optimised using Bayesian optimisation, a sequential model-based optimisation strategy for hyperparameter tuning. Bayesian optimisation constructs a surrogate model to approximate the unknown objective function and uses an acquisition function (often based on expected improvement or information-theoretic principles) to balance exploration (sampling in regions of high uncertainty) and exploitation (sampling near the current optimum). At each iteration, the acquisition function is maximised, and the surrogate model is updated with the new result [271].

XGBoost and RF share several common hyperparameters, such as the number of trees and the number of features at each split. Besides, XGBoost also relies on some unique hyperparameters. These include the learning rate, subsampling ratios for features and training data, the minimum gain required to make a split, and L1/L2 regularisation terms on leaf weights. These hyperparameters offer fine-grained control over model complexity and allow for performance optimisation.

3.4.5 Imbalanced data

Class imbalance refers to a condition in classification problems where one class significantly outnumbers the others, typically described as the majority class versus the minority class. Class imbalance is a common challenge that hinders the accurate estimation of instances from the minority class. Most machine learning algorithms are designed to maximise overall accuracy and minimise errors, which generally performs well when the classes are evenly distributed. However, in imbalanced scenarios, these algorithms tend to favour the majority class, leading to poor sensitivity and inadequate classification of the minority class [272]. This issue is especially prevalent in the biomedical field, where the positive class (disease diagnosis or event occurrence) is often much less frequent than the negative class. Techniques for addressing class imbalance are generally divided into two categories: data level approaches and algorithm level approaches [273].

Data level approaches are external techniques that address class imbalance by resampling the dataset, aiming to achieve a more balanced class distribution. These methods either oversample

the minority class or undersample the majority class to reach an equitable representation. In oversampling, additional instances of the minority class are generated, while in undersampling, a portion of the majority class is removed to match the minority class size. A widely adopted oversampling technique is the Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic samples of the minority class rather than simply duplicating existing ones [274]. SMOTE has been successfully applied in various domains, including network intrusion detection, breast cancer diagnosis, and miRNA gene prediction [275]. However, resampling methods also raise concerns. Undersampling may lead to the loss of information by removing potentially informative majority class instances. Oversampling, such as SMOTE, may fail to accurately preserve the original data distribution and could introduce the risk of overfitting and misclassification [276].

Algorithm level approaches address class imbalance by incorporating class weights into the cost function during training. This strategy achieves class balance by down-weighting the majority class or up-weighting the minority class, penalising the model more heavily for misclassifying minority class instances. Compared to data-level approaches, introducing class weights into the algorithm has the advantage of using the entire dataset without discarding samples or without interference from synthetic noises, thereby avoiding the risks of information loss or overfitting due to synthetic noise. The class weight is based on the Equation 3.38:

$$\omega_C = \frac{N}{K * N_C} \tag{3.38}$$

where ω_C is the weight assigned to class C, N denotes the total number of instances, K represents the total number of classes, and N_C is the total number of instances within class C [277]. The equation can be simply as a basic inverse-frequency weighting scheme:

$$\omega_C = \frac{1}{N_C} \tag{3.39}$$

which assigns larger weight to minority classes. Class weights are incorporated into the cost function for most models used in this thesis, except for XGBoost. This is because XGBoost handles class imbalance more flexibly. Its boosting framework naturally emphasises misclassified instances and therefore does not require the explicit introduction of class weights for regularisation in the same way as other models.

3.4.6 Performance measurement

The performance of Experiment 2 in this thesis, which employs various machine learning models, is evaluated using the following strategies. Model performance in classifying CVD

mortality outcomes is assessed through metrics derived from the confusion matrix, including sensitivity, specificity, accuracy, and F1 score. To ensure reliable and stable performance estimates, 10-fold cross-validation is applied throughout the experiments. In addition, to evaluate the statistical significance of performance differences between models and feature combinations, Wilcoxon signed ranks test are conducted. While overall performance is quantified through these standard classification metrics, the individual contributions of features are interpreted using SHAP analysis, providing insights into feature importance and improving model interpretability. The details of each evaluation strategy are discussed in this section.

3.4.6.1 Confusion matrix

A confusion matrix is a widely used tool for visualising and evaluating the performance of classification models. It compares the predicted class labels with the actual labels, providing a summary of model performance in classification. An example confusion matrix for a binary classification problem is presented in **Table 3.5** [278]. In a binary classification task, such as disease diagnosis or the task in Experiment 2 of this thesis, the outcome is typically classified as either positive or negative. There are four possible outcomes when comparing the predicted and actual labels, as summarised in **Table 3.5**:

- True Positive (TP): the actual label is positive, and the model correctly predicts it as positive.
- True Negative (TN): the actual label is negative, and the model correctly predicts it as negative.
- False Positive (FP): the actual label is negative, but the model incorrectly predicts it as positive.
- False Negative (FN): the actual label is positive, but the model incorrectly predicts it as negative.

Based on the four possible outcomes in the confusion matrix, the performance of classification models can be evaluated using several key metrics: Sensitivity, Specificity, Accuracy, and F1 Score. Sensitivity, also known as Recall, measures the proportion of individuals who have the condition and are correctly identified by the model as positive. It is calculated as:

$$Sensitivity = \frac{TP}{TP+FN} = \frac{TP}{N_{P*}} \tag{3.40}$$

Specificity, or Selectivity, quantifies the proportion of individuals who do not have the condition and are correctly classified as negative:

$$Specificity = \frac{TN}{TN+FP} = \frac{TN}{N_{N*}} \quad (3.41)$$

Accuracy reflects the overall correctness of the model by measuring the proportion of total predictions that are correct:

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} = \frac{TN+TP}{N_{**}} \quad (3.42)$$

The F1 Score is a widely used metric for imbalanced datasets, especially when the focus is on correctly identifying positive cases, as is the case in Experiment 2 of this thesis. It is defined as the harmonic mean of Positive Predictive Value (PPV) and Sensitivity:

$$F1\ Score = 2 \times \frac{PPV \times Sen}{PPV+Sen} = \frac{2TP}{2TP+FP+FN} \quad (3.43)$$

where PPV measures the proportion of people who predicted positive for a disease and have the disease:

$$Positive\ Predictive\ Value = \frac{TP}{TP+FP} = \frac{TP}{N_{*P}} \quad (3.44)$$

Table 3.5 An example of a binary classification confusion matrix.

		Predicted class label		
		Positive	Negative	Sum
Actual label	Positive	True Positive (TP)	False Negative (FN)	N_{P*}
	Negative	False Positive (FP)	True Negative (TN)	N_{N*}
	Sum	N_{*P}	N_{*N}	N_{**}

3.4.6.2 Receiver operating characteristic (ROC) curve and Area under the ROC curve (AUC)

The ROC curve is a graphical tool for evaluating the performance of a binary classification method, and is useful for assessing a model's ability to discriminate between classes. By plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at different thresholds, the ROC curve reveals the trade-off between sensitivity and specificity across the full range of operating points. The AUC, representing the area under the ROC curve, quantifies the overall performance of the classifier across all decision thresholds. The AUC value ranges from 0 to 1, with a value of 0.5 or below indicating no predictive ability (random guessing), and a value close to 1 indicating strong predictive ability [279].

As shown in **Figure 3.9**, the averaged ROC curve obtained via 10-fold cross validation illustrates the performance of an XGBoost model in predicting CVD mortality. The mean ROC curve (blue) represents the model's performance across all possible decision thresholds, with a curve closer to the top-left corner indicating better performance due to higher sensitivity and lower false positive rate. The curve also facilitates the selection of an appropriate decision threshold to meet specific clinical needs. For instance, if high sensitivity is prioritised, the selected threshold should correspond to a point on the curve where sensitivity is maximised, even if it comes at the cost of a higher false positive rate.

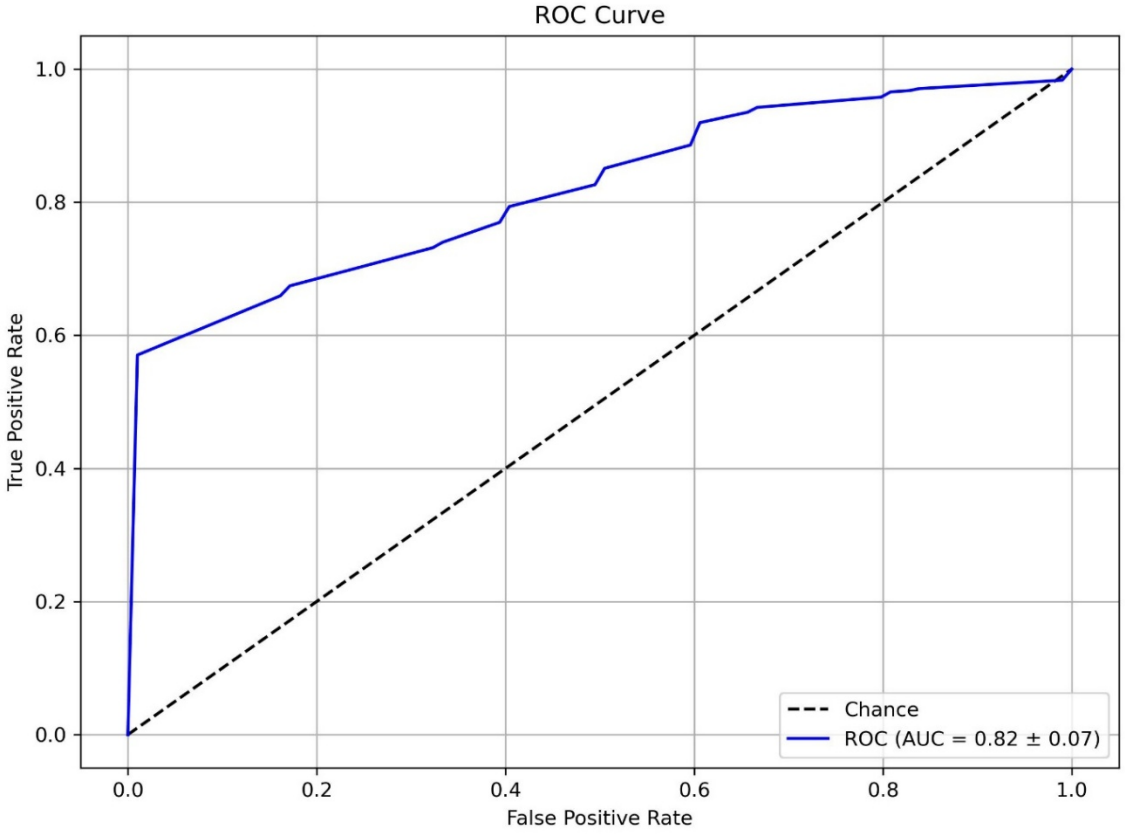


Figure 3.9 An example of the ROC curve and AUC for a binary classifier. The false positive rate (1 − specificity) represents the proportion of actual negative cases incorrectly classified as positive, while the true positive rate indicates the proportion of actual positive cases correctly classified as positive. The diagonal grey line (Chance) represents the performance of random guessing, whereas the blue ROC curve illustrates the classifier's performance across all possible decision thresholds.

3.4.6.3 10-fold cross validation

When evaluating the performance of a predictive model, the strategy of using data plays a critical role in ensuring accurate and unbiased assessment. For example, if the same dataset is

used for both training and testing, the model is likely to memorise the training data, leading to an overestimated performance. In such cases, the model may appear highly accurate but is unlikely to generalise well to unseen data, resulting in overfitting. The ideal approach is to train the model on one subset of data and evaluate its performance on a separate, unseen subset, ensuring a more realistic estimate of how the model would perform in real-world applications [280]. To achieve this, 10-fold cross validation is employed in Experiment 2 of this thesis.

In 10-fold cross validation, the input dataset is randomly divided into 10 equally sized subsets, referred to as 10 folds. For each of the 10 iterations, one fold is used as the testing set, while the remaining nine folds are used to train the model. This process is repeated 10 times, with each fold serving as the testing set exactly once [280]. At the end of this procedure, every data point is used for training and testing. For each fold, the model is trained on the corresponding training set and evaluated on the held-out testing set, with performance metrics (e.g. sensitivity, specificity, accuracy, and F1 score) calculated for each round. The overall performance of the model is then obtained by averaging the metrics across all folds, as shown in **Figure 3.10**.

The key advantage of 10-fold cross validation, compared to other data partitioning strategies such as the 80/20 hold-out method, is that it ensures all data are used for both training and testing. This means that no information is excluded from the training process, reducing the risk of underfitting and preventing performance underestimation due to an undertrained model. At the same time, the method avoids testing on seen data, thereby offering a reliable and unbiased estimate of predictive performance and minimising the risk of overfitting.

In scenarios involving imbalanced datasets, 10-fold cross validation can be adapted to preserve class distribution across folds through a method known as stratified 10-fold cross validation. In stratified cross validation, each fold contains approximately the same proportion of samples from each class as in the original input dataset. This approach ensures that in every iteration, the model is trained on representative samples from all class labels and evaluated on a test set that also reflects the class distribution. As a result, stratified 10-fold cross validation retains the core advantages of standard cross validation while suitable for imbalanced classification problems.

Figure 3.10 Demonstration of 10-fold cross validation. The dataset is divided into 10 equal-sized subsets (folds). In each iteration, one fold is used as the testing set, while the remaining nine folds are used for training. Performance metrics are computed for each iteration, and the overall model performance is obtained by averaging the results across all folds.

3.4.6.4 Wilcoxon signed ranks test

Since this thesis involves comparing multiple machine learning models to identify the best-performing model for predicting CVD mortality outcomes, the statistical significance of performance differences is also assessed. Specifically, when an improvement in performance is observed between two models, the Wilcoxon signed ranks test is conducted to rank the performance difference between two models, as shown in Equation 3.45 [281].

$$T = \min(R^+, R^-) \tag{3.45}$$

where T is the Wilcoxon-statistic, R^+ denotes the sum of ranks where Model 2 outperforms Model 1, and R^- represents the sum of ranks for the opposite case. R^+ and R^- are calculated as:

$$R^+ = \sum_{d_i>0} rank(d_i) \tag{3.46}$$

$$R^+ = \sum_{d_i<0} rank(d_i) \tag{3.47}$$

where d_i is the difference between each pair i , and the standard Wilcoxon signed ranks test considers $d_i = 0$ excluded [282]. The advantage of using the Wilcoxon signed ranks test over the traditional paired t-test is that it does not assume normality of the differences between observations, making it more robust to a variety of data distribution [281]. In this context, the p-value represents the probability of observing a test statistic as extreme or more extreme than the calculated Wilcoxon-statistic, assuming the null hypothesis (the median difference between

models is 0) is true. A p-value less than 0.05 indicates a statistically significant difference between model performances.

3.4.6.5 SHapley Additive exPlanations (SHAP)

Beyond seeking for an effective machine learning model for classifying CVD mortality outcomes, this thesis also emphasises the importance of model explainability. Transparency in model decision-making is essential for fostering trust and understanding among end users, particularly clinicians and patients. Accordingly, a range of machine learning models were selected based on both predictive performance and interpretability. While simpler models such as LDA offer high transparency, they often exhibit suboptimal performance. In contrast, more complex models, such as those based on ensemble learning, tend to achieve superior predictive accuracy but are inherently more difficult to interpret. To address this challenge, SHAP analysis is employed in this thesis to interpret the individual contribution of each feature to the model's prediction, enhancing the interpretability of advanced models without compromising on performance [283].

SHAP analysis is a commonly used method within the class of additive feature attribution techniques, grounded in Shapley values from cooperative game theory. Additive feature attribution methods explain a model's prediction by assigning interpretable contributions to each input feature, under the assumption that the model output can be expressed as the sum of individual feature effects. This is represented by Equation 3.48:

$$g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z'_i \tag{3.48}$$

where $g(z')$ is the explanation model, $z \in \{0,1\}^M$ M is the number of simplified input features, $\emptyset_i$ denotes the individual effect of features, and $\emptyset_0$ represents the expected model output over the background dataset. It serves as the baseline prediction prior to the contribution of individual features. The SHAP value is based on the additive feature attribution equation, seeking the solution of $\emptyset_i(f, x)$:

$$\emptyset_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} [f_x(z') - f_x(z' \backslash i)] \tag{3.49}$$

$$f_x(z') = E[f(z)|z_S] \tag{3.50}$$

where f is the original model, and x is the original input features, $E[f(z)]$ represents the baseline value, and S is the set of non-zero indexes in z'. Compared to other additive feature attribution methods, SHAP uniquely satisfies three desirable properties for local explanations:

1. Local accuracy: the sum of SHAP values equals the model output.
2. Missingness: if a feature is not used in a prediction, its SHAP value is zero.

3. Consistency: if a model change such that the contribution of a feature increases or stays the same, its SHAP value does not decrease.

In general, a larger absolute SHAP value (regardless of positive or negative) indicates a stronger contribution of a feature to the model's prediction. A positive SHAP value suggests that the feature pushes the prediction towards the positive class, whereas a negative SHAP value implies that the feature drives the prediction towards the negative class. An informative way to visualise SHAP values is through a beeswarm plot, as shown in **Figure 3.11** [284]. In this plot, features are listed on the y-axis (left), ranked from top to bottom based on their overall contribution to the predictive model (from the most to the least important). Each dot represents an individual data instance, and its position along the x-axis corresponds to the SHAP value for that instance and feature. The colour gradient, shown along the y-axis (right), reflects the original feature value, with red indicating high values and blue indicating low values.

For example, consider the row corresponding to the feature MedInc in **Figure 3.11**. This feature appears at the top of the plot, indicating its high overall importance in the predictive model. The red dots positioned on the right side of the SHAP axis show that higher values of MedInc are associated with positive SHAP values, thereby increasing the likelihood of a positive class prediction. In contrast, blue dots, representing lower MedInc values, are mostly located on the left side, indicating a contribution to predicting the negative class. Therefore, the ranking of features along the y-axis reflects their general importance, while the distribution and colour of dots for each feature provide insight into how individual instances influence the model's output.

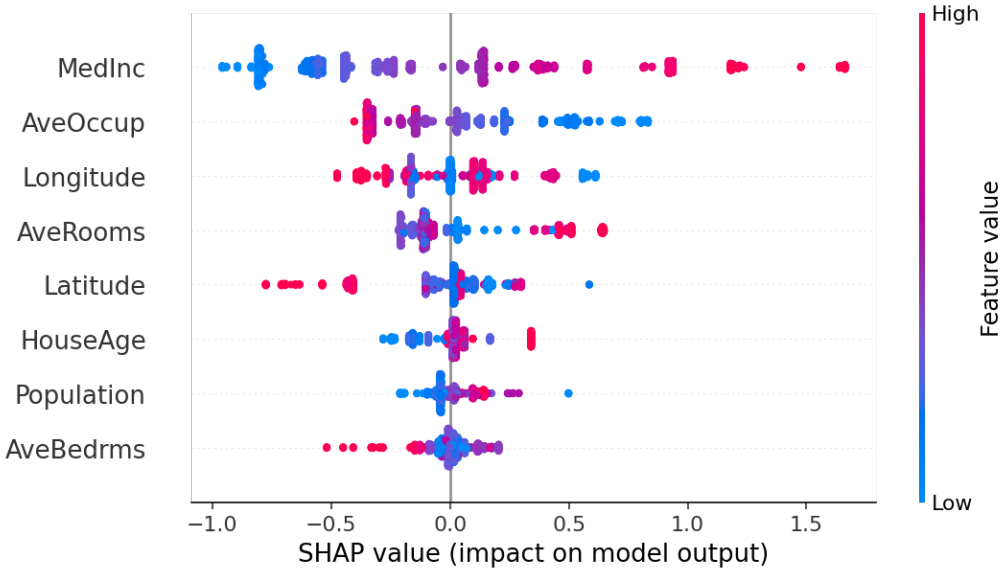


Figure 3.11 An example of SHAP analysis using beeswarm plots, Figure adapted from [284].

3.5 Summary

This chapter outlines the methodology and rationale underlying the two experiments conducted in this thesis. The chapter serves two primary purposes. First, Sections 3.1 and 3.2 provide a detailed review of PSG-derived parameters, as well as demographic and medical history information commonly used in OSA–CVD analysis. Section 3.1 focuses on optimal oximetry-derived parameters for CVD prediction, including their definitions, computational methods, and current limitations. Based on this analytical review, the motivation for Experiment 1 and part of the rationale for Experiment 2 are introduced in Section 3.1.5. Section 3.2 expands the discussion to additional parameters that may serve as covariates in Experiment 1 or as baseline features in Experiment 2.

The second part of the chapter, Sections 3.3 and 3.4 shift focus to the evaluation tools used across both experiments. Section 3.3 presents the Cox proportional hazards model, which is used to examine the relative hazard associated with oximetry-derived parameters in predicting CVD mortality. This section addresses the key research question: *Can oximetry-derived parameters effectively predict CVD outcomes?* The Cox model is introduced along with its theoretical foundations, ability to assess independent effects, and potential limitations when applied inappropriately. To overcome the constraints of the Cox model, particularly its focus on relative risk rather than individual-level prediction, Section 3.4 introduces machine learning techniques. These models are applied in Experiment 2, alongside the features described in Sections 3.1 and 3.2, to enable individualised prediction of CVD mortality and to explore the combined predictive performance of PSG-derived parameters.

Chapter 4
Experiment 1

4 Comparison of oxygen desaturation area-based methods in predicting cardiovascular disease mortality outcomes

This chapter particularly focuses on the oximetry signals and use of desaturation area-based parameters in predicting CVD mortality. Desaturation area-based parameters have emerged as novel predictor of CVD mortality [12, 28]. Existing algorithms estimate the area under the oximetry curve but differ in computational aspects due to variations in baseline, sampling window, and event choice. These differences result in varying computational complexity and predictive performance. This chapter systematically characterises the published desaturation area-based algorithms and evaluates the fifteen possible combinations of event choices, baseline, and sampling windows to identify the most effective method for predicting CVD mortality.

This experiment presents the first comprehensive comparison of these algorithms within the same patient population, using a standardised definition of events to identify the most effective method for predicting CVD mortality. Furthermore, the comparison between manually scored respiratory events and automatically detected desaturation events offers valuable insights for improving future automated algorithm development.

4.1 Rationale

AHI, the standard measure of OSA severity and a primary diagnostic tool, quantifies the frequency of apnoea and hypopnoea events during sleep [66, 285-288]. However, as discussed in previous chapters, many studies suggest that AHI may not adequately elucidate the association between OSA and CVD. It fails to capture factors such as respiratory event duration, sleep fragmentation, arousal events, and oxygen saturation, all of which exert critical impacts on the cardiovascular system [12, 27, 114]. With the growing understanding of the association between OSA and CVD, a range of oximetry-based parameters has been employed to analyse this relationship [289]. Among these, T90 and ODI are routinely calculated in sleep studies and have shown promise as predictors of CVD outcomes in patients with OSA [6, 14].

Over the last decade, a novel set of oximetry parameters has been introduced: the desaturation area-based parameters. Evidence so far suggests that these parameters may provide superior prediction of CVD outcomes in community populations and chronic heart failure-free populations [12, 15, 290], compared to the T90 and ODI parameters. Desaturation area-based parameters measure the cumulative area of the oximetry trace within a sampling window beneath a baseline associated with sleep events.

To date, desaturation area-based parameters have generally been classified according to the type of event used as the trigger: respiratory events, typically annotated manually by sleep experts following AASM guidelines, and blood oxygen desaturation events, which can be identified automatically by algorithms. To capture both approaches, this experiment selected three published desaturation area-based algorithms. HB and REDTA are based on manually scored respiratory events, whereas DesSev is calculated directly from automated algorithms [12, 28, 166]. DesSev was chosen over many other automated methods because it is open-sourced and supported by published software. This ensures reproducibility and alignment with the original design. By contrast, other automated algorithms, such as HBoxi, have not disclosed full implementation details. This limits reproducibility and raises the risk that differences in performance may reflect variations in reimplementation rather than differences in the algorithms themselves [291]. Although HBoxi has shown promise in predicting CVD mortality, its limited methodological transparency placed it beyond the scope of this thesis.

These three algorithms also differ in how the sampling window and baseline are defined. DesSev employs an event-specific sampling window, HB applies a record-specific window, and REDTA uses a fixed window. In terms of baselines, both DesSev and HB adopt event-specific baselines, whereas REDTA assumes a fixed baseline of 100% [11, 15, 28, 166]. The visual demonstrations of sampling windows and baselines are presented in **Figure 3.2** (HB), **Figure 3.3** (REDTA), and **Figure 3.4** (DesSev). Event-specific windows and baselines allow the area calculation to be tailored precisely to each event, but they are more susceptible to noise [28]. Moreover, when events occur in close succession, the residual effects of earlier events can interfere with baseline estimation for subsequent ones, reducing the reliability of event-specific approaches [292]. By contrast, fixed windows and baselines adopt a “one-size-fits-all” approach that is more robust to noise but is not customised for the recording/event under study. The record-specific approach represents a compromise between the two extremes but requires

additional human interpretation to determine how the windows should be adapted for each recording.

To our knowledge, no study has systematically examined how computational variations affect desaturation area-based parameters or whether such differences alter their ability to capture the association between OSA and CVD. Existing studies vary in their computational methods, definitions of sleep events (based on different AASM criteria), and study populations, making cross-study comparisons difficult [16, 17]. Moreover, the current parameters have not been jointly assessed in the same database for predicting the same outcomes, thus limiting direct performance comparisons.

This experiment addresses these gaps by conducting a comprehensive comparison of major desaturation area-based methods within the same patient population, using consistent definitions of respiratory and desaturation events. To ensure accurate implementation, the algorithms were applied using their original methods proposed by the authors (REDTA), validated replication methods (HB), and publicly available software (DesSev). The aim of this experiment is to examine the effects of event selection, sampling window, and baseline calculation on desaturation area-based measures for predicting CVD mortality and to identify the method that best suits this prediction.

4.2 Database

4.2.1 Study samples

This experiment used data from the SHHS, a community-based multi-centre cohort study conducted by the National Heart, Lung, and Blood Institute. The dataset is publicly available and was designed to investigate the consequences of sleep-disordered breathing, including cardiovascular outcomes [293, 294]. A total of 6441 men and women aged 40 years and older participated in an unattended PSG at baseline, with a follow-up examination conducted over the subsequent decade. The follow-up collected information on CVD outcomes, demographics, smoking history, and alcohol use through interviews, self-reported questionnaires, telephone contacts, and adjusted surveillance methods. Of the initial 6441 participants who met the inclusion criteria (no history of sleep apnoea treatment, current home oxygen therapy, or tracheostomy), 5804 completed the study.

During the PSG examination, SpO2 signals were recorded using a fingertip pulse oximeter (Nonin, Minneapolis, MN) at a sampling rate of 1 Hz. Respiratory events were manually scored by the SRC in Boston, USA, in accordance with the SHHS Reading Centre Manual of Operations. Desaturation events were identified based on amplitude attenuation, apnoea and hypopnoea events were detected using airflow measurements, and arousals associated with respiratory events were identified through EEG signals [294, 295]. The event scoring criteria are detailed in Chapter 3.

4.2.2 Sample selection and characteristics

The experiment selected participants with completed PSG, CVD mortality outcomes, and covariates. According to the experimental design, the covariates included demographic variables, smoking status, alcohol intake, non-CVD-related medical history, AHI, T90, MinSat, and concurrent cardio-metabolic diseases (heart failure, stroke, angina, coronary revascularisation, and myocardial infarction). Of the 5,804 participants who completed the study, 11 were excluded due to missing PSG data, 711 due to missing covariate data, 393 due to missing CVD mortality outcome data, and 206 due to the missing desaturation areas that the DesSev software failed to compute. This resulted in a final sample of 4,483 participants eligible for analysis, as depicted in **Figure 4.1**. Within this cohort, 311 deaths were attributable to CVD [58]. The sample characteristics are presented in **Table 4.1**. In the CVD survivor group, females comprised 53.91% of participants, whereas in the CVD mortality group, males were more prevalent, accounting for 53.38%. Participants were predominantly Caucasian in both the survivor group (88.49%) and the mortality group (88.75%). The mean age was 63.47 years in the survivor group and 75.79 years in the mortality group. The mean AHI indicated moderate OSA, with values of 17.73 in the survivor group and 20.94 in the mortality group.

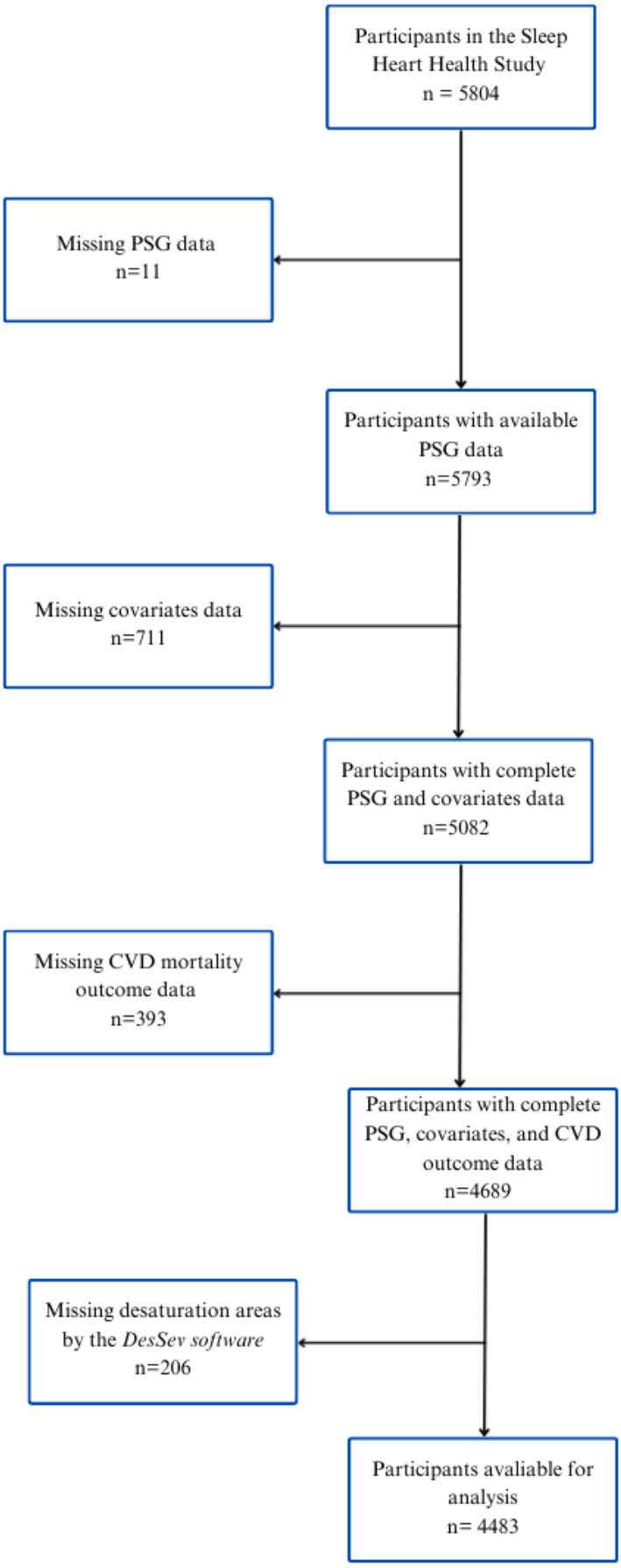


Figure 4.1 Flow chart for the study sample identified for inclusion from SHHS cohort database.

Table 4.1 Sample characteristics of the SHHS involved in the analysis

Variables	Total n= 4483 (100%)	
	CVD survivor n = 4172 (93.1%)	CVD death n = 311 (6.9%)
Age (years), mean (SD)	63.47 (10.64)	75.79 (7.60)
BMI (kg/m^2), mean (SD)	28.34 (5.09)	27.33 (4.79)
Race		
Caucasian, n (%)	3692 (88.49)	276 (88.75)
Other, n (%)	480 (11.51)	35 (11.25)
Gender		
Male, n (%)	1923 (46.09)	166 (53.38)
Female, n (%)	2249 (53.91)	145 (46.62)
Smoking status		
Never, n (%)	1941 (46.53)	142 (45.66)
Former, n (%)	1830 (43.86)	148 (47.59)
Current, n (%)	401 (9.61)	21 (6.75)
Total time of sleep (TST), n (%)		
≤ 5h	626 (15.00)	79 (25.40)
5-8h	3499 (83.87)	230 (73.95)
≥ 8h	47 (1.13)	2 (0.65)
T90 (%TST), mean (SD)	3.36 (9.96)	6.01 (15.53)
AHI (events/h), mean (SD)	17.73 (15.73)	20.94 (16.00)
COPD, n (%)	50 (1.20)	3 (0.96)
Stroke, n (%)	125 (3.00)	32 (10.29)
Heart failure, n (%)	57 (1.37)	21 (6.75)
Diabetes, n (%)	254 (6.09)	65 (20.90)
Hypertension, n (%)	1576 (37.78)	218 (70.10)
Lipid-lowering medication use, n (%)	507 (12.15)	50 (16.08)

4.3 Methodology

4.3.1 Desaturation area-based methods

To achieve the goal of assessing the impact of different computational approaches on desaturation area-based parameters in predicting CVD mortality, and of identifying the method best suited for this prediction, the experiment began with the implementation of three published algorithms: DesSev, HB, and REDTA (the implementation details are provided in Chapter 3). These three oximetry-derived algorithms all characterise the overnight desaturation area: the area between the oximetry trace and the baseline within sampling windows, which are associated with sleep events. However, they differ in computational methodology and can be categorised along three dimensions: the choice of events, the definition of sampling windows, and baselines. As shown in **Table 4.2**, DesSev employs event-specific sampling windows and event-specific baselines, with windows triggered by automatically detected desaturation events [29, 176]. In contrast, HB and REDTA both rely on manually scored respiratory event annotations. HB is calculated using record-specific sampling windows and event-specific baselines, whereas REDTA is designed as a "one-size-fits-all" method, applying fixed sampling windows and a fixed baseline [12, 28]. The detailed computational steps for each algorithm, including event scoring criteria, were detailed in previous chapters.

The implementation provided the basis for the subsequent experiment. To maintain consistency with the published algorithms, the same standards were applied for event annotation, sampling windows, and baselines.

For the events:

- **Automatically detected desaturation events** with a 3% SpO2 drop were annotated using the open-source ABOSA package, developed by the original team for calculating DesSev [176].
- **Respiratory events** were manually scored by the SRC and provided by the SHHS database.

For the sampling windows:

- The **event-specific window** was defined as the start and end of each respiratory or desaturation event [176].
- The **record-specific window** was defined by the two peaks of the averaged SpO2 trace [12].

- The **fixed window** started at the midpoint of the event and extended to 2.5 times the event duration [28].

For the baselines:

- The **event-specific baseline** was set as the maximum SpO2 value within 100 seconds before the end of the event [12], except for DesSev, whose baseline was calculated directly by the ABOSA software.
- The **record-specific baseline** was defined as the 99th percentile of the SpO2 signal within a single recording [292].
- The **fixed baseline** was set at 100% [28].

Table 4.2 Summary of the desaturation area calculation methods implemented*. Due to limitations of software used for processing signals we were unable to implement the 3 methods shaded grey. All other methods were implemented.

	Manually scored respiratory events			**Automatically detected desaturation events**		
Sampling window	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**
Event-specific sampling window	A_{EEM}	A_{ERM}	A_{EFM}	A_{EEA} (DesSev)	A_{ERA}	A_{EFA}
Record-specific sampling window	A_{REM} (HB)	A_{RRM}	A_{RFM}	A_{REA}	A_{RRA}	A_{RFA}
Fixed sampling window	A_{FEM}	A_{FRM}	A_{FFM} (REDTA)	A_{FEA}	A_{FRA}	A_{FFA}

* $A_{a,b,c}$: A denotes desaturation area methods, a represents the sampling window (E: event-specific; R: record-specific; F: fixed), b is the choice of baseline (E: event-specific; R: record-specific; F: fixed), and c indicates the type of events (M: manually scored respiratory events; A: automatically detected desaturation events).

With standards set, this experiment generated 15 possible combinations of desaturation area-based methods, varying among event choices, sampling window definition, and baseline calculations, as summarised in **Table 4.2**. This experiment introduced a unified notation for all these methods. A method is denoted with the $A_{a,b,c}$, where a represents the sampling window (E: event-specific; R: record-specific; F: fixed), b is the choice of baseline (E: event-specific; R: record-specific; F: fixed), and c indicates the type of events (M: manually scored respiratory events; A: automatically detected desaturation events). Using this notation, three base algorithms are denoted as follows: DesSev (A_{EEA}), HB (A_{REM}), and REDTA (A_{FFM}).

While most methods were computed in MATLAB following the predefined standards above, the computation of A_{EEA} (DesSev), A_{ERA}, and A_{EFA} were an exception, as these were implemented using the ABOSA software. The rationale for using ABOSA for these methods, rather than manual implementation as with the others, was to ensure that observed performance differences were attributable solely to the algorithms, not to potential inconsistencies introduced during algorithm reimplementation.

A_{EEA} (DesSev):

- Implemented directly via ABOSA software.
- The event-specific baseline was defined as the maximum SpO2 value during an event (usually at the start of the event), representing a minor deviation from the definition above.

A_{EFA}:

- Implemented directly via ABOSA software, without deviation from the standard methodology.

A_{ERA}:

- Calculated using ABOSA software.
- The value of the record-specific baseline was first determined. A constant value equal to the difference between this value and 100% was then added to every SpO2 sample, so that the record-specific baseline was adjusted to exactly 100%. The adjusted SpO2 signal was then processed by ABOSA, and the resulting fixed baseline output was taken as A_{ERA}.

Additionally, as the ABOSA software is currently unable to accept a respiratory event list, this experiment could not calculate three methods associated with the event-specific sampling

window and manually scored respiratory events, as indicated in grey in **Table 4.2**. The software functions as a closed and highly integrated package, with no facility to modify event annotations or import externally respiratory event lists. As a result, the event-specific sampling windows and baselines can only be computed within the software's native workflow, as originally implemented by the authors. However, because ABOSA cannot accept manually scored respiratory events as input, it is not possible to combine manually score respiratory events timing with the software's automated baseline and sampling window calculations. Thus, the three desaturation area–based methods shaded in grey could not be evaluated in this experiment.

Consequently, a total of 15 combinations were evaluated. Although two event variations combined with three sampling window variations and three baseline variations could theoretically yield 18 combinations, three of these could not be implemented due to software limitations.

4.3.2 Statistical analysis

The results of each desaturation area method were normalised (z-scores) and treated as distinct SpO2 predictors of CVD mortality [148]. Normalisation was necessary to allow direct comparison across parameters with different units. It was performed by subtracting the mean value of each set of desaturation areas and dividing by its corresponding standard deviation. HR, p-values, and associated 95% confidence intervals (95% CI), derived from the Cox proportional hazards regression analysis [215], were used to compare the ability of each desaturation area method to predict CVD mortality. A p-value threshold of 0.05 was applied to determine statistical significance, and a threshold of 0.10 was applied to indicate statistical trend. A higher hazard ratio with statistical significance indicated a stronger predictive ability of the desaturation area method for CVD mortality.

The Cox proportional hazards regression models in this study were adjusted using the same covariates as those in Model 4 of the study by Azarbarzin et al. [12]. These covariates included age, race, gender, total sleep time, smoking status, alcohol use, existing COPD, AHI, T90, event-related MinSat, and concurrent cardio-metabolic diseases (heart failure, stroke, angina, coronary revascularisation, and myocardial infarction). In total, three Cox regression models were fitted. The unadjusted model estimated HRs for CVD mortality using desaturation areas alone. The partially adjusted model incorporated demographic factors, smoking status, alcohol

intake, and non-CVD–related medical history as covariates. The fully adjusted model included the same covariates as the partially adjusted model, with the additional adjustment for concurrent cardiometabolic disease.

4.4 Results

To evaluate the predictive efficacy of fifteen desaturation area-based methods, we conducted a comparison of their performance in predicting CVD mortality using the same covariates. **Table 4.3** shows the unadjusted HRs with corresponding 95% CI, while **Table 4.4** and **Table 4.5** present the adjusted HRs with corresponding 95% CI for normalized desaturation areas as predictors of CVD mortality.

The results in **Table 4.4** and **Table 4.5** indicate that the oximetry desaturation areas based on automatically detected desaturation events were unsuccessful in predicting CVD mortality as all p-values of the HRs were greater then 0.1, for partially and fully adjusted covariate models. In contrast, the p-values of the HRs for oximetry desaturation areas based on manually scored respiratory events were less than 0.1 for the partially adjusted model (**Table 4.4**) suggesting a trend of this group of oximetry parameters providing independent predicting performance of CVD mortality. The A_{RRM}, A_{RFM}, and A_{FRM} models all demonstrated statistical significance. These results did not hold up for the fully adjusted model (**Table 4.5**). The only parameter that achieved statistical significance was A_{RRM} (HR of 1.79 (95% CI: 1.00–3.19)). A_{RFM}, and A_{FRM} demonstrated a statistical trend with a p-value less than 0.1 and all other parameters resulted in a p-value greater 0.1.

Table 4.3 Desaturation area-based methods predicting CVD mortality in the SHHS with unadjusted model. The hazard ratios and corresponding 95% confidence intervals are shown for evaluating the performance. All methods are statistically significant.

	Manually scored respiratory events			**Automatically detected desaturation events**		
Sampling window	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**
Event-specific sampling window	A_{EEM}	A_{ERM}	A_{EFM}	A_{EEA} (DesSev) 1.45 (1.23-1.70)	A_{ERA} 1.49 (1.27-1.75)	A_{EFA} 1.51 (1.28-1.77)
Record-specific sampling window	A_{REM} (HB) 2.59 (1.94-3.48)	A_{RRM} 3.45 (2.50-4.68)	A_{RFM} 4.10 (2.87-5.85)	A_{REA} 1.33 (1.13-1.56)	A_{RRA} 1.38 (1.18-1.62)	A_{RFA} 1.41 (1.20-1.66)
Fixed sampling window	A_{FEM} 1.82 (1.46-2.27)	A_{FRM} 2.10 (1.67-2.66)	A_{FFM} (REDTA) 2.25 (1.75-2.89)	A_{FEA} 1.25 (1.07-1.47)	A_{FRA} 1.29 (1.10-1.52)	A_{FFA} 1.30 (1.11-1.53)

Table 4.4 Desaturation area-based methods predicting CVD mortality in the SHHS with partially adjusted model. The hazard ratios and corresponding 95% confidence intervals are shown for evaluating the performance. P-values shown when less than 0.1.

	Manually scored respiratory events			**Automatically detected desaturation events**		
Sampling window	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**
Event-specific sampling window	A_{EEM}	A_{ERM}	A_{EFM}	A_{EEA} (DesSev) 1.06 (0.89-1.27)	A_{ERA} 1.10 (0.92-1.31)	A_{EFA} 1.08 (0.90-1.29)
Record-specific sampling window	A_{REM} (HB) 1.34 (0.96-1.87) p=0.09	A_{RRM} 1.57 (1.09-2.27) p=0.02	A_{RFM} 1.62 (1.08-2.42) p=0.02	A_{REA} 1.03 (0.87-1.23)	A_{RRA} 1.06 (0.89-1.27)	A_{RFA} 1.05 (0.88-1.26)
Fixed sampling window	A_{FEM} 1.19 (0.94-1.51)	A_{FRM} 1.30 (1.00-1.69) p=0.04	A_{FFM} (REDTA) 1.31 (0.99-1.73) p=0.06	A_{FEA} 1.00 (0.88-1.14)	A_{FRA} 1.03 (0.89-1.18)	A_{FFA} 1.01 (0.89-1.16)

The hazard model is partially adjusted by demographic covariates, smoking status, alcohol intake, and non-CVD-related medical history.

Table 4.5 Desaturation area-based methods predicting CVD mortality in the SHHS with fully adjusted model. The hazard ratios and corresponding 95% confidence intervals are shown for evaluating the performance. P-values are shown when less than 0.1.

	Manually scored respiratory events			**Automatically detected desaturation events**		
Sampling window	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**	**Event-specific baseline**	**Record-specific baseline**	**Fixed baseline**
Event-specific sampling window	A_{EEM}	A_{ERM}	A_{EFM}	A_{EEA} (DesSev) 1.02 (0.79-1.32)	A_{ERA} 1.08 (0.84-1.39)	A_{EFA} 1.02 (0.79-1.32)
Record-specific sampling window	A_{REM} (HB) 1.56 (0.84-2.89)	A_{RRM} 1.79 (1.00-3.19) p=0.04	A_{RFM} 1.69 (0.91-3.15) p=0.1	A_{REA} 0.96 (0.74-1.23)	A_{RRA} 1.00 (0.78-1.28)	A_{RFA} 0.98 (0.76-1.25)
Fixed sampling window	A_{FEM} 1.24 (0.79-1.96)	A_{FRM} 1.53 (0.93-2.52) p=0.09	A_{FFM} (REDTA) 1.47 (0.88-2.46)	A_{FEA} 0.94 (0.82-1.07)	A_{FRA} 0.96 (0.83-1.11)	A_{FFA} 0.95 (0.82-1.09)

The hazard model is fully adjusted by demographic covariates, smoking status, alcohol intake, non-CVD-related medical history, AHI, T90, MinSat, and concurrent cardio-metabolic disease (heart failure, stroke, angina, coronary revascularisation, and myocardial infarction).

4.5 Discussion

This experiment conducted a comprehensive comparison of desaturation area-based methods to evaluate their effectiveness in predicting CVD mortality among middle-aged and older adult cohorts. Variations in event choice, sampling window, and baseline for area-based computational methods were examined, and the predictive performance of each was assessed for CVD mortality. The results indicate that parameters associated with manually scored respiratory events generally demonstrated some evidence of predictive ability for CVD mortality. Among these, methods using record-specific sampling windows achieved the strongest performance in both partially and fully adjusted models. Specifically, in the fully adjusted model, the A_{RRM} method outperformed all others with a significant hazard ratio. Besides, A_{RFM} and A_{FRM} showed significance in the partially adjusted model but became only marginally significant after adjustment for all covariates. By contrast, there was minimal evidence linking CVD mortality to parameters derived from automatically detected desaturation events. Methods associated with the automated desaturation detection algorithms chosen in this thesis consistently produced p-values greater than 0.1, regardless of the level of covariate adjustment, indicating negligible predictive ability for CVD mortality.

With the goal of identifying a desaturation area method that reliably predicts CVD mortality, A_{RRM} emerges as the most promising approach. Its record-specific sampling window is patient-oriented and tailored to each recording, reducing the risk of incomplete capturing of desaturation events, especially in patients with prolonged respiratory events. Additionally, the record-specific baseline offers greater noise tolerance than event-by-event baselines, minimising the influence of recording quality on predictive performance. A_{RRM} strikes an optimal balance between individualised analysis and noise resistance, showing strong potential for CVD mortality prediction. This method represents a significant step toward clinical implementation and could be valuable for early risk stratification in CVD patients.

Moreover, the unadjusted Cox models (**Table 4.3**) generally yielded larger hazard ratios than the covariate-adjusted models (**Table 4.4** and **Table 4.5**). This attenuation after adjustment suggests that part of the association between desaturation area-based metrics and CVD mortality may be explained by established CVD risk factors such as age, BMI, hypertension, and diabetes. In other words, the unadjusted estimates capture both the direct prognostic contribution of the desaturation area-based metrics and its shared variation with traditional

clinical covariates. After controlling for these factors, the adjusted hazard ratios more specifically reflect the independent association of the desaturation area-based metrics with mortality risk. These findings highlight the importance of multivariable modelling when evaluating sleep-related predictors, as apparent effects in univariable analyses may be partly attributable to confounding or overlapping risk pathways. Nonetheless, the persistence of statistically significant associations in adjusted models would support the added prognostic value of desaturation area-based metrics beyond conventional cardiovascular risk factors.

In addition to evaluating predictive performance, this experiment also discussed the reasons underlying the distinct differences between methods based on manually scored respiratory events and those based on automatically detected desaturation events. As shown in **Table 4.4** and **Table 4.5**, methods triggered by automatically detected desaturation events yielded insignificant HR values across both Cox model adjustments, suggesting no predictive ability for CVD mortality. By contrast, methods with the same paired sampling windows and baseline choices but based on manually scored respiratory events demonstrated promising outcomes. For example, A_{FRM} and A_{FRA} showed substantial variation in predictive performance, with A_{FRM} achieving a significant HR (1.53) and AFRA showing an insignificant HR (0.96) in the fully adjusted models. The only difference between these methods was the choice of events, despite both using the same fixed sampling window and record-specific baseline. This finding suggests that event choice can greatly influence the predictive ability of desaturation area-based methods for CVD mortality.

A possible explanation for this inconsistency is the exclusion of arousal events by the automated desaturation event detection algorithm. The automatically detected desaturation events used in this experiment were obtained directly from the oximetry signal. The system applied a moving sampling window across the recording and captured any oxygen desaturation greater than 3%. By contrast, manually scored respiratory events were annotated using multiple signals to identify apnoea and hypopnoea events. As described in previous chapters, the scoring criteria for hypopnoea events incorporate not only oxygen desaturation but also arousals. This difference in annotation means that hypopnoea events associated with arousal may be undercounted when using automated detection algorithms. As exampled in **Figure 4.2**, a hypopnoea event scored in the SHHS database was associated with an arousal that occurred within 5 seconds after the event ended, despite the absence of a desaturation greater than 3%. If the area method employs manually scored respiratory events, this event is still counted

toward the area calculation. However, when using automated algorithms that rely solely on oxygen desaturation, this event is neglected in the calculation, leading to inaccurate estimation of overnight hypoxemic burden. Within the same SHHS recording, manually scored respiratory events outnumbered automatically detected desaturation events by an average factor of five, suggesting that a substantial number of events are excluded when applying automated approaches to desaturation area-based methods.

Similar outcomes were observed by Esmaeili et al., who developed an automated desaturation method, HBoxi, based solely on the oximetry signal with a 2% desaturation threshold, and reported promising results in predicting CVD outcomes [291]. According to Esmaeili et al., applying a 3% threshold (as recommended by the AASM criteria) to the automated method did not yield statistically significant results. This suggests that automated methods using thresholds aligned with AASM criteria may not provide a valid characterisation of overnight hypoxia, thereby reducing their predictive performance for CVD mortality. Although HBoxi was not included in this experiment due to technical limitations, as the authors did not disclose detailed implementation instructions, placing it beyond the scope of this study, their observations regarding the desaturation threshold align with the findings of this experiment. Both indicate that desaturation area-based methods are highly sensitive to the sleep events. This means that even when the sampling window and baseline are held the same, the definition of annotated sleep events can substantially influence the performance of parameters in predicting CVD outcomes. Therefore, in the future development of automated desaturation area-based algorithms, particular attention should be given to the approach used to annotate sleep events.

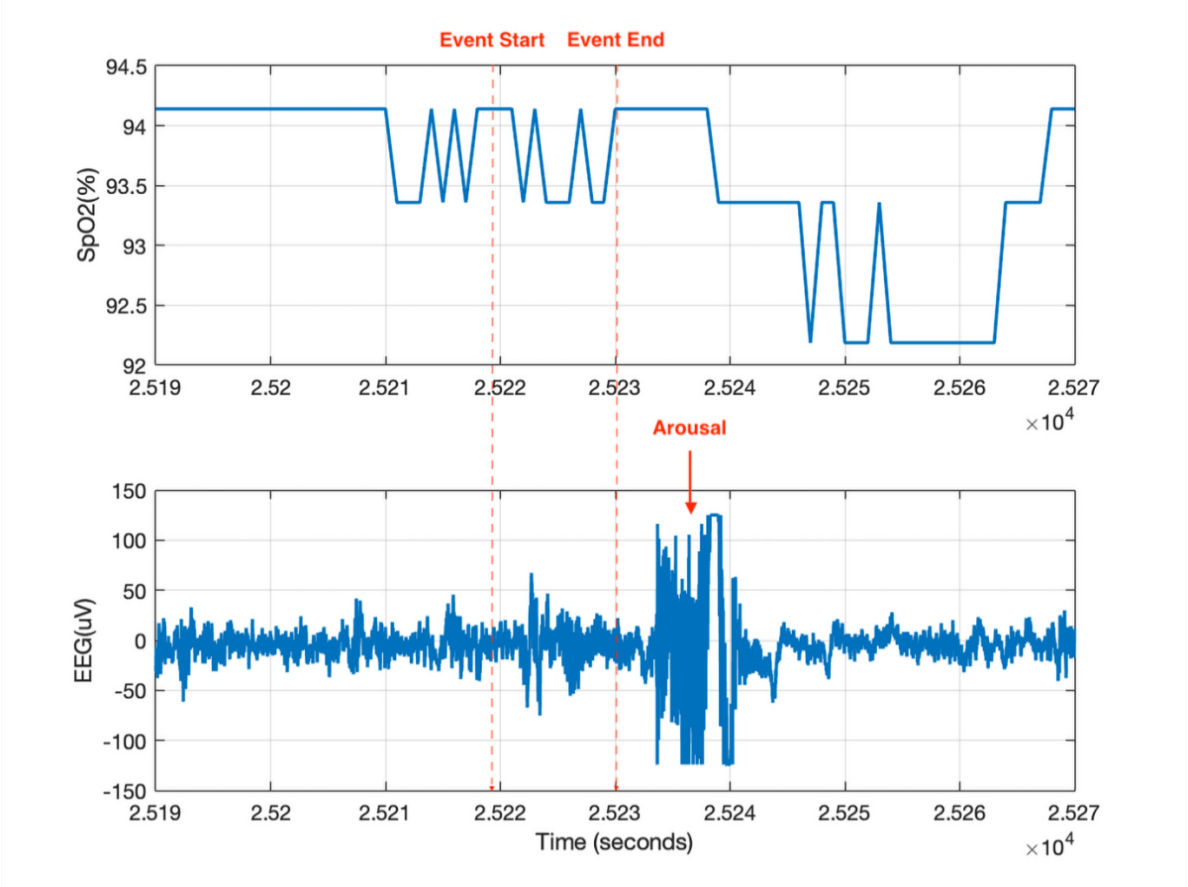


Figure 4.2 An example of a hypopnoea event is one accompanied by an arousal event (identified based on EEG signal) that begins within 5 seconds after the hypopnoea event ends. If the associated desaturation is less than 3% (as illustrated in this figure), such events can only be identified through manual respiratory event scoring, which relies on expert input and multiple signals from overnight PSG. Consequently, the number of manually scored respiratory events is typically higher than that of automatically detected desaturation events.

4.6 Limitations

This experiment has some limitations. The SHHS represents a community-based population of middle-aged and older adults, predominately Caucasian, with limited information on OSA treatment history. Additionally, the duration of OSA in patients is unknown, which may influence the adaption of the cardiovascular system to long-term hypoxemia and hence may influence the oximetry desaturation area-based parameters [295]. In future studies, the comparison could be made in larger and more diverse databases, and there is a need to conduct such experiment on clinical populations or populations with OSA patients only.

Furthermore, the discussion of event scoring has its limitations. Comparisons between manually scored respiratory events and automatically detected desaturation events were conducted in one database. To derive more reliable conclusions regarding the impact of automated desaturation event detection on CVD mortality prediction, further investigations using multiple databases are necessary. Additionally, the automatic detection algorithm used in this study is only one of several available methods and does not represent the performance of all published algorithms. Terrill et al. introduced an algorithm for detecting desaturation events and calculating a dynamic baseline for oximetry data [296] and Esmaeili et al. developed HBoxi [291]. Both approaches could be explored in future research.

Moreover, one potential explanation for the comparatively lower performance of the automated algorithms observed in this experiment is their sensitivity to the choice of desaturation threshold. The DesSev method was originally developed using a 3% desaturation criterion, consistent with the AASM scoring manual. However, previous studies have suggested that a 2% threshold may improve performance in certain contexts [291]. Therefore, it would be valuable to explore whether applying a lower desaturation threshold within the ABOSA software implementation of DesSev could enhance predictive accuracy. Nevertheless, such sensitivity analyses could not be undertaken in the present study because the ABOSA software operates as a closed and highly integrated package. To ensure that the algorithm was implemented exactly as proposed by the original authors, the software had to be used in its native form, which does not permit modification of desaturation thresholds or event annotation criteria. Consequently, threshold-based investigations were not feasible within the current experimental framework.

Chapter 5
Experiment 2

5 Using PSG-derived parameters and explainable machine learning approaches to predict CVD mortality

This chapter investigates the predictive ability of combined PSG-derived parameters and aims to provide individualised CVD mortality outcome estimates for patients through a machine learning approach. The experiment comprises two phases:

- Phase 1 builds on previous analyses by examining whether oximetry-derived parameters can predict 3-year CVD mortality outcomes.
- Phase 2 focuses on identifying an explainable machine learning model that predicts individual-level CVD mortality outcome, incorporating features from demographics, medical history, lifestyle factors, and PSG-derived parameters.

This experiment is distinctive in its holistic assessment of sleep study parameters as an integrated set, moving beyond the traditional focus on single-parameter analysis. By exploring whether internal interactions among PSG-derived features can enhance predictive performance, this study offers new insights into the complex relationship between OSA and CVD. Furthermore, it introduces individualised outcome estimates, a significant step towards practical clinical application that bridges the gap between sleep study data and actionable cardiovascular risk management.

5.1 Rationale

In the previous chapters, OSA has been increasingly recognised as an independent risk factor for CVD, presenting a valuable opportunity for future CVD risk stratification. Overnight PSG, the gold standard for OSA assessment, records a comprehensive set of physiological signals, including respiratory flow, blood oxygen saturation, heart rate, muscle and brain activity, and eye movements. These signals characterise the OSA condition and can provide insights for predicting CVD outcomes. While the traditional metric AHI remains the gold standard for OSA diagnosis, it is a poor independent predictor of CVD mortality, primarily due to its inability to capture the full consequences of respiratory disturbances. Alternative PSG-derived metrics have been explored, and several have demonstrated promising performance in predicting CVD outcomes. Among these, oximetry-derived parameters, such as T90 and ODI3, have

consistently shown strong predictive ability across various CVD outcomes. Desaturation area-based parameters, despite their computational variations, have also emerged as novel and promising indicators of CVD mortality. These are discussed in detail in Chapter 4 [15, 28, 132, 188, 200]. Additionally, the association between sleep metrics and CVD risk extends beyond oximetry-derived measures. For instance, EEG-derived TST demonstrates a U-shaped relationship with CVD risk [297].

However, existing studies predominantly evaluate parameters in isolation, assessing the predictive ability of each parameter individually rather than in combination. This single-parameter focus fails to capture the multifaceted nature of OSA, wherein each parameter reflects only a specific physiological dimension. Relying solely on partial representations of the OSA condition may limit the accuracy of predicting future CVD outcomes [19]. Supporting this, Baumert et al. demonstrated that combining two T90-based measures significantly improved the prediction of CVD mortality compared to single-parameter models, highlighting the value of multi-dimensional predictors that better characterise nocturnal hypoxaemia [6]. While combining features can enhance predictive performance, the relationship is not strictly additive, as the inclusion of excessive or irrelevant variables may degrade model performance. Therefore, rigorous and systematic parameter selection remains critical. Nevertheless, in the context of disease prediction, particularly for complex conditions like heart disease, multi-dimensional parameters that provide complementary insights are often necessary for robust risk stratification [298].

Methodologically, most current studies rely on the Cox model, a traditional survival analysis technique that estimates the relationship between predictors and time-to-event outcomes. The widely used Cox proportional hazards model estimates relative effects in the form of hazard ratios rather than absolute risk probabilities. Although survival models remain essential for time-to-event analysis and can support dynamic assessment of how risk evolves with changes in risk factors over time, relative hazard measures are often difficult for non-statistical end-users, such as patients, to interpret in terms of the severity of their condition. This limitation may reduce the clinical utility of such models for personalised risk assessment. In clinical communication, statements such as *"your sleep measurement corresponds to a hazard ratio of X, indicating an X-fold higher hazard of CVD mortality compared with the baseline level"* may provide limited practical meaning for patients, as the concept of the baseline level is often unclear. As a result, patients may struggle to interpret what such relative risk estimates imply

about the severity of their condition or their personal prognosis. Additionally, modelling the combined effects of multiple variables requires explicit interaction terms, increasing model complexity and potentially reducing interpretability [18].

In contrast, machine learning approaches inherently capture nonlinear interactions between features and can demonstrate personalised survival probabilities at specific time points. These methods provide a more intuitive and actionable framework for clinical decision-making by presenting clear, individual-level CVD outcome estimates. Studies have shown that machine learning survival models achieve performance comparable to, and often exceeding, that of the Cox model, particularly in complex, multivariate contexts such as disease prediction [21-26]. More recent work applying machine learning approaches to metastatic relapse prediction further supports their utility for complex outcome modelling beyond traditional regression-based survival analysis [24]. Given the concerns of single-parameter analyses and the goal of improving individual-level CVD outcome prediction, a methodological shift towards machine-learning approaches may serve as a valuable complement to conventional statistic models. Such approaches can integrate multiple parameters and provide more intuitive risk estimates, thereby supporting clearer communication of cardiovascular risk to patients.

To address the common concern regarding the “black-box” nature of certain machine learning algorithms, this experiment prioritises the use of explainable models. Given the goal of supporting clinical application, it is essential to select models that are interpretable and transparent for end-users, including clinicians and patients. While previous studies have demonstrated strong predictive performance using less-interpretable methods, particularly deep learning, the lack of transparency poses a significant barrier to clinical applications, as clinicians are unlikely to base decisions on models that are difficult to interpret. Explainable machine learning offers a promising solution to these concerns, as it provides greater transparency while also outperforming traditional statistical methods in various medical contexts [18]. Therefore, explainable machine learning provides a flexible, data-driven, and clinically relevant framework for predicting CVD outcomes using PSG-derived parameters.

5.2 Database

5.2.1 Sample selection

To ensure consistency in comparing the predictive abilities of parameters for CVD mortality, this experiment also utilised PSG recordings, corresponding CVD outcomes, demographic data, and medical history from the SHHS cohort. Details of the SHHS cohort have been described in the previous chapter. Samples included in this experiment were selected based on predefined criteria, as outlined in **Figure 5.1**. Specifically, only participants with complete PSG data, covariate data, and CVD outcome data were considered in the experiment. The covariates used in the sample selection process included age, race, BMI, gender, smoking status, daily alcohol intake, and medical history of hypertension, diabetes, heart failure, and hyperlipidaemia. Importantly, while the feature sets differ among the two phases of the experiment, the same samples were used in both phases to control for potential bias introduced by varying sample selection in the training and testing datasets.

5.2.2 Sample characteristics

Of 5804 participants who completed the study, 11 were excluded due to missing PSG data, 711 due to missing covariate data, 393 due to missing CVD mortality outcome data, and 151 due to censored CVD mortality outcome data. This resulted in a final sample of 4.538 participants eligible for the analysis, as depicted in **Figure 5.1**. Within this cohort, samples were categorised using a 3-year cut-off threshold: 3-year CVD death group and 3-year CVD survivor group for the first two phases of the experiment. The 3-year CVD mortality group explicitly excludes 151 censored participants whose follow-up ended within 3 years without an observed CVD event. **Table 5.1** summarised the sample characteristics. In the 3-year CVD survivor group, females comprised 53.20% of the participants, whereas in the 3-year CVD death group, males were more prevalent, representing 58.93%. Participants were predominantly Caucasian in both the survivor group (87.29%) and the mortality group (85.71%). The mean age was 64.23 years in the survivor group and 76.54 years in the mortality group. The mean AHI (17.95 in the survivor group and 23.41 in the mortality group) indicated moderate OSA.

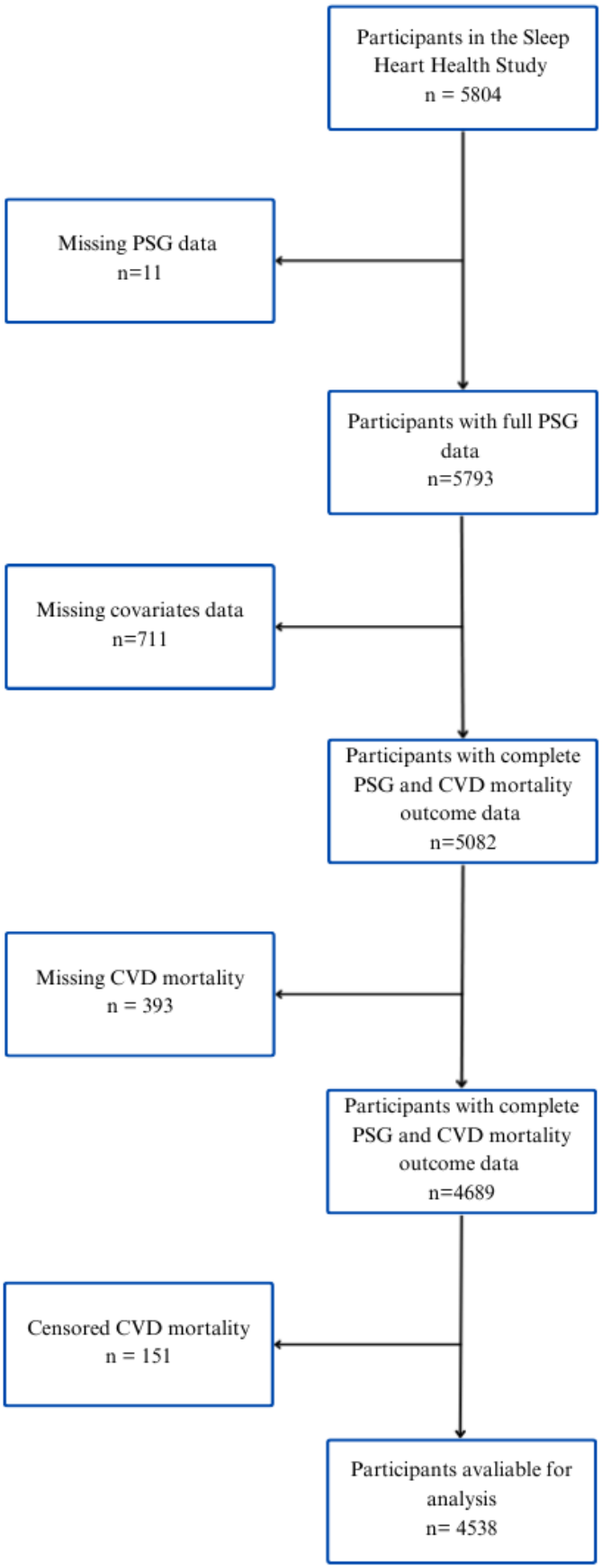


Figure 5.1 Flow chart for the study sample identified for inclusion from SHHS cohort database.

Table 5.1 Sample characteristics of the SHHS involved in the analysis.

Variables	Total n= 4538 (100%)	
	3-year CVD survivor n = 4482 (98.8%)	3-year CVD death n = 56 (1.2%)
Age (years), mean (SD)	64.23 (10.76)	76.54 (7.31)
BMI (kg/m^2), mean (SD)	28.30 (5.08)	26.44 (5.27)
Race		
Caucasian, n (%)	3961 (87.29)	48 (85.71)
Other, n (%)	521 (12.71)	8 (14.29)
Gender		
Male, n (%)	2124 (46.80)	33 (58.93)
Female, n (%)	2414 (53.20)	23 (41.07)
Smoking status		
Never, n (%)	2088 (46.01)	26 (46.43)
Ever, n (%)	2450 (53.99)	30 (53.57)
Alcohol intake (drinks per day)	2.70 (5.74)	2.18 (5.43)
Total time of sleep (TST), n (%)		
5-8h	3854 (84.93)	50 (89.29)
≤ 5h	684 (15.07)	6 (10.71)
T90 (%TST), mean (SD)	3.38 (10.02)	10.07 (20.59)
AHI (events/h), mean (SD)	17.95 (15.72)	23.41 (17.23)
ODI3*(events), mean (SD)	76.17 (75.71)	81 (68.87)
Heart failure, n (%)	70 (1.54)	11 (19.64)
Diabetes, n (%)	313 (6.90)	18 (32.14)
Hypertension, n (%)	1807 (39.82)	41 (73.21)
Lipid-lowering medication use, n (%)	573 (12.63)	12 (21.43)

*The ODI3 used in this experiment was calculated according to the SHHS cumulative definition, where ODI3 represents the total number of oxygen saturation with desaturation >= 3%.

5.3 Methodology

To achieve the goal of personalised CVD mortality outcome forecasting, the experiment was divided into two phases. Phase 1 validated whether a selected group of oximetry-derived parameters can enhance predictive performance for future 3-year CVD mortality, while Phase 2 developed an explainable machine learning model that balances predictive accuracy, computational simplicity, and interpretable decision-making for end-users.

The experiments in both phases followed the same classification system, as illustrated in **Figure 5.2** and detailed in the following sections. The system comprised four main steps, from signal pre-processing to performance evaluation. In the first step, signals (different in two phases) from the SHHS database underwent pre-processing to prepare for subsequent feature extraction. Then, the calculated features, alongside demographic information and medical history, were selected and passed to the classification stage. Finally, the proposed model was evaluated for its predictive performance in forecasting individualised CVD mortality outcomes. In addition, the combined and individual contributions of features were analysed to enhance model interpretability. These steps were iteratively performed to identify optimal feature and classifier combinations that maximise the predictive performance while minimising reliance on specialised clinical inputs and reducing classifier complexity.

This strategy was designed to maximise the model's applicability across a broader population. Models that rely exclusively on features obtained through clinical assessments or specialist manual annotations are inherently limited to individuals with access to professional sleep testing: resources that may not be available in medically underserved regions or for individuals facing financial constraints (as discussed in Chapter 2). By focusing on general features such as basic medical history, lifestyle factors, or unattended measurable parameters, the model has the potential for broader application in primary care settings or home environments, thereby supporting large-scale population screening. Furthermore, prioritising simpler and more interpretable models enhances the likelihood of adoption by end-users, including both clinicians and patients, and improves feasibility in low-resource settings or portable platforms.

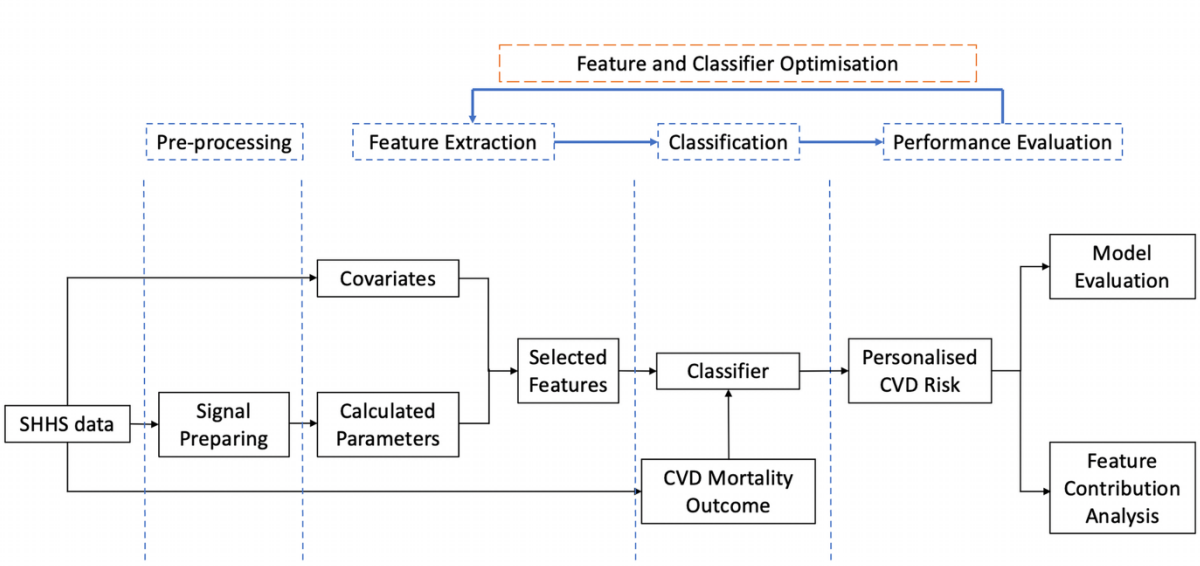


Figure 5.2 Block diagram of the classification system used in this experiment.

5.3.1 Phase 1: Can oximetry-derived parameters effectively predict CVD outcomes?

Phase 1 served as an extension of the previous chapters, where the individual predictive abilities of oximetry-derived parameters were summarised, and the influence of computational differences on CVD mortality prediction was discussed. As demonstrated in Section 3.1.5, although oximetry-derived parameters have shown promise in predicting CVD outcomes, existing studies have rarely explored the combined utility of these parameters and their interactive effects on prediction. Therefore, Phase 1 specifically focused on evaluating the combined predictive power of oximetry-derived parameters in relation to CVD mortality. This analysis also aims to provide preliminary guidance for feature selection in the subsequent phase of the study.

5.3.1.1 Data preparation

This study analysed 4,246 SpO2 recordings, along with corresponding demographic information, lifestyle habits, and 3-year CVD mortality outcomes. The overnight SpO2 signals were pre-processed using a 50% SpO2 cut-off to remove sensor movement artefacts [299]. Manually annotated sleep stages were incorporated to distinguish sleep from wake periods, ensuring all features were calculated during sleep only [16, 80]. Participants were categorised into two classes based on 3-year CVD mortality outcomes: those who died within 3 years (n = 56) [mortality class] and those who lived for 3 or more years (n = 4,190) [survivor class].

Due to the highly imbalanced nature of the dataset, comprising 56 positive cases and 4190 negative cases, class weighting was applied to mitigate the bias toward the majority class. As detailed in Chapter 3.4.5, the class weights were calculated using a simplified inverse-frequency weighting scheme, as shown in Equation 3.39:

$$\omega_C = \frac{1}{N_C} \tag{3.39}$$

where ω_C is the weight assigned to class C, N_C is the total number of samples within class C [277]. In this study, the minority class was assigned a weight of $\omega_{mortality} = \frac{1}{56} \approx 0.0179$, while the majority class received $\omega_{survivor} = \frac{1}{4190} \approx 0.00024$. These weights were incorporated into the class prior probabilities, which influence the discriminant function of the LDA classifier, thereby balancing the contribution of each class during classification.

5.3.1.2 Feature extraction

To evaluate the interactive contribution of oximetry-derived parameters in predicting 3-year CVD mortality, this study selected features from self-reported data and oximetry signals, following a primary feature selection principle: prioritising oximetry-derived parameters while including only simple, low-risk features unlikely to be misreported, such as basic demographics and lifestyle habits. This approach aims to minimise the influence of potentially inaccurate self-reported data and better reflects the true predictive value of oximetry-derived features.

The selected features were grouped into three models for comparative analysis, as summarised in **Table 5.2**. Model A (baseline) included demographic and lifestyle features; Model B included only oximetry-derived parameters; and Model C combined all features from Models A and B. Age were log-transformed, ensuring they are suitable for the subsequent classification.

Demographic features included age (in years), BMI, race (Caucasian or non-Caucasian), and gender (male or female). Age was recorded at the time of the SHHS experiment, with any value above 90 capped at 90 years. BMI was calculated using height and weight measured at the time of study, using the standard formula:

$$BMI\ (kg/m^2) = \frac{Weight\ in\ kilogram}{Height\ in\ metre^2} \tag{5.1}$$

Lifestyle features comprised two major CVD risk factors: smoking status (categorised as never or ever) and alcohol intake (quantified as the number of drinks per day). The oximetry-derived parameters: T90, ODI3, and HB, were computed for the sleep period only, as prior findings (in Chapter 3 and 4) highlighted their strong predictive performance for CVD mortality.

As shown in **Figure 5.3**, T90 quantifies cumulative hypoxemic insult by measuring the total duration of sleep during which oxygen levels fall below 90% [15]. ODI3 is a commonly used metric to indicate intermittent hypoxemia. It measures the number of oxygen desaturation events greater than 3% per during sleep. The calculation method involves identifying desaturation events with at least a 3% decrease in oxygen levels from the baseline [15]. This experiment employs SHHS cumulative definition of ODI3, which is the total number of oxygen saturation with desaturation >= 3%. HB is a widely used desaturation area-based parameter that calculates the sum of desaturation areas per hour of sleep. Each desaturation area is associated with apnoea or hypopnoea events that involve at least a 3% oxygen levels drop or arousal events occurring within 5 seconds. The HB method applied in this experiment follows the original approach proposed by Azarbarzin et al. The area is measured as the space between the SpO2 trace and the pre-event baseline, within a recording-specific sampling window. This sampling window is defined as the interval between two peaks of the averaged events in a recording and remains consistent for all events in that recording. The pre-event baseline is determined as the maximum SpO2 value observed within 100 seconds prior to the end of each event [12].

Table 5.2 Summary of three feature combinations used in Phase 1.

Model	Features
A	Demographic information: Age, BMI, Race, and Gender Lifestyle habits: Smoking status and Alcohol intake
B	Oximetry-derived features: T90, ODI, and HB
C	Demographic information Lifestyle habits Oximetry-derived features.

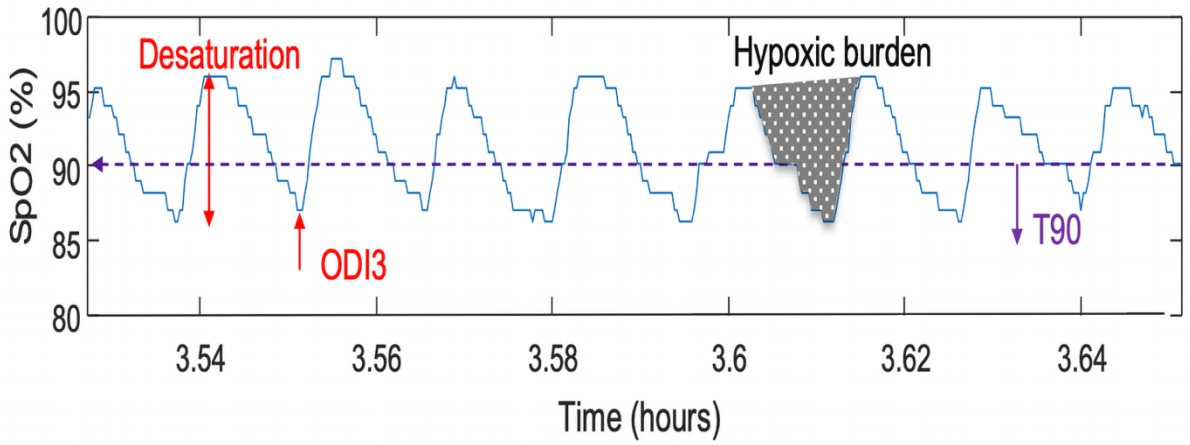


Figure 5.3 The oximetry-derived parameters used in this study are illustrated with each parameter highlighted in a different colour for clarity. T90 (in purple) represents the total duration of sleep with SpO2 levels below 90%. ODI3 (in red) is an event-based parameter that counts the number of desaturation events where SpO2 decreases by more than 3%. Hypoxic burden (in grey) is a desaturation area-based parameter that measures the total area between the baseline and SpO2 trace associated with desaturation events.

5.3.1.3 Classification

As Phase 1 focuses on evaluating the predictive ability of oximetry-derived features, a simple and interpretable classification model was employed: LDA, a traditional discriminant classifier. LDA aims to identify a linear decision boundary where the prior probabilities of the classes are equal, under the assumption of multivariate normality and homoscedasticity. This choice offers computational efficiency and allows for straightforward assessment of both individual and combined feature contributions. Given the highly imbalanced nature of the dataset, Weighted LDA was applied by incorporating class weights into the calculation of the within-class and between-class scatter matrices. This weighting scheme increases the influence of the minority class (positive cases) during the determination of the discriminant direction.

Mathematically, the objective of Weighted LDA is to identify the projection vector ω that maximises the ratio of between-class variance to within-class variance, also known as the weighted multivariate Fisher criterion:

$$J(\omega) = \frac{\omega^T S_B^{weighted} \omega}{\omega^T S_W^{weighted} \omega} \tag{5.2}$$

where ω is the optimal projection direction that enhances class separability. $S_B^{weighted}$ and $S_W^{weighted}$ denote the weighted between-class and within-class scatter matrices, respectively, as detailed in Chapter 3.4.2 [227].

5.3.1.4 Performance evaluation

As this study specifically aimed to evaluate the interactive effects of oximetry-derived parameters in predicting CVD mortality, the optimisation step shown in the classification system (**Figure 5.2**) was not applied in this phase. Model performance was assessed using 10-fold cross-validation. The dataset was randomly partitioned into 10 folds, with each fold serving once as the testing set while the remaining nine folds were used for training. To secure a fair comparison, the same random fold separation was maintained across all models. For each model, performance metrics including sensitivity, specificity, accuracy, and F1 score were calculated based on the confusion matrix from each fold and then averaged. Standard errors (SE) of the metrics were also computed to quantify the variability. To ensure that performance comparisons were statistically robust, the Wilcoxon signed-rank test was employed to assess whether differences between models were statistically significant [281].

While the combined predictive ability of oximetry-derived parameters was assessed using performance metrics, the individual contribution of each parameter was evaluated using the univariate Fisher score:

$$J = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{5.3}$$

where μ and σ represent the mean and variance of a given feature within each class, respectively [300]. Although the multivariate Fisher criterion shares the same underlying principle: maximising class separability, the univariate version evaluates each parameter independently and allows for individual ranking. In this study, the univariate Fisher score was employed to quantify the discriminatory power of each oximetry-derived parameter and to inform feature selection in the subsequent Phase 2 analysis.

5.3.2 Phase 2: Explainable machine learning model for predicting 3-year CVD mortality outcome

At this phase of the experiment, the research focus shifted from evaluating the predictive performance of oximetry-derived parameters for CVD mortality to developing an explainable model that utilises sleep measurements to provide personalised assessments suitable for clinical

application. Current OSA–CVD analyses predominantly rely on the Cox model to assess the association between sleep measurements and CVD outcomes. While informative, this approach estimates relative hazard ratios across groups defined by a particular sleep metric and does not generate individualised predictions, limiting its clinical applicability. To serve as a complementary, this study proposed an explainable machine learning model that builds upon the previously identified PSG-derived parameters to provide accurate, individual-level predictions of CVD mortality outcome. Importantly, the model maintains transparency in its decision-making process while minimising reliance on medical resources, thereby enhancing both its trustworthiness and broader utility for end-users. This phase of the experiment began with building the predictive model for 3-year CVD mortality. Once the optimal combination of models and features was established, the analysis was extended to 5-year and 10-year CVD mortality to evaluate the generalisability of the proposed machine learning model.

5.3.2.1 Data preparation

This phase of the study retained the data preparation strategies employed in Phase 1. The SpO2 signal was pre-processed using a 50% cut-off threshold, and CVD mortality outcomes were categorised based on a 3-year period. To address class imbalance, class weights were again applied during model training. In addition to the methods used in Phase 1, this phase incorporated manually annotated EEG and airflow signals, which were used to identify non-REM and REM sleep stages as well as key sleep events (arousals, apnoea, and hypopnoea) for feature extraction. Sleep stages and arousal events were scored by experts using 30-second epochs based on EEG signals, while apnoea and hypopnoea events were identified using airflow measurements over 2- or 5-minute windows [168, 169]. Obstructive apnoea was defined as a ≥75% reduction in airflow lasting at least 10 seconds, whereas hypopnoea was defined as a ≥30% reduction in airflow of the same duration.

5.3.2.2 Feature extraction

The feature extraction process in this study followed a key principle: prioritising features that require minimal expert input while remaining concise and informative, yet still achieving strong predictive performance for CVD mortality outcome. This strategy aims to broaden the model's applicability, enabling effective CVD outcome assessments for populations in medically underserved regions or for individuals experiencing financial burden, without necessitating extensive clinical resources.

The optimisation process for feature selection in this study comprised two stages: preliminary feature selection and comprehensive feature selection, as detailed in **Table 5.3**. In the preliminary stage, features were manually selected based on their minimal reliance on clinical resources. These included: (i) variables that can be self-reported without medical assessment (e.g., demographic information and lifestyle habits); (ii) common medical history items that are widely understood by the general population and are recognised as independent risk factors for both CVD and OSA; and (iii) unattended measurable parameters that can be obtained using portable devices (e.g., smartwatches and fitness trackers), thereby reducing dependence on specialised equipment. The final model then underwent comprehensive feature selection, which was guided by performance outcomes from both Phase 1 and Phase 2. Feature combinations were iteratively refined to optimise predictive performance while maintaining interpretability and clinical applicability.

Table 5.3 Summary of features used in two stages in Phase 2.

Stage of feature selection	Features
Preliminary stage	(i) Age, BMI, race, gender, smoking status, and alcohol intake (ii) Hypertension, diabetes, heart failure, and hyperlipidaemia (iii) HB, T90, ODI3, TST, and AHI
Comprehensive stage	(i) Age, BMI, race, gender, and alcohol intake (ii) Hypertension, diabetes, and heart failure (iii) HB, T90, TST, and AHI

In addition to the features used in Phase 1 (**Table 5.2**), this study incorporated two additional PSG-derived parameters: TST, and AHI. TST, derived from EEG recordings, is calculated by summing the duration of manually identified non-REM and REM sleep stages. It has been recognised as a promising predictor of adverse CVD outcomes [203]. AHI, the standard metric for OSA diagnosis, is manually calculated based on airflow, with specific scoring criteria detailed in Section 5.3.2.1. Although it has been widely acknowledged that AHI inadequately captures the full respiratory disturbances, often resulting in suboptimal predictive performance, it was still included in this study. This decision is based on the recognition that univariate analyses may overlook important interactive effects, and that features with limited individual predictive power can nonetheless contribute meaningfully to model performance when used in combination with other variables [11, 27, 301].

Medical history in this study was intentionally simplified to include only conditions that are widely understood by the general population. The response format was binary ("ever" or "never") to minimise ambiguity and facilitate self-reporting. The included conditions: hypertension, diabetes, hyperlipidaemia, and heart failure, are all strongly associated with both CVD and OSA, as detailed in Chapter 2.

5.3.2.3 Classifiers

Given that the objective of this study is to classify participants into appropriate CVD mortality groups (i.e., 3-year CVD death versus 3-year CVD survivor), the choice of classifiers was restricted to supervised learning methods. Furthermore, to align with the study's goal of developing an explainable model with a transparent decision-making process, only interpretable models were considered. These included two single classifiers: LDA and SVM, and two decision tree-based ensemble learning methods: RF (bagging) and XGBoost (boosting). Neural networks and deep learning models were excluded due to their inherent complexity and black-box decision logic, which can undermine trust and hinder practical application by end-users. Additionally, for tabular data as used in this study, boosting methods, particularly XGBoost, have consistently demonstrated superior performance [266, 302, 303].

This study considered LDA as the baseline model, as detailed in Phase 1. The linear SVM included in this study is a margin-based classifier that identifies a linear decision boundary by maximising the margin between classes. This is achieved by solving a quadratic programming problem under the soft-margin formulation:

$$min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{l} \xi_i \tag{3.20}$$

where ξ_i is the slack variable that allow margin violation and C is the user-defined regulation term that controls the trade-off between maximising the margin and minimising errors [242, 243]. The key difference between LDA and SVM is that SVM does not rely on any distributional assumptions about the input features, and is generally less prone to overfitting due to its margin-maximising framework and regularisation. Class weights can also be incorporated into SVM to address class imbalance during training.

Two decision tree-based ensemble learning models employed in this study were RF and XGBoost. While both utilise decision trees as base learners, they differ fundamentally in their training strategies, as detailed in Chapter 3. RF follows a bagging approach, constructing multiple fully grown trees on bootstrapped subsets of the training data and aggregating their

predictions through majority voting. To mitigate the effects of class imbalance, RF typically incorporates class weights during model training [250]. In contrast, XGBoost implements a boosting strategy by sequentially building an ensemble of shallow trees, where each successive tree aims to correct the residual errors of its predecessors. Owing to its gradient-based optimisation framework, XGBoost is capable of handling class imbalance intrinsically, and thus does not mandatorily require the application of class weights [251]. Both models underwent hyperparameter tuning using Bayesian Optimisation (See Chapter 3, section 3.4.4.2) to efficiently enhance predictive performance [271]. Although these models are less interpretable than linear classifiers such as LDA or SVM, an external explainability framework was applied to improve transparency and support interpretability of the predictive results.

5.3.2.4 Model optimisation and Feature selection

To develop an explainable model, the optimisation process in this study comprised two consecutive steps: a preliminary analysis and a comprehensive analysis, as shown in **Figure 5.4**.

In the preliminary stage, a subset of features was manually selected based on their minimal reliance on clinical expertise. These included demographic information, lifestyle factors, common medical history, and unattended measurable parameters, as described in **Table 5.3**. Four machine learning models: LDA (baseline), SVM, RF, and XGBoost, were trained and evaluated using identical feature sets and random seed. Performance metrics and SHAP analysis (discussed in the next section) were used to assess model performance and feature contributions. This stage enabled an initial comparison and identification of the top two performing models and potential optimal feature combinations for the next stage of analysis.

In the comprehensive analysis, feature selection was refined using insights from Phase 1 and the preliminary results. Features demonstrating limited individual contribution or high collinearity were excluded from further consideration. The top two models from the preliminary stage were re-evaluated using the optimised feature subsets. SHAP analysis was employed throughout this phase to support iterative refinement of feature selection. This iterative process was essential, as univariate Fisher scores do not account for interactions between features, and SHAP values, while informative, reflect marginal contributions and can be influenced by feature collinearity. Therefore, relying solely on these two analyses for feature selection may be insufficient and potentially hinder model performance, as features with

seemingly low individual importance may contribute significantly when combined with others. Repeated evaluations enabled the identification of an optimal feature set for the selected classifier, ultimately facilitating the development of a high-performing, explainable model suitable for clinical end-users.

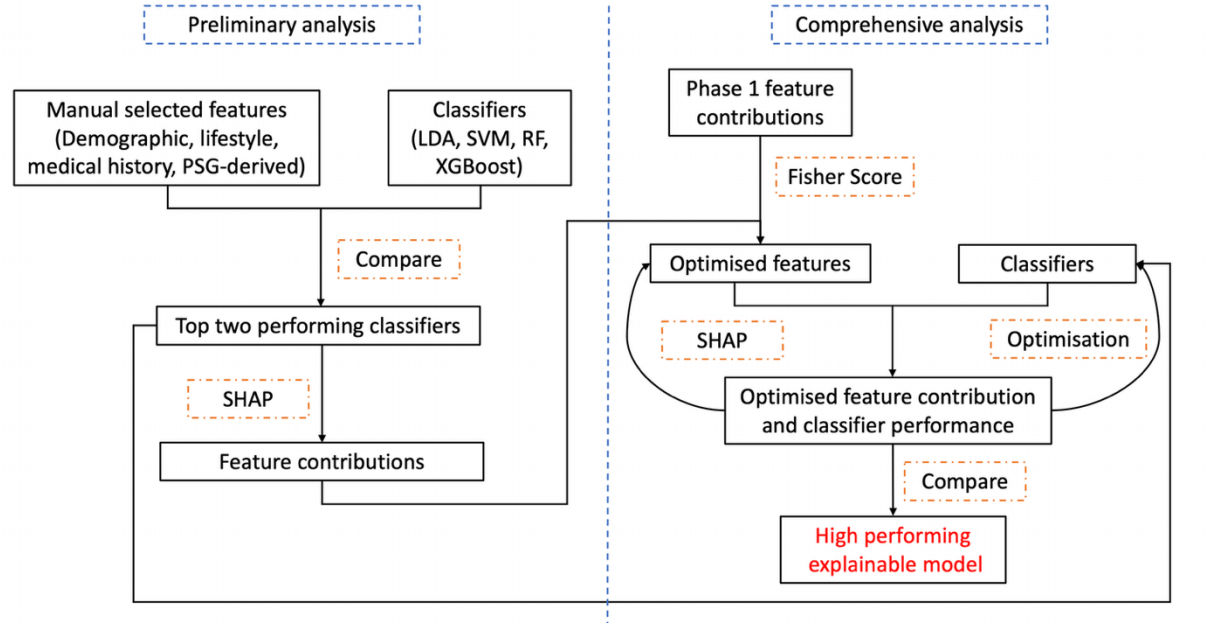


Figure 5.4 Summary of the two-stage analysis conducted in Phase 2, detailing the iterative process of feature and classifier optimisation. The preliminary stage involved manual selection of features requiring minimal clinical input and comparison of selected explainable classifiers (LDA, SVM, RF, XGBoost). The comprehensive stage refined feature selection based on Phase 1 insights, SHAP analysis, and model performance in preliminary stage, ultimately identifying the optimal feature set and best-performing explainable model for predicting 3-year CVD mortality.

5.3.2.5 Performance evaluation

The performance evaluation stage primarily followed the strategies outlined in Phase 1, including 10-fold cross-validation and the use of standard performance metrics: sensitivity, specificity, accuracy, and F1 score. To ensure the reliability of performance comparisons, SE was calculated, and the Wilcoxon signed-rank test was applied to assess statistical significance. Unlike Phase 1, this phase incorporated the optimisation steps shown in **Figure 5.2**, which involved iterative feature selection and classifier tuning to identify the optimal model configuration.

In addition to standard performance metrics, this study also incorporated the ROC curve and the AUC to provide a more comprehensive evaluation of model performance. The ROC curve shows the trade-off between sensitivity and specificity across a range of decision thresholds (with 0.5 as the default), offering insight into the model's performance under varying classification criteria. AUC, as a threshold-independent metric, quantifies the model's overall ability to discriminate between positive and negative cases. This is particularly valuable in the context of imbalanced datasets, as in the present study.

The ensemble learning models were interpreted using SHAP analysis, which quantifies the contribution of each feature to the model's output. SHAP assumes that a model's prediction can be represented as the sum of individual feature effects. Feature importance was visualised using a bee-swarm plot, in which features are ranked from top to bottom according to their mean absolute SHAP values. Higher absolute SHAP values indicate greater influence on the model's predictions, while the horizontal spread of points illustrates the variability in each feature's contribution across individual instances.

5.3.3 Extension of Phase 2: Application of the best-performing explainable model to predict 5-year and 10-year CVD mortality outcomes

Phase 2 proposed an explainable model with an optimised feature combination for predicting 3-year CVD mortality outcomes. While this time horizon was selected as a preliminary step to facilitate model development and evaluation, a 3-year cutoff may not be ideal for all clinical purposes. The impact of a best-performing machine learning model based on features with minimal clinical reliance could be maximised if the model not only demonstrated strong predictive ability for a particular time horizon but also generalised to longer time frames. Such flexibility would enable the model to be applied to broader populations while accommodating different clinical needs. Therefore, an extension of Phase 2 was conducted to assess whether the proposed model could reliably predict CVD mortality outcomes across different time horizons. Specifically, this extended study evaluated the model's ability to predict 5-year and 10-year CVD mortality outcomes.

5.3.3.1 Data preparation

This study followed the same data preparation protocol as Phase 2, with the only modification being the selection of time horizons (5 and 10 year). Of the 4246 samples used in Phases 1 and 2, 4035 were eligible for the 10-year CVD mortality analysis, as 211 more participants

discontinued follow-up within 10 years and were therefore excluded from outcome classification. For the 5-year CVD mortality prediction, 113 participants were categorised as 5-year CVD deaths, and 4133 as 5-year CVD survivors. For the 10-year prediction, 289 participants were identified as CVD deaths within 10 years, while 3746 were classified as survivors.

5.3.3.2 Summary of this study

The results of Phase 2 demonstrated that XGBoost, utilising the comprehensively selected feature set (**Table 5.3**), was the optimal explainable model for predicting 3-year CVD mortality outcomes. Building upon these findings, the extended study adopted the same model (XGBoost) and feature set—comprising age, gender, BMI, alcohol intake, hypertension, diabetes, heart failure, HB, T90, TST, AHI, and ODI3. The performance evaluation framework remained consistent with that of Phase 2, with the proposed model assessed separately for 5-year and 10-year CVD mortality predictions using standard performance metrics, AUC and ROC curves, and SHAP analysis.

5.4 Results

This section presents the results of the proposed experiments conducted in Phase 1 and Phase 2. The results from Phase 1 are reported independently, focusing on the evaluation of oximetry-derived parameters. In contrast, the results from Phase 2 build upon those findings by incorporating the outcomes of Phase 1 as a reference in the feature optimisation process.

5.4.1 Phase 1: Can oximetry-derived parameters effectively predict CVD outcomes?

To evaluate the combined predictive ability of oximetry-derived parameters for 3-year CVD mortality and their potential to enhance model performance, three models were developed and tested using Weighted LDA as the classifier. The predictive performance of these models is summarised in **Table 5.4**, and the individual contribution of each feature is visualised in **Figure 5.5**.

As shown in **Table 5.4**, Model A, which incorporates demographic information and lifestyle habits, demonstrates high sensitivity (83.87% for the training set and 80.36% for the testing set) with slightly lower specificity (71.59% and 71.50%, respectively). The small SE values of all models indicate stable performance estimates. The F1 score of 75.66% highlights the

model's balanced ability to capture both precision and sensitivity. In contrast, Model B, which includes only oximetry-derived parameters, achieves the highest accuracy across both datasets but yields lower F1 scores (58.10% for training and 53.65% for testing). Model C, which integrates features from both Models A and B, achieves the best overall performance. It delivers the highest sensitivity (85.97% for training and 82.50% for testing) and maintains good specificity. The F1 scores improve by approximately 2% compared to Model A, and the improvements are statistically significant ($p < 0.001$) in both datasets.

The corresponding feature contributions in Model C are shown in **Figure 5.5**, based on univariate Fisher scores, which assess each feature's discriminative power independently. Features are ranked from most to least important according to their Fisher scores. Age exhibited the highest discriminative ability (Fisher score = 1.05), followed by T90 (0.10), BMI (0.07), HB (0.06), and gender (0.04). In contrast, alcohol intake, race, ODI3, and smoking status had Fisher scores below 0.01, indicating limited individual discriminative power.

Table 5.4 Performance of three feature combinations using Weighted LDA classifier predicting 3-year CVD mortality.

Model**	**Training***				**Testing***			
	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE
A	83.87 0.30	71.59 0.06	71.75 0.06	77.24 0.14	80.36 1.46	71.50 0.18	71.62 0.17	75.66 0.63
B	45.54 0.92	80.28 0.38	79.85 0.37	58.10 0.67	40.36 2.69	80.19 0.35	79.71 0.36	53.65 2.48
C	85.97 0.23	73.07 0.05	73.24 0.04	79.00 0.09	82.50 1.41	73.04 0.27	73.16 0.25	77.47 0.56

* Values are expressed as percentages (%). SE: Standard Error

** Model A includes demographic information, smoking status, and alcohol intake as features. Model B uses oximetry-derived features: T90, ODI, and HB. Model C has all features used in Model A and Model B. The performance improvements observed across models are statistically significant.

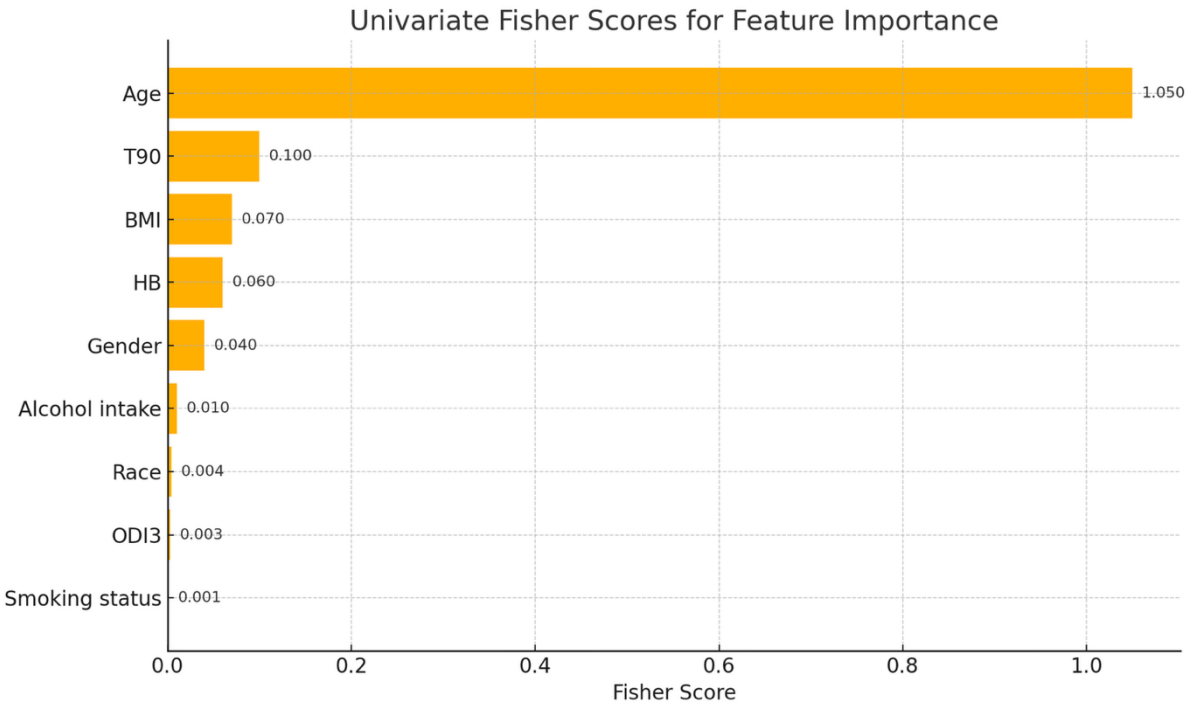


Figure 5.5 Univariate Fisher scores demonstrating individual feature contributions to the performance of Model C. Features are ranked from highest to lowest contribution, with greater Fisher scores indicating higher predictive importance.

5.4.2 Phase 2: Explainable machine learning model for predicting 3-year CVD mortality outcome

To determine an explainable machine learning model, Phase 2 comprised two stages: preliminary analysis and comprehensive analysis. In the preliminary stage, 3-year CVD mortality was predicted using manually selected features applied to a set of explainable classifiers (LDA, SVM, RF, XGBoost), along with an evaluation of individual feature contributions. In the comprehensive stage, the feature set was further refined based on insights from Phase 1 and the preliminary results. The top two performing models were then re-evaluated using the optimised feature sets (**Table 5.3**), and their feature contributions were reassessed.

The preliminary stage results encompass the performance metrics of selected machine learning models in predicting 3-year CVD mortality using manually selected feature sets. As summarised in **Table 5.5**, the baseline model LDA achieved F1 scores of 84.11% (training) and 81.53% (testing). SVM showed improved performance, with F1 scores of 84.65% and

82.89%. Notably, SVM maintained specificity and accuracy above 80% across both datasets, indicating a more balanced classification compared to LDA, and was identified as one of the top-performing models in this stage. In contrast, RF exhibited slightly inferior performance relative to LDA, with lower metrics across most categories except for training sensitivity (93.07% vs. 88.69%). Despite acceptable training performance, RF's generalisability was compromised, as evidenced by a notable drop in testing F1 score (84.06% vs. 78.55%), suggesting potential overfitting. XGBoost outperformed all other models, achieving the highest metrics overall, with accuracy of 84.70% (training) and 84.50% (testing). Its F1 scores surpassed those of the baseline by over 4%, confirming its superior predictive capability in this preliminary evaluation. These improvements were statistically significant ($P < 0.05$), and the consistently small SEs across performance metrics showed the stability of the results.

In addition to the performance metrics reported in **Table 5.5**, the best-performing model, XGBoost, was further evaluated using the ROC curve, AUC, and SHAP analysis to assess its overall discriminative ability and to interpret feature contributions for latter comprehensive analysis. The averaged ROC curve from 10-fold cross-validation yielded a mean AUC of 0.87 ± 0.08, with the curve showing a consistently high true positive rate and low false positive rate across thresholds, indicating strong classification performance (**Figure 5.6**A). SHAP analysis was used to rank the features by their contribution to the model's predictions, as shown in **Figure 5.7**A. Age, hypertension, and diabetes emerged as the most influential predictors, with higher values associated with an increased risk of CVD mortality. Other key contributors included PSG-derived features such as HB, TST, and AHI, as well as demographic BMI. The presence of both positive and negative SHAP values for these features suggests non-linear relationships or interactions with other variables. In contrast, features such as gender, hyperlipidaemia, smoking status, and race contributed minimally, as evidenced by SHAP values concentrated near zero.

Based on insights gained from the preliminary analysis and Phase 1 results, the comprehensive study iteratively refined the feature set to include: age, gender, BMI, alcohol intake, hypertension, diabetes, heart failure, HB, T90, TST, AHI, and ODI3, as detailed in **Table 5.3**. While not all features exhibited strong individual contributions in earlier analyses, their interactions with higher-performing features were shown in results to enhance overall model performance, justifying their inclusion. The top two models from the preliminary stage were re-evaluated using these optimised features. The best-performing model was further assessed

using ROC curve, AUC, and SHAP analysis to evaluate both discriminative performance and model interpretability.

As shown in **Table 5.6**, XGBoost consistently maintained the highest performance, outperforming SVM across all metrics in both training and testing datasets, with minimal SEs indicating stable model performance. Notably, XGBoost achieved F1 scores of 88.17% (training) and 86.20% (testing), approximately 4% higher than those of SVM. These improvements were statistically significant (P-value < 0.05). The steadily rising ROC curve and the improved mean AUC of 0.89 ± 0.05, compared to 0.87 in the preliminary analysis, demonstrate the effectiveness of the feature selection optimisation (**Figure 5.6**B). The cumulative testing confusion matrix for SVM and XGBoost are shown in **Table 5.7** and **Table 5.8**.

Regarding feature contributions, the ranking remained largely consistent, as shown in **Figure 5.7**B. Age, hypertension, TST, and diabetes continued to be the dominant predictors. Although the relative contribution of PSG-derived parameters was somewhat reduced compared to earlier models, they still provided added value, particularly in supporting the classification of one class. For instance, higher values of these parameters were associated with an increased likelihood of positive cases, while lower values did not show a comparable effect. The contribution of alcohol intake increased in this model, with higher intake aiding the classification of positive cases. Although gender appeared to contribute minimally with SHAP values around zero, its exclusion resulted in decreased model performance.

Table 5.5 Performance of selected explainable machine learning models for predicting 3-year CVD mortality, following preliminary feature selection.

Model**	**Training***				**Testing***			
	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE
LDA	88.69	79.97	80.09	84.11	85	79.78	79.84	81.53
	0.29	0.29	0.29	0.27	5.24	0.89	0.87	2.53
SVM	88.89	**80.80**	**80.91**	**84.65**	**86.67**	**80.74**	**80.81**	**82.89**
	0.30	**0.27**	**0.27**	**0.25**	**4.84**	**0.68**	**0.64**	**2.24**
RF	**93.07**	76.76	76.98	84.06	82.67	76.75	76.82	78.55
	0.73	1.53	1.51	1.03	5.57	1.77	1.74	2.92
XGBoost	**91.86**	**84.60**	**84.70**	**88.06**	**89.33**	**84.44**	**84.50**	**85.97**
	0.60	**0.54**	**0.53**	**0.36**	**4.61**	**1.15**	**1.11**	**2.21**

* Values are expressed as percentages (%).

** The performance improvements observed from LDA to XGBoost are statistically significant across all evaluation metrics. (p value<0.05)

Table 5.6 Performance of selected explainable machine learning models for predicting 3-year CVD mortality, following comprehensive feature selection.

Model**	**Training***				**Testing***			
	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE
SVM	87.70	80.68	80.77	84.04	87.77	80.50	80.59	83.72
	0.38	0.20	0.20	0.24	2.72	0.40	0.39	1.14
XGBoost	**91.87**	**84.80**	**84.89**	**88.17**	**88.00**	**85.89**	**85.21**	**86.20**
	0.52	**0.54**	**0.53**	**0.36**	**3.44**	**0.89**	**0.88**	**1.77**

* Values are expressed as percentages (%).

** The performance improvements observed are statistically significant across all evaluation metrics. (p value<0.05)

Table 5.7 The cumulative confusion matrix of SVM predicting 3-year CVD mortality, following the comprehensive feature selection.

		Predicted class label*	
		Positive	Negative
Actual label*	Positive	49	7
	Negative	874	3608

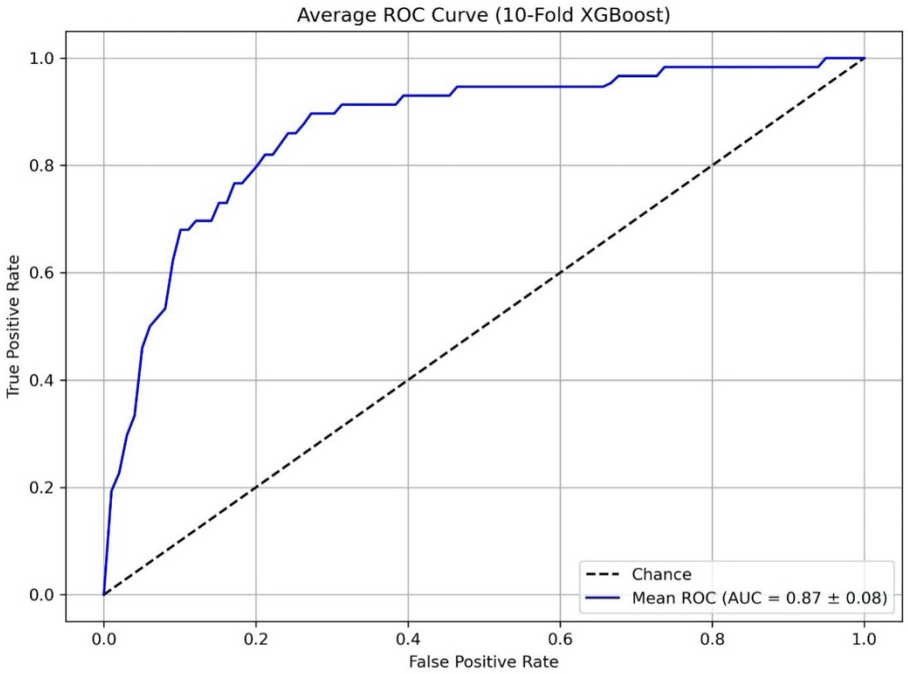
* The positive class represents participants who experienced 3-year CVD death, whereas the negative class represents individuals who survived beyond 3 years.

Table 5.8 The cumulative confusion matrix of XGBoost predicting 3-year CVD mortality, following the comprehensive feature selection.

		Predicted class label*	
		Positive	Negative
Actual label*	Positive	49	7
	Negative	634	3848

* The positive class represents participants who experienced 3-year CVD death, whereas the negative class represents individuals who survived beyond 3 years.

(A) Preliminary Feature Selection



(B) Comprehensive Feature Selection

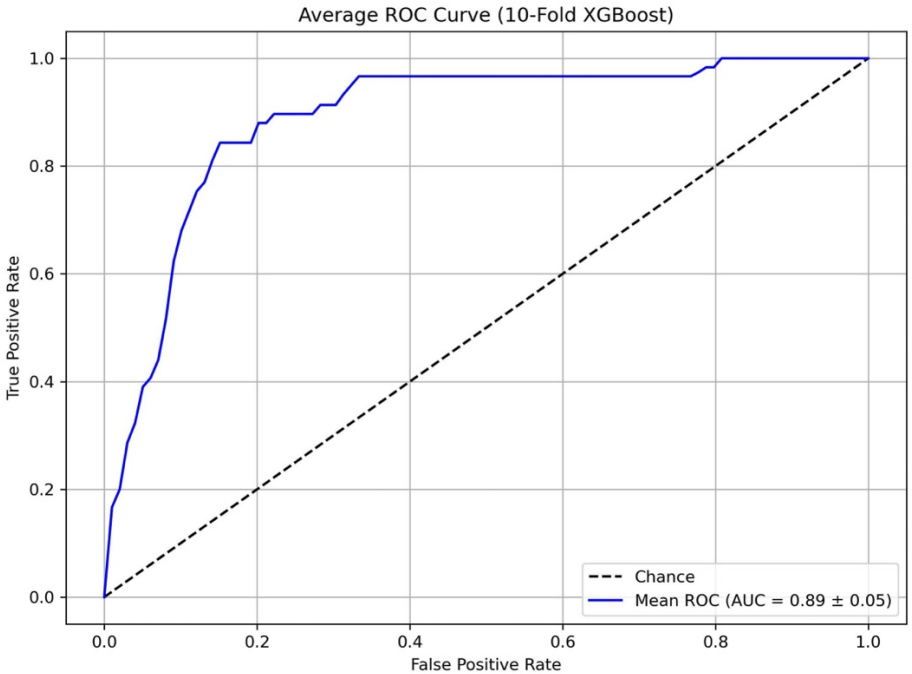


Figure 5.6 Average ROC curves for the XGBoost model. (A) shows the ROC curve following preliminary feature selection, where features were manually chosen based on minimal reliance on clinical expertise (demographics, basic medical history, lifestyle habits, and unattended measurable parameters). (B) illustrates the ROC curve after comprehensive feature selection, including optimisation steps and informed by the results from Phase 1

(A) Preliminary Feature Selection

(B) Comprehensive Feature Selection

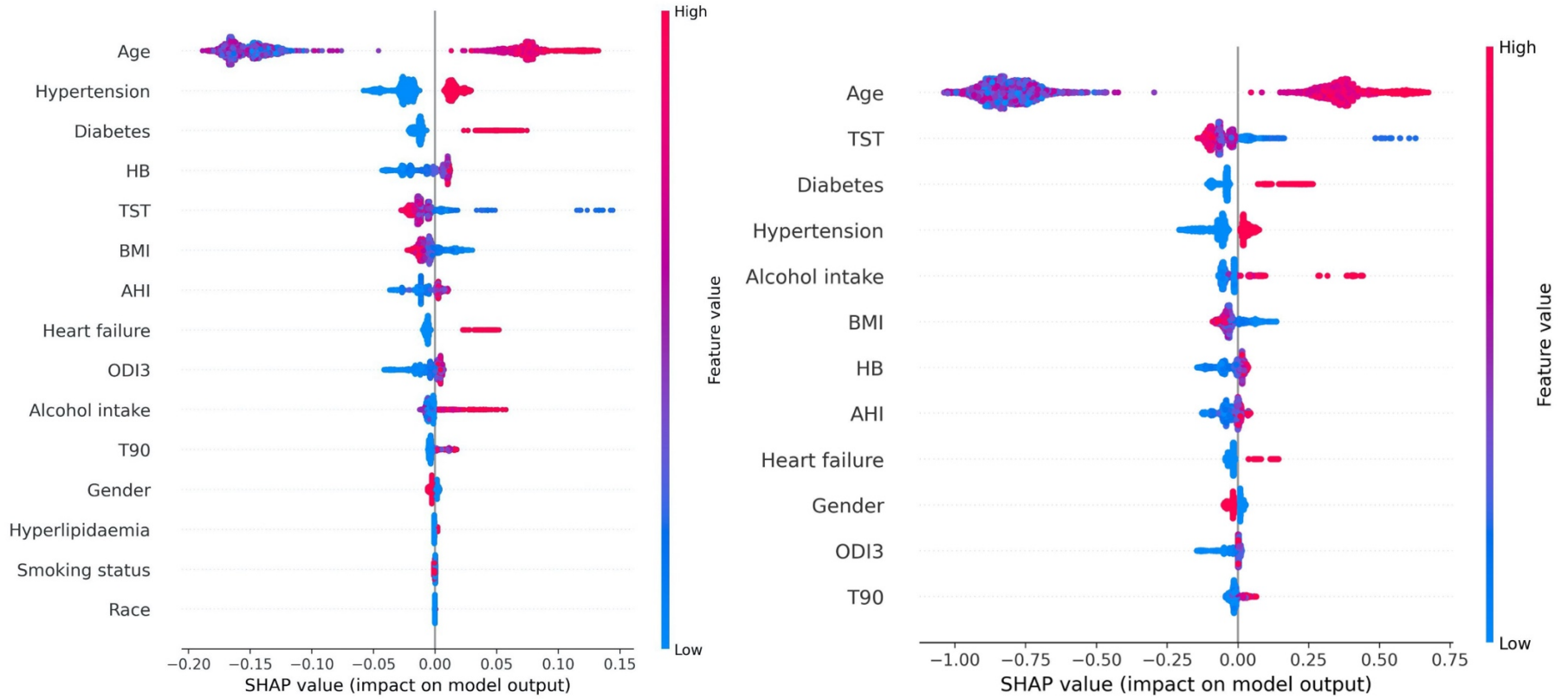


Figure 5.7 SHAP analysis illustrating individual feature contributions to the XGBoost model's predictions. (A) presents the SHAP analysis following preliminary feature selection, wherein features were manually selected to minimise reliance on clinical expertise (including demographics, basic medical history, lifestyle habits, and unattended measurable parameters). (B) shows the SHAP analysis after comprehensive feature selection, including optimisation procedures informed by Phase 1 results.

5.4.3 Extension of Phase 2: Application of the best-performing explainable model to predict 5-year and 10-year CVD mortality outcomes

The best-performing explainable model for predicting 3-year CVD mortality was proposed in Phase 2. The same model was also applied to predict 5-year and 10-year CVD mortality to test its generalisability. As shown in **Table 5.9**, XGBoost with the selected features (age, gender, BMI, alcohol intake, hypertension, diabetes, heart failure, HB, T90, TST, AHI, and ODI3) performed consistently well in predicting 5-year and 10-year CVD mortality. The training and testing F1 scores were 83.51% and 83.03% for 5-year prediction, and 84.36% and 81.17% for 10-year prediction, respectively. The ROC curves also rose steadily, with mean AUC values of 0.87 ± 0.05 for 5-year prediction and 0.88 ± 0.03 for 10-year prediction. The cumulative testing confusion matrix for the 5- and 10-year prediction horizons are shown in **Table 5.10** and **Table 5.11**.

The contributions of individual features to model performance were also assessed using SHAP analysis. Age was the predominant predictor for both the 5-year and 10-year predictions. Hypertension and diabetes were two key medical history variables that contributed to prediction across both time horizons. Among the oximetry-derived parameters, T90 was more informative for categorising 5-year CVD mortality, while HB contributed more to the 10-year prediction. By contrast, ODI3 ranked lowest for both predictions. TST, as a well-performing PSG-derived predictor, also demonstrated a meaningful contribution to both predictions.

Table 5.9 Performance of proposed explainable machine learning model and feature selections (XGBoost with comprehensive feature selection) for predicting 5-year and 10-year CVD mortality.

Model	**Training***				**Testing***			
	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE	Sensitivity SE	Specificity SE	Accuracy SE	F1 Score SE
5-year	78.63	89.08	78.91	83.51	78.55	88.48	78.82	83.03
	0.39	0.50	0.38	0.22	0.75	2.46	0.73	1.16
10-year	82.26	86.51	82.58	84.36	81.90	80.74	81.81	81.17
	0.15	0.33	0.13	0.15	0.60	2.04	0.54	0.96

* Values are expressed as percentages (%).

Table 5.10 The cumulative confusion matrix of proposed explainable machine learning model and feature selections (XGBoost with comprehensive feature selection) predicting 5-year CVD mortality.

		Predicted class label*	
		Positive	Negative
Actual label*	Positive	89	24
	Negative	494	3792

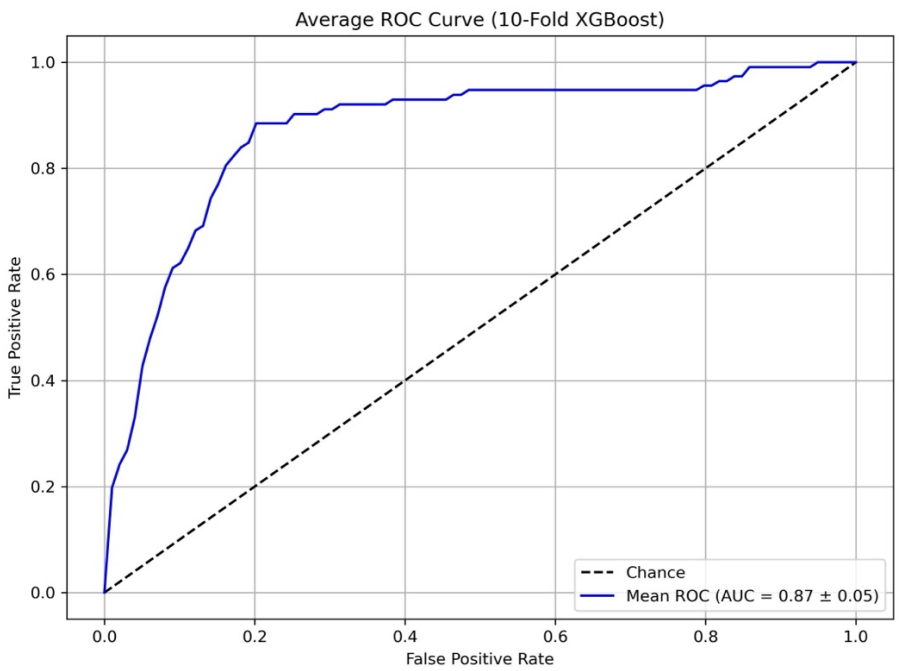
* The positive class represents participants who experienced 5-year CVD death, whereas the negative class represents individuals who survived beyond 5 years. For the 5-year analysis, a total of 4,399 participants were included, of whom 113 experienced CVD death.

Table 5.11 The cumulative confusion matrix of proposed explainable machine learning model and feature selections (XGBoost with comprehensive feature selection) predicting 10-year CVD mortality.

		Predicted class label*	
		Positive	Negative
Actual label*	Positive	232	51
	Negative	688	2883

* The positive class represents participants who experienced 10-year CVD death, whereas the negative class represents individuals who survived beyond 10 years. For the 10-year analysis, a total of 3854 participants were included, of whom 283 experienced CVD death.

(A) 5-year CVD mortality



(B) 10-year CVD mortality

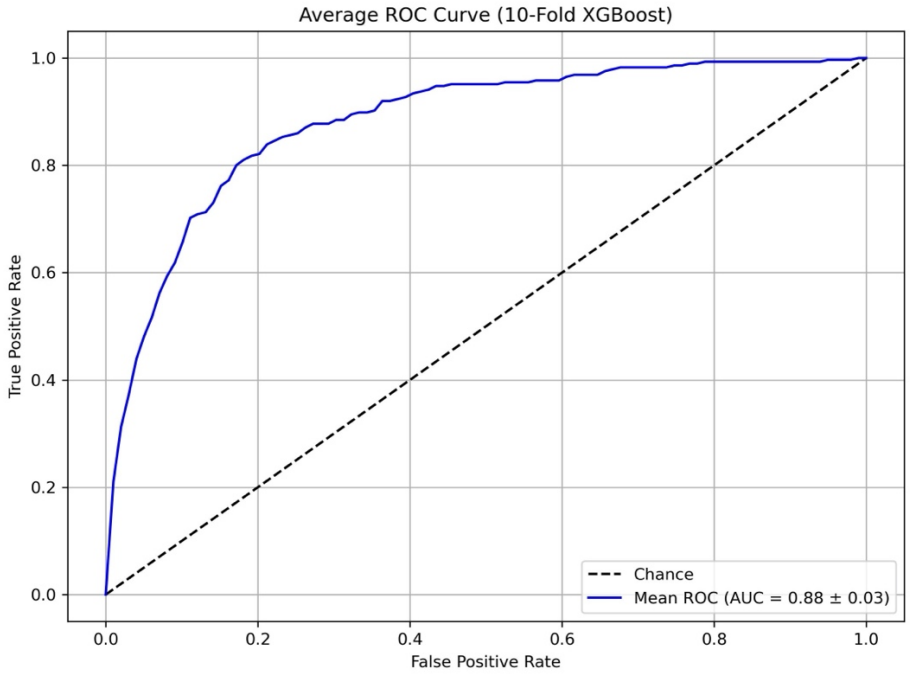


Figure 5.8 Average ROC curves for the proposed XGBoost model with comprehensively selected feature combinations predicting CVD mortality outcomes. (A) shows the ROC curve of model predicting 5-year CVD mortality. (B) shows the ROC curve of model predicting 10-year CVD mortality.

(A) 5-year CVD mortality

(B) 10-year CVD mortality

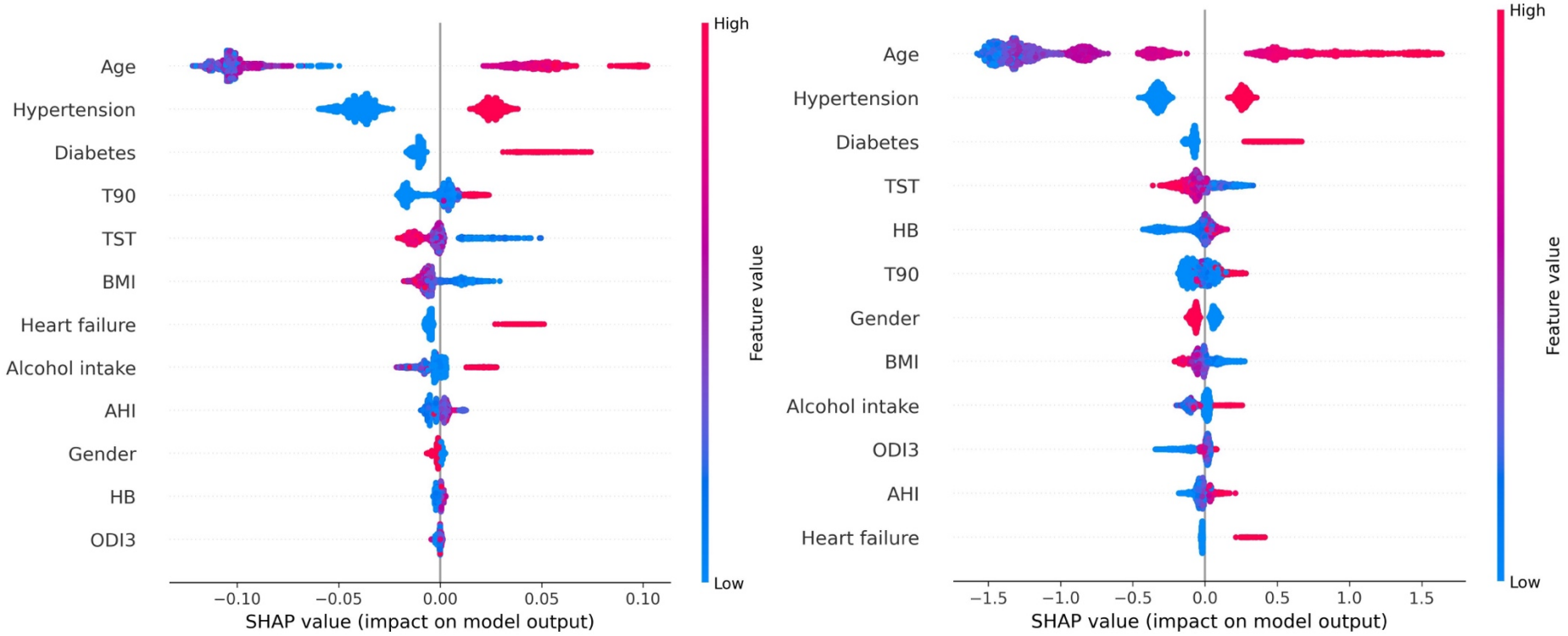


Figure 5.9 SHAP analysis illustrating individual feature contributions to the proposed XGBoost model's predictions. (A) presents the SHAP analysis of model predicting 5-year CVD mortality outcome. (B) shows the SHAP analysis of model predicting 10-year CVD mortality outcome.

5.5 Discussion

5.5.1 Phase 1: Can oximetry-derived parameters effectively predict CVD outcomes?

Phase 1 evaluated the combined predictive ability of oximetry-derived parameters in predicting 3-year CVD mortality outcomes. As shown in **Table 5.4**, the results highlight the extent to which these parameters enhance model performance when integrated with demographical and lifestyle predictors.

The baseline model (Model A), which included only demographic and lifestyle features, achieved reasonably good performance. In contrast, Model B, comprising only oximetry-derived parameters, demonstrated suboptimal predictive capability, suggesting that oximetry features alone may be insufficient for accurate CVD mortality prediction. However, when the features from Models A and B were combined in Model C, performance improved with a statistically significant 2% increase in the F1 score. This outcome suggests that while oximetry-derived parameters may not be strong standalone predictors, their additive value effectively enhances prediction when combined with demographic and lifestyle variables. In other words, these parameters contribute meaningful supplementary information, supporting the statement that medical outcome prediction benefits from a multi-dimensional, multi-variable approach. The more diverse the input features, the better the model performance.

To further explore the contributions of individual predictors underlying this performance boost, **Figure 5.5** presents the feature contributions based on Fisher scores. Age emerged as the most influential predictor, which is physiologically intuitive given the experiment is predicting the life expectancy. Among oximetry-derived features, T90 and HB were the top two contributors to the model. Although their Fisher scores were markedly lower than that of age, their relative importance supports their discriminative ability, aligning with prior research findings [6, 12]. In contrast, features such as smoking status, race, and ODI3 yielded near-zero Fisher scores, suggesting limited individual predictive value.

However, the univariate Fisher score used in this phase has inherent limitations. It evaluates features in isolation and does not capture potential interaction effects. In predictive modelling, it is not uncommon for a feature to appear irrelevant when considered alone, yet become highly

informative in conjunction with other variables [301]. For instance, although ODI3 ranked low individually, its combination with other oximetry-derived features contributed positively to model performance. This underscores the importance of recognising that univariate methods, such as the Fisher score, may overlook synergistic feature interactions. As such, Fisher scores should be regarded as a preliminary reference, not a definitive criterion for feature selection in Phase 2.

In summary, this phase confirmed the additive value of oximetry-derived parameters in enhancing the prediction of 3-year CVD mortality and underscored the importance of incorporating features that represent multiple dimensions of patient health. Although the improvement was modest under the simplest setting (approximately a 2% gain using LDA with oximetry features alone), the results provided preliminary evidence that multivariable integration can yield measurable performance benefits. The findings therefore advocate for a comprehensive modelling approach that combines demographic information, lifestyle habits, physiological signals, and medical history, and directly informed the feature selection strategy in Phase 2, where predictive performance improved further with the inclusion of additional parameters prior to model optimisation. While the univariate Fisher scores offered valuable insights consistent with both model performance and physiological relevance, they should be interpreted with caution due to their inability to capture feature interactions.

5.5.2 Phase 2: Explainable machine learning model for predicting 3-year CVD mortality outcome

While Phase 1 extended previous studies by evaluating the combined predictive value of oximetry-derived parameters, Phase 2 aimed to deliver personalised CVD outcome predictions suitable for public screening. This phase was conducted in two stages: a preliminary analysis to support feature selection and classifier optimisation, followed by a comprehensive analysis to propose a robust and explainable machine learning model for individualised CVD mortality prediction.

Using the selected feature set (age, BMI, gender, TST, T90, ODI3, HB, diabetes, hypertension, heart failure, and alcohol intake), XGBoost demonstrated superior predictive performance for predicting 3-year CVD mortality. It consistently surpassed alternative models across all evaluation metrics and maintained a balance between sensitivity and specificity. The model achieved an average AUC of 0.89 (**Figure 5.6**B) in the testing dataset, reflecting strong

capability to accurately stratify patients into 3-year CVD death versus 3-year CVD survivor groups, thus fulfilling the personalised prediction goal.

Although XGBoost is inherently less transparent than simpler models (e.g., LDA), its interpretability was significantly enhanced through SHAP, which quantified each feature's contribution to individual predictions. Among all predictors, age emerged as the most dominant feature, contributing substantially to both positive and negative outcome classes. The model using age as the sole input feature achieved an F1 score of 67.95% and an average AUC of 0.80 ± 0.07, indicating that age alone exhibits strong predictive ability, consistent with the findings from the SHAP analysis. Beyond this, **Figure 5.7**B reveals class-specific patterns in feature relevance: ODI3 and HB were more influential in predicting negative outcomes, whereas alcohol intake and heart failure were key drivers of positive predictions. Gender, despite its limited standalone predictive power, improved overall model performance through interactions with other variables, aligning with established epidemiological links between male sex and elevated CVD risk. While visualising individual tree splits remains complex, the SHAP outputs offer a practical interpretability layer by identifying modifiable risk factors. For instance, if a patient is flagged as high risk of 3-year CVD death, SHAP analysis may highlight high alcohol intake as a major contributor, providing clinicians with a clear, actionable target for intervention.

The feature contributions (**Figure 5.7**) observed in this study generally align with findings from previous studies. Established cardiovascular risk factors, including age, diabetes, and hypertension, demonstrated substantial contributions to the prediction of CVD mortality. However, notable differences were observed for specific sleep-associated measures. For instance, TST and HB, which demonstrated strong predictive utility in traditional Cox regression models, continued to serve as key contributors in the XGBoost model [12, 203]. In contrast, ODI3 and T90, which previously exhibited strong performance, particularly ODI3, showed reduced contributions in the current XGBoost-based analysis. Conversely, AHI, which was previously considered a suboptimal predictor due to its limited ability to reflect the full extent of hypoxic burden, contributed meaningfully in the current XGBoost model [6, 11, 27]. An unexpected pattern was observed for BMI. Higher BMI is generally considered a well-established risk factor for cardiovascular disease through recognised physiological pathways; however, the SHAP analysis indicated that lower BMI values contributed to the prediction of CVD mortality in this model. This finding appears contradictory to conventional clinical

expectations. One possible explanation is the presence of interaction effects among predictors, whereby BMI may influence risk differently depending on other covariates included in the model. In addition, this pattern may reflect residual confounding or non-linear associations.

These discrepancies may reflect fundamental differences in how predictor relationships are typically modelled. Although regression-based survival approaches such as Cox proportional hazards regression are multivariable in nature, OSA–CVD research has often emphasised the independent contribution of individual predictors, with limited systematic investigation of complex interactions among variables. While Cox models can accommodate interaction terms and non-linear transformations, these effects must be explicitly specified a priori, which may constrain the exploration of higher-order relationships in practice. In contrast, machine learning methods such as XGBoost can more flexibly capture non-linear associations and feature interdependencies without requiring manual definition of interaction structures. Consequently, variables that appear weak when considered individually may still contribute meaningfully through synergistic interactions, whereas predictors with strong marginal associations may show reduced importance once collinearity and shared information are accounted for in a multivariable setting. Therefore, analyses focused primarily on isolated predictor effects are valuable for preliminary interpretation but should not be viewed as definitive for feature selection when the goal is personalised prediction using multivariable machine learning frameworks.

The ROC curve of the XGBoost model (**Figure 5.6**) illustrates the trade-off between true positive and false positive rates across varying decision thresholds. At the default decision threshold of 0.5, the model achieved a sensitivity of 91.87% (**Table 5.6**). However, this threshold can be adjusted to better align with clinical priorities. In the context of this study, where the primary objective is to identify individuals at high risk of dying from CVD within three years, a high sensitivity is critical to minimise missed cases. To achieve a sensitivity exceeding 99%, as might be required in public health screening settings, the decision threshold would need to be lowered below 0.05, as indicated by the ROC curve. Nonetheless, increasing sensitivity often comes at the expense of specificity and overall accuracy. Therefore, any adjustment to the decision threshold must carefully balance clinical priorities with the risk of over-prediction, ensuring that the model remains both sensitive and practically useful for broad application.

The model proposed in this experiment successfully identifies a balance between predictive performance and the complexity of required input features. A core principle guiding this work was to minimise reliance on clinical resources, thereby ensuring that the model remains applicable in medically underserved regions and among individuals with limited access to healthcare due to financial constraints. Compared to existing studies that employed less comprehensive feature sets for CVD prediction, the proposed model achieved an 8% higher AUC [304]. This improvement is largely attributable to the inclusion of sleep-related measurements, specifically oximetry-derived indicators, which underscore the role of nocturnal hypoxia as an important predictor of CVD outcomes. While some other studies reported even greater predictive performance, with AUC improvements of up to 9%, their models typically relied on an extensive range of clinically obtained variables, such as fasting blood glucose, lipid profiles, and other laboratory-based assessments [305]. In contrast, this study adopted a minimalist approach: aside from PSG-derived features (which can be captured through portable devices such as smartwatches), all other inputs were simplified into binary responses (e.g., yes/no for medical history), making the model far more amenable to self-reported data collection. Although this strategy may compromise peak predictive performance, it significantly enhances accessibility and scalability. The resulting model represents a practical trade-off, slightly reduced performance in exchange for the potential to reach a broader population, including those without access to specialised medical assessments.

5.5.3 Extension of Phase 2: Application of the best-performing explainable model to predict 5-year and 10-year CVD mortality outcomes

The proposed explainable machine learning model, requiring minimal clinical resources, demonstrated strong predictive ability for 3-year CVD mortality. The next step was to evaluate its generalisability to broader time horizons, testing whether the model could maintain high performance in predicting CVD mortality outcomes across different timeframes and thereby support multiple clinical applications. Accordingly, the extension of Phase 2 applied the proposed best-performing explainable model to predict 5-year and 10-year CVD mortality outcomes. For the Phase 2 extension, XGBoost with the selected features (age, gender, BMI, alcohol intake, hypertension, diabetes, heart failure, HB, T90, TST, AHI, and ODI3) was applied to predict 5-year and 10-year CVD mortality. The mean AUCs were 0.87 for the 5-year prediction and 0.88 for the 10-year prediction (**Figure 5.8**). Compared with the mean AUC of 0.89 (**Figure 5.6**B) for the 3-year prediction in Phase 2, the proposed model demonstrated consistent and robust performance for CVD mortality prediction across different time horizons.

In previous studies, proposed machine learning approaches for predicting CVD outcomes have shown a key limitation: these classification approaches targeted a specific selected time horizon, without addressing how varying time horizons affect relative predictive performance [306]. Researchers have hypothesised that the reported strong predictive ability of many machine learning approaches may, in fact, be highly sensitive to the choice of time horizon. Theoretically, shorter time horizons increase the relevance of predictors to outcomes and may therefore result in better model performance. By contrast, when models are applied to longer durations, predictive performance is not guaranteed [306]. This study addressed this limitation by testing the generalisation ability of the proposed model and demonstrated that the explainable model maintained strong predictive performance across different time horizons. In addition, the results aligned with the previous hypothesis that shorter time horizons lead to better model performance. The testing F1 scores for 3-year (86.20%), 5-year (83.03%), and 10-year (81.18%) CVD prediction showed that, with the same combination of classifier and features, predictive performance declined as the prediction horizon increased (3-year: **Table 5.6** and 5- & 10- year: **Table 5.9**).

Apart from model performance, the explainability of the 5-year and 10-year CVD mortality predictions was also assessed using SHAP analysis (**Figure 5.9**). The results showed that, consistent with the 3-year prediction (**Figure 5.7**B), age was the predominant predictor for both time horizons. Notably, although the same model was applied across all three predictions, the SHAP values for individual features ranked differently. For example, HB contributed moderately to the 3-year and 10-year predictions but showed less contribution to the 5-year prediction. This reflects the fact that SHAP analysis is influenced not only by the model itself but also by the underlying data distribution [283]. As the time horizon varies, so does the class distribution. Therefore, it is expected that the order of feature contributions may not remain the same across different prediction horizons. Nevertheless, since SHAP values reflect only the individual contribution of each feature and do not account for feature interactions, it remains worthwhile to retain features that perform less optimally in a particular prediction [283].

In summary, Phase 2 proposed an explainable machine learning approach with minimal clinical reliance for predicting CVD mortality. Phase 2, together with its extension, evaluated the proposed model's ability across three different time horizons, demonstrating that this model, unlike previously proposed approaches, not only achieved strong predictive performance but also adapted flexibly to variations in time frames. Given its minimal reliance on clinical

resources and flexibility across time horizons, the model is well suited for large-scale population screening and on-site adjustments, making it applicable to multiple clinical purposes.

5.6 Limitations

Despite the promising findings, several limitations are present across phases of this experiment. The SHHS dataset was selected to maintain consistency with previous research; however, as discussed in Chapter 4, it presents notable limitations. These include potential participant selection bias and variability in data quality. Importantly, because the dataset is population-based sample, there is a very high proportion of CVD mortality survivors, leading to a pronounced class imbalance. While class weighting was applied to address this issue, the small proportion of CVD deaths may still lead to underrepresentation of the minority class. This underrepresentation can affect prediction accuracy when the model is applied more broadly in clinical settings, as some examples of positive cases may remain unseen during model training and therefore may not be correctly identified during prediction. As such, future studies may benefit from employing alternative datasets with more balanced class distributions and a greater representation of CVD mortality cases to improve model generalisability and applicability.

This experiment aims not only to develop an explainable machine learning framework for predicting CVD mortality, but also to reduce reliance on laboratory-based assessment and specialised clinical expertise by favouring simpler and more scalable feature inputs. Many of the selected PSG-derived parameters support this objective, as oximetry-based parameters can, in principle, be derived from portable monitoring devices in home settings. However, AHI and TST represent important exceptions. Although respiratory event indices can be approximated using home sleep apnoea testing devices, these systems typically estimate the respiratory event index based on recording time rather than true sleep time, and their accuracy relative to laboratory PSG-derived AHI may be reduced, particularly in mild or complex cases [74]. Accordingly, PSG-derived AHI was used in this experiment to ensure accurate event quantification, while acknowledging that model performance may differ when using portable estimates. Similarly, TST in this study was obtained from laboratory EEG-based sleep staging. While portable EEG systems with fewer electrodes may provide approximate sleep–wake estimation, their accuracy remains lower than full PSG, and the impact of substituting estimated TST on predictive performance requires further investigation.

The feature selection strategy in Phase 2 also introduces limitations. In the preliminary stage, features were manually selected based on their minimal reliance on clinical resources to promote broader applicability. However, this manual selection process may have introduced subjective bias, as the perceived degree of "clinical reliance" can vary across researchers and healthcare settings. Some features considered accessible in one context may be difficult to obtain in medically underserved regions or may be misreported due to a lack of public awareness. Thus, future studies adopting this principle should incorporate expert consultation and potentially region-specific evaluations to guide feature inclusion more rigorously.

From a modelling perspective, the choice of classifier represents a trade-off between accuracy and explainability. While the XGBoost model demonstrated the highest performance and acceptable interpretability through SHAP analysis, it requires greater computational resources than simpler models such as LDA. Specifically, XGBoost needs more memory to store the ensemble of trees during deployment. Its dependence on hyperparameter optimisation also adds implementation complexity at the training stage. For example, when the outcome time horizon changes, the hyperparameters must be re-tuned to maintain optimal performance. In contrast, LDA does not require such additional hyperparameter searches, even when time horizons are adjusted. For scenarios requiring low-latency or resource-constrained deployment (portable devices), simpler models may offer more practical alternatives. Although cloud-based services can reduce some computational barriers and enable the use of more complex models, they do not fully eliminate practical constraints. Large-scale data processing remains costly and energy-intensive, and reliance on cloud infrastructure may be unsuitable in settings requiring immediate response, enhanced privacy, or offline functionality, such as wearable or bedside monitoring devices. Therefore, lightweight and efficient predictive algorithms may still be essential for real-world translation, even when input feature dimensionality is modest. Indeed, preliminary results in Phase 2 showed that LDA and linear SVM models, while not outperforming XGBoost, achieved competitive performance, highlighting their potential for future use in resource-limited or portable applications.

Moreover, the evaluation of feature selection and classifier optimisation in this study carries an inherent risk of optimistic bias. Although 10 folds cross validation was employed to mitigate the possible overfitting, both the feature selection phase and the subsequent classifier training and testing were conducted within the same dataset, without an entirely independent external test cohort. This design choice was largely driven by the highly imbalanced nature of the SHHS

dataset, in which positive cases account for only approximately 1.2% of the total population. Under such conditions, a conventional hold-out split would yield very few outcome events in the test set, limiting the stability and interpretability of performance estimates. Consequently, the present framework prioritised the use of a consistent dataset across experimental stages to enable like-for-like comparison.

Beyond challenges in experimental design, the methodological shift from traditional time-to-event survival analysis to machine learning based prediction also raises important considerations. Although conventional survival models typically output relative outcomes such as hazard ratios, which may be difficult for non-statistical patients to interpret, they remain highly valuable for modelling time-to-event outcomes. In particular, survival analysis provides a dynamic framework in which risk can be characterised over time and updated as risk factors change. Clinically, such models support the visualisation of risk trajectories and can assist clinicians in evaluating the effectiveness of risk stratification strategies or therapeutic interventions for CVD outcomes. Therefore, rather than viewing survival analysis and machine learning approaches as competing methodologies, they should be regarded as complementary. Integrating both frameworks may support more comprehensive monitoring of risk development while also facilitating clearer communication of disease severity and prognosis at the individual level.

Machine learning models, compared with traditional survival approaches, are often more flexible and may therefore carry a greater risk of overfitting if not carefully regularised and rigorously validated. Evaluating predictive performance across multiple time horizons, as undertaken in this experiment, may provide additional evidence of model robustness and temporal generalisability. Nevertheless, consistent performance across horizons does not preclude overfitting, particularly when models are developed and evaluated within the same cohort. Consequently, rigorous validation remains essential to ensure reliable and generalisable predictive performance.

Chapter 6
Conclusion and future work

6 Conclusion and future work

The previous two chapters presented the methodological and applied developments of this thesis. Chapter 4 established a unified computational framework for comparing desaturation area–based algorithms, discussing the impacts of algorithmic discrepancies on predictive performance even under the same parameter definition, and identifying the most robust and best-performing method for CVD mortality prediction. Chapter 5 extended the analysis to individual-level prediction of CVD mortality using explainable machine learning. This chapter concludes the thesis by summarising the main findings, the limitations identified in each experiment, and the corresponding directions for future research.

6.1 Experiment 1: Comparison of Oxygen Desaturation Area–Based Methods in Predicting Cardiovascular Disease Mortality Outcomes

Experiment 1 conducted the first systematic comparison of three major desaturation area–based algorithms (HB, REDTA, and DesSev) within a unified computational framework using the SHHS database, along with additional methodological variations inspired by these algorithms. In total, fifteen methodological combinations were examined. The analysis demonstrated that variations in event definition, sampling window, and baseline calculation substantially influenced predictive ability for CVD mortality. The results clarified that methodological discrepancies were largely responsible for the inconsistent findings reported in previous research. This was evidenced by the three algorithms (HB, REDTA, and DesSev) resulting HRs that ranged from statistically significant to insignificant when applied to the same dataset for predicting CVD mortality.

Among all desaturation area–based methods evaluated in Chapter 4, A_{RRM} (the area method employing a record-specific sampling window and baseline based on manually scored respiratory events) achieved the highest HR of 1.79 (p-value = 0.04) after full covariate adjustment. Consequently, A_{RRM} was identified as the best performing algorithm for large-scale analysis. This study established a reproducible computational benchmark for future investigations of desaturation area–based parameters and provided valuable insights for the continued refinement and automation of oximetry-based algorithms.

However, there are limitations in this experiment. First, the SHHS cohort is a community-based sample of predominantly older Caucasian adults, with limited information on OSA treatment history and unknown disease duration. These factors may influence cardiovascular adaptation to long-term hypoxaemia and limit the generalisability of the findings to broader clinical populations. Second, comparisons between manually scored respiratory events and automated desaturation detection were conducted in only one database using a single algorithm. Validation across multiple datasets and alternative published detection methods is needed to draw more reliable conclusions. Finally, automated algorithm performance may be sensitive to the choice of desaturation threshold, as lower criteria (e.g., 2%) have been suggested to improve prediction in some contexts. However, such sensitivity analyses were not feasible because the ABOSA software is a closed implementation that does not allow modification of detection parameters.

Future studies should address the limitations of this experiment by validating desaturation area–based methods in larger and more demographically diverse cohorts, including broader age groups and ethnic backgrounds. In addition, clinical datasets consisting of OSA-only populations would be valuable to reduce confounding effects from mixed community-based samples and better isolate OSA-specific CVD risk. Further investigations should also evaluate multiple automated desaturation detection frameworks to determine whether the observed findings are consistent across different algorithmic approaches.

Moreover, the sensitivity of automated algorithms to desaturation thresholds should be systematically explored, for example by comparing 2%, 3%, and 4% criteria. Such analysis would require more flexible implementations than the current closed ABOSA software, allowing adjustment of event annotation and detection criteria. Beyond threshold effects, future work could extend model evaluation beyond HRs, 95% CI and p-values by incorporating model fit statistics such as AIC, enabling comparison of which desaturation area metrics best explain CVD outcomes. The role and significance of covariates within these multivariable models could also be examined in greater detail.

Beyond the scope of this thesis, it would be informative to investigate potential collinearity between desaturation area metrics and clinical covariates, for example through PCA-based dimension reduction or correlation-based feature analysis. Sensitivity analysis could also examine robustness within each baseline/sampling window category, rather than only

comparing broad frameworks. For example, parameters within the event-specific approach, such as window length (100 seconds vs 150 seconds) or baseline estimation rules, could be varied to assess how sensitive performance is to minor implementation differences. These investigations would provide deeper insight into the desaturation area computation.

Finally, future research could move beyond algorithmic comparisons to explore the physiological interpretation of oximetry-derived severity metrics. For instance, the same desaturation area may be calculated from prolonged mild desaturations or shorter but deeper events, yet these patterns may not carry equivalent CVD consequences. Disentangling the relative contributions of desaturation depth, duration, and event frequency may improve understanding of the mechanisms linking nocturnal hypoxaemia to adverse CVD outcomes.

6.2 Experiment 2: Using PSG-Derived Parameters and Explainable Machine Learning Approaches to Predict CVD Mortality

Experiment 2 aimed to translate the methodological findings of Experiment 1 into applied outcome prediction and to address current limitations wherein PSG-derived parameters are typically assessed individually, with evaluation methods restricted to relative hazard estimation rather than individual-level prediction. Phase 1 evaluated the predictive ability of established oximetry-derived parameters (ODI3, T90, and desaturation area–based measures) for 3-year CVD mortality. Results confirmed that while individual parameters demonstrated predictive value, their combination enhanced predictive performance for CVD mortality. Specifically, the model incorporating combined oximetry-derived parameters achieved a 2% improvement in performance under the LDA classifier, supporting the use of multivariable modelling strategies.

Building on these findings, Phase 2 developed an explainable machine learning framework based on XGBoost, integrating PSG-derived parameters with demographic, lifestyle, and medical information. Among all classifiers tested, XGBoost achieved the best overall predictive performance, yielding an AUC of 0.89 ± 0.05 and an F1 score of 86.20% for 3-year CVD mortality prediction. The optimal feature subset combined key PSG-derived measures (TST, HB, T90, and ODI3) with clinical covariates (age, BMI, gender, hypertension, diabetes, heart failure, and alcohol intake). The framework generalised effectively across extended time horizons, maintaining AUCs of 0.87 ± 0.05 and 0.88 ± 0.03 for 5-year and 10-year CVD

mortality predictions, respectively, with corresponding F1 scores of 83.03% and 81.17%. These results demonstrate strong temporal stability and robustness of the proposed model. Model interpretability analysis using SHAP identified age, hypertension, HB, and T90 as dominant contributors, reaffirming their physiological and clinical relevance.

Importantly, the XGBoost-based framework achieved performance comparable to that of more resource-intensive approaches while requiring substantially fewer specialised clinical inputs (showing up to 9% improvement compared to previous studies). Notably, it required only binary responses to medical history (e.g., yes/no) rather than detailed numerical inputs, thereby reducing medical reliance and enhancing scalability for population-level screening. This adaptability supports its potential for deployment across diverse healthcare settings, including resource-limited and community-based environments.

Despite the promising findings, several limitations should be acknowledged. First, the SHHS dataset is population-based and highly imbalanced, with very few CVD mortality cases, which may limit minority-class representation and reduce generalisability to clinical settings despite the use of class weighting. Second, although reducing medical reliance was a key goal, there is a trade-off between simplicity and accuracy. While many features were scalable oximetry-based measures, key predictors such as AHI and TST were derived from full laboratory PSG, which may limit direct transferability to portable or home-based settings. Third, the feature selection process involved manual judgement to prioritise clinically accessible variables, which may introduce subjective bias and may vary across healthcare contexts.

From a modelling perspective, the strongest-performing classifier, XGBoost, involves greater computational cost and tuning complexity compared with simpler linear models, which may be more suitable for resource-constrained deployment. In addition, although cross-validation was applied, feature selection and model optimisation were conducted within the same dataset without external validation, introducing a risk of optimistic bias. Finally, the shift from traditional survival modelling to machine learning prediction highlights a trade-off: while ML offers flexible individual-level absolute risk estimates, survival analysis remains valuable for time-to-event interpretation, and both approaches should be viewed as complementary rather than competing. Overall, future work should prioritise external validation, improved cohort diversity, and further evaluation of portability and robustness across settings.

To address these limitations, future work should prioritise validation of the proposed framework in independent cohorts with more balanced outcome distributions and broader demographic diversity to improve generalisability. Feature selection should be further refined through expert consultation and context-specific evaluation to enhance reproducibility across healthcare settings. In addition, future studies could explore lightweight predictive algorithms that retain interpretability while reducing computational demands, supporting potential translation to portable or home-based monitoring devices. Given the trade-off between traditional survival models such as Cox regression and machine learning–based prediction, future research should also consider integrating survival-focused machine learning approaches, such as tree-based survival models or discrete-time survival frameworks, which can provide absolute risk estimates over time and complement conventional time-to-event analysis.

Beyond model development, evaluation strategies could be extended to incorporate clinically meaningful, cost-sensitive decision thresholds. In practice, the consequences of misclassification are asymmetric: failing to identify a truly high-risk individual may carry substantially greater clinical cost than incorrectly flagging a low-risk patient for further assessment. Besides, predictive performance assessment in this experiment was primarily based on sensitivity, specificity, accuracy, and F1-score. Although PPV and negative predictive value are also informative in medical applications, the extremely low prevalence of CVD mortality in SHHS (~1.2%) limits their interpretability. In highly imbalanced settings, PPV, in particular, can remain low even when overall discrimination of model is strong, making comparisons between models less informative. Future studies using cohorts with higher event representation or more balanced outcome distributions should incorporate PPV and negative predictive value to better reflect clinical utility.

Finally, additional outcomes beyond CVD mortality may warrant investigation, as mortality events are rare and represent a distant endpoint for many individuals. Alternative clinically relevant outcomes, such as broader CVD events, hospitalisation, or transitions to aged-care support, may provide richer and more actionable opportunities for risk stratification. For many patients, understanding the likelihood of maintaining independent living versus requiring nursing-home care may be more immediately meaningful than long-term mortality prediction, thereby enhancing the practical impact of modelling.

6.3 Conclusion

This thesis advances both the methodological and applied understanding of how PSG-derived parameters, particularly oximetry-based parameters, can be used for predicting CVD mortality. It highlights major computational inconsistencies in the implementation of desaturation area–based parameters and identifies the best-performing and most robust model, with A_{RRM} achieving the highest statistically significant HR of 1.79. The findings also demonstrate the advantage of integrating multiple PSG-derived metrics for prediction and establish an interpretable machine learning framework capable of accurate, individual-level outcome prediction with minimal clinical reliance across multiple time horizons. The best-performing model, XGBoost, achieved an AUC of 0.89 ± 0.05 and an F1 score of 86.20% for 3-year prediction, with consistent performance at 5-year (AUC = 0.87) and 10-year (AUC = 0.88) horizons using PSG-derived parameters and clinical covariates.

Collectively, these contributions provide a foundation for scalable and clinically relevant CVD mortality prediction based on sleep data. They enable early identification of high-risk individuals and promote the broader integration of sleep assessment into cardiovascular risk stratification. Together, these advances represent a step toward clinically deployable, data-driven tools for cardiovascular risk prediction, enhancing the translational value of sleep studies in preventive cardiology.

7 References

[1] "Heart, stroke and vascular disease: Australian facts." Australian Institute of Health and Welfare https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/hsvd-facts/contents/summary (accessed Oct. 23rd, 2025).

[2] "Cardiovascular diseases (CVDs)." WHO. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Oct. 23rd, 2025).

[3] A. V. Benjafield *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," (in eng), *Lancet Respir Med,* vol. 7, no. 8, pp. 687-698, Aug 2019, doi: 10.1016/s2213-2600(19)30198-5.

[4] G. Jean-Louis, F. Zizi, D. Brown, G. Ogedegbe, J. Borer, and S. McFarlane, "Obstructive sleep apnea and cardiovascular disease: evidence and underlying mechanisms," (in eng), *Minerva Pneumol,* vol. 48, no. 4, pp. 277-293, Dec 2009.

[5] R. Alvarez-Sala, F. García-Río, F. Del Campo, C. Zamarrón, and N. C. Netzer, "Sleep apnea and cardiovascular diseases," (in eng), *Pulm Med,* vol. 2014, p. 690273, 2014, doi: 10.1155/2014/690273.

[6] M. Baumert *et al.*, "Composition of nocturnal hypoxaemic burden and its prognostic value for cardiovascular mortality in older community-dwelling men," (in eng), *Eur Heart J,* vol. 41, no. 4, pp. 533-541, Jan 21 2020, doi: 10.1093/eurheartj/ehy838.

[7] K. Ouriel, "Peripheral arterial disease," (in eng), *Lancet,* vol. 358, no. 9289, pp. 1257-64, Oct 13 2001, doi: 10.1016/s0140-6736(01)06351-6.

[8] S. Luong, L. Lezama, and S. Khan, "Diagnosis and Management of Obstructive Sleep Apnea: Updates and Review," *Journal of Otorhinolaryngology, Hearing and Balance Medicine,* vol. 5, no. 2, p. 16, 2024. [Online]. Available:https://www.mdpi.com/2504-463X/5/2/16.

[9] R. Allen, "Home sleep studies," *Australian Prescriber,* vol. 35, pp. 62-64, 04/01 2012, doi: 10.18773/austprescr.2012.027.

[10] AASM, "Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. American Academy of Sleep Medicine Task Force," (in eng), *Sleep,* vol. 22, no. 5, pp. 667-89, Aug 1 1999.

[11] N. M. Punjabi, "COUNTERPOINT: Is the Apnea-Hypopnea Index the Best Way to Quantify the Severity of Sleep-Disordered Breathing? No," (in eng), *Chest,* vol. 149, no. 1, pp. 16-9, Jan 2016, doi: 10.1378/chest.14-2261.

[12] A. Azarbarzin *et al.*, "The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study," (in eng), *Eur Heart J,* vol. 40, no. 14, pp. 1149-1157, Apr 7 2019, doi: 10.1093/eurheartj/ehy624.

[13] P. I. Terrill, "A review of approaches for analysing obstructive sleep apnoea-related patterns in pulse oximetry data," (in eng), *Respirology,* vol. 25, no. 5, pp. 475-485, May 2020, doi: 10.1111/resp.13635.

[14] L. Wang *et al.*, "Independent Association Between Oxygen Desaturation Index and Cardiovascular Disease in Non-Sleepy Sleep-Disordered Breathing Subtype: A Chinese Community-Based Study," (in eng), *Nat Sci Sleep,* vol. 14, pp. 1397-1406, 2022, doi: 10.2147/nss.S370471.

[15] S. He, P. A. Cistulli, and P. de Chazal, "A Review of Novel Oximetry Parameters for the Prediction of Cardiovascular Disease in Obstructive Sleep Apnoea," (in eng), *Diagnostics (Basel),* vol. 13, no. 21, Oct 26 2023, doi: 10.3390/diagnostics13213323.

[16] R. B. Berry *et al.*, "Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine," (in eng), *J Clin Sleep Med,* vol. 8, no. 5, pp. 597-619, Oct 15 2012, doi: 10.5664/jcsm.2172.

[17] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. Quan, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications," *Westchester, IL: American Academy of Sleep Medicine,* 01/01 2007.

[18] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," (in eng), *Sci Rep,* vol. 11, no. 1, p. 6968, Mar 26 2021, doi: 10.1038/s41598-021-86327-7.

[19] A. Abidov and O. Chehab, "Cardiovascular risk assessment models: Have we found the perfect solution yet?," *Journal of Nuclear Cardiology,* vol. 27, no. 6, pp. 2375-2385, 2020/12/01/ 2020, doi: 10.1007/s12350-019-01642-x.

[20] A. Sashegyi and D. Ferry, "On the Interpretation of the Hazard Ratio and Communication of Survival Benefit," (in eng), *Oncologist,* vol. 22, no. 4, pp. 484-486, Apr 2017, doi: 10.1634/theoncologist.2016-0198.

[21] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," (in eng), *BMC Med Res Methodol,* vol. 18, no. 1, p. 24, Feb 26 2018, doi: 10.1186/s12874-018-0482-1.

[22] I. K. Omurlu, M. Ture, and F. Tokatli, "The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer," *Expert Systems with Applications,* vol. 36, no. 4, pp. 8582-8588, 2009.

[23] F. Datema, A. Moya, P. Krause, and T. Bäck, "Random survival forests versus cox proportional hazards regression. Surv. Predict," *Head Neck Cancer Impact Tumor Patient Spec. Charact,* vol. 94, p. 20, 2012.

[24] C. Nicolò *et al.*, "Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer," (in eng), *JCO Clin Cancer Inform,* vol. 4, pp. 259-274, Mar 2020, doi: 10.1200/cci.19.00133.

[25] S. Pölsterl, N. Navab, and A. Katouzian, "Fast Training of Support Vector Machines for Survival Analysis," in *Machine Learning and Knowledge Discovery in Databases*, Cham, A. Appice, P. P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, and C. Soares, Eds., 2015// 2015: Springer International Publishing, pp. 243-259.

[26] H. Ishwaran, U. Kogalur, E. Blackstone, and M. Lauer, "Random Survival Forests," *The Annals of Applied Statistics,* vol. 2, 12/11 2008, doi: 10.1214/08-AOAS169.

[27] M. P. Butler *et al.*, "Apnea-Hypopnea Event Duration Predicts Mortality in Men and Women in the Sleep Heart Health Study," (in eng), *Am J Respir Crit Care Med,* vol. 199, no. 7, pp. 903-912, Apr 1 2019, doi: 10.1164/rccm.201804-0758OC.

[28] P. d. Chazal *et al.*, "Predicting Cardiovascular Outcomes Using the Respiratory Event Desaturation Transient Area Derived from Overnight Sleep Studies," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* pp. 5496-5499, 2021.

[29] A. Kulkas, P. Tiihonen, K. Eskola, P. Julkunen, E. Mervaala, and J. Töyräs, "Novel parameters for evaluating severity of sleep disordered breathing and for supporting diagnosis of sleep apnea-hypopnea syndrome," (in eng), *J Med Eng Technol,* vol. 37, no. 2, pp. 135-43, Feb 2013, doi: 10.3109/03091902.2012.754509.

[30] E. Olvera Lopez, B. D. Ballard, and A. Jan, "Cardiovascular Disease," in *StatPearls*. Treasure Island (FL): StatPearls Publishing Copyright © 2025, StatPearls Publishing LLC., 2025.

[31] P. Libby and P. Theroux, "Pathophysiology of Coronary Artery Disease," *Circulation,* vol. 111, no. 25, pp. 3481-3488, 2005/06/28 2005, doi: 10.1161/CIRCULATIONAHA.105.537878.
[32] P. Libby, "Inflammation in atherosclerosis," (in eng), *Nature,* vol. 420, no. 6917, pp. 868-74, Dec 19-26 2002, doi: 10.1038/nature01323.
[33] B. T. William, "Cerebrovascular Disease," *Veterinary Clinics of North America: Small Animal Practice,* vol. 26, no. 4, pp. 925-943, 1996/07/01/ 1996, doi: 10.1016/S0195-5616(96)50112-9.
[34] A. S. Khaku and P. Tadi, "Cerebrovascular Disease," in *StatPearls*. Treasure Island (FL): StatPearls Publishing Copyright © 2025, StatPearls Publishing LLC., 2025.
[35] R. V. Krishnamurthi *et al.*, "Stroke Prevalence, Mortality and Disability-Adjusted Life Years in Adults Aged 20-64 Years in 1990-2013: Data from the Global Burden of Disease 2013 Study," (in eng), *Neuroepidemiology,* vol. 45, no. 3, pp. 190-202, 2015, doi: 10.1159/000441098.
[36] C. Cimminiello, "PAD. Epidemiology and pathophysiology," (in eng), *Thromb Res,* vol. 106, no. 6, pp. V295-301, Jun 1 2002, doi: 10.1016/s0049-3848(01)00400-5.
[37] P. Song *et al.*, "The Global and Regional Prevalence of Abdominal Aortic Aneurysms: A Systematic Review and Modeling Analysis," (in eng), *Ann Surg,* vol. 277, no. 6, pp. 912-919, Jun 1 2023, doi: 10.1097/sla.0000000000005716.
[38] N. Sakalihasan, R. Limet, and O. D. Defawe, "Abdominal aortic aneurysm," (in English), *The Lancet,* vol. 365, no. 9470, pp. 1577-89, Apr 30-May 6, 2005, doi:10.1016/S0140-6736(05)66459-8.
[39] B. G. Bruneau, "The developmental genetics of congenital heart disease," (in eng), *Nature,* vol. 451, no. 7181, pp. 943-8, Feb 21 2008, doi: 10.1038/nature06801.
[40] C. Antzelevitch and A. Burashnikov, "Overview of Basic Mechanisms of Cardiac Arrhythmia," (in eng), *Card Electrophysiol Clin,* vol. 3, no. 1, pp. 23-45, Mar 1 2011, doi: 10.1016/j.ccep.2010.10.012.
[41] G. Y. H. Lip *et al.*, "Hypertension and cardiac arrhythmias: executive summary of a consensus document from the European Heart Rhythm Association (EHRA) and ESC Council on Hypertension, endorsed by the Heart Rhythm Society (HRS), Asia-Pacific Heart Rhythm Society (APHRS), and Sociedad Latinoamericana de Estimulación Cardíaca y Electrofisiología (SOLEACE)," (in eng), *Eur Heart J Cardiovasc Pharmacother,* vol. 3, no. 4, pp. 235-250, Oct 1 2017, doi: 10.1093/ehjcvp/pvx019.
[42] F. Khan, T. Tritschler, S. R. Kahn, and M. A. Rodger, "Venous thromboembolism," (in eng), *Lancet,* vol. 398, no. 10294, pp. 64-77, Jul 3 2021, doi: 10.1016/s0140-6736(20)32658-1.
[43] A. Khoja *et al.*, "Modifiable and Non-Modifiable Risk Factors for Premature Coronary Heart Disease (PCHD): Systematic Review and Meta-Analysis," (in eng), *Heart Lung Circ,* vol. 33, no. 3, pp. 265-280, Mar 2024, doi: 10.1016/j.hlc.2023.12.012.
[44] M. Vaduganathan, G. A. Mensah, J. V. Turco, V. Fuster, and G. A. Roth, "The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health," (in eng), *J Am Coll Cardiol,* vol. 80, no. 25, pp. 2361-2371, Dec 20 2022, doi: 10.1016/j.jacc.2022.11.005.
[45] R. Marinigh, G. Y. H. Lip, N. Fiotti, C. Giansante, and D. A. Lane, "Age as a Risk Factor for Stroke in Atrial Fibrillation Patients: Implications for Thromboprophylaxis," *Journal of the American College of Cardiology,* vol. 56, no. 11, pp. 827-837, 2010/09/07/ 2010, doi: 10.1016/j.jacc.2010.05.028.
[46] M. Garcia, S. L. Mulvagh, C. N. Merz, J. E. Buring, and J. E. Manson, "Cardiovascular Disease in Women: Clinical Perspectives," (in eng), *Circ Res,* vol. 118, no. 8, pp. 1273-93, Apr 15 2016, doi: 10.1161/circresaha.116.307547.

[47] T. Keteepe-Arachi and S. Sharma, "Cardiovascular Disease in Women: Understanding Symptoms and Risk Factors," (in eng), *Eur Cardiol,* vol. 12, no. 1, pp. 10-13, Aug 2017, doi: 10.15420/ecr.2016:32:1.

[48] N. Chaturvedi, "ETHNIC DIFFERENCES IN CARDIOVASCULAR DISEASE," *Heart,* vol. 89, no. 6, p. 681, 2003, doi: 10.1136/heart.89.6.681.

[49] M. Gupta, S. Brister, and S. Verma, "Is South Asian ethnicity an independent cardiovascular risk factor?," *Canadian Journal of Cardiology,* vol. 22, no. 3, pp. 193-197, 2006/03/01/ 2006, doi: 10.1016/S0828-282X(06)70895-9.

[50] M. R. Kolber and C. Scrimshaw, "Family history of cardiovascular disease," (in eng), *Can Fam Physician,* vol. 60, no. 11, p. 1016, Nov 2014.

[51] S. S. Khan *et al.*, "Association of Body Mass Index With Lifetime Risk of Cardiovascular Disease and Compression of Morbidity," *JAMA Cardiology,* vol. 3, no. 4, pp. 280-287, 2018, doi: 10.1001/jamacardio.2018.0022.

[52] P. Dikaiou *et al.*, "Obesity, overweight and risk for cardiovascular disease and mortality in young women," *European Journal of Preventive Cardiology,* vol. 28, no. 12, pp. 1351-1359, 2020, doi: 10.1177/2047487320908983.

[53] J. R. Sowers, M. Epstein, and E. D. Frohlich, "Diabetes, hypertension, and cardiovascular disease: an update," (in eng), *Hypertension,* vol. 37, no. 4, pp. 1053-9, Apr 2001, doi: 10.1161/01.hyp.37.4.1053.

[54] D. Mozaffarian, P. W. F. Wilson, and W. B. Kannel, "Beyond Established and Novel Risk Factors," *Circulation,* vol. 117, no. 23, pp. 3031-3038, 2008/06/10 2008, doi: 10.1161/CIRCULATIONAHA.107.738732.

[55] M.-C. Tsai, C.-C. Lee, S.-C. Liu, P.-J. Tseng, and K.-L. Chien, "Combined healthy lifestyle factors are more beneficial in reducing cardiovascular disease in younger adults: a meta-analysis of prospective cohort studies," *Scientific Reports,* vol. 10, p. 18165, 10/23 2020, doi: 10.1038/s41598-020-75314-z.

[56] E. Tasdighi *et al.*, "Association between cigarette smoking status, intensity, and cessation duration with long-term incidence of nine cardiovascular and mortality outcomes: The Cross-Cohort Collaboration (CCC)," *PLOS Medicine,* vol. 22, no. 11, p. e1004561, 2025, doi: 10.1371/journal.pmed.1004561.

[57] Y. Yang *et al.*, "Joint association of smoking and physical activity with mortality in elderly hypertensive patients: A Chinese population-based cohort study in 2007–2018," (in English), *Frontiers in Public Health,* Original Research vol. Volume 10 - 2022, 2022-September-29 2022, doi: 10.3389/fpubh.2022.1005260.

[58] D. Linz *et al.*, "The importance of sleep-disordered breathing in cardiovascular disease," (in eng), *Clin Res Cardiol,* vol. 104, no. 9, pp. 705-18, Sep 2015, doi: 10.1007/s00392-015-0859-7.

[59] S. C. Nata and M. V. Launico, "Anatomy, Airway," in *StatPearls*. Treasure Island (FL): StatPearls Publishing Copyright © 2025, StatPearls Publishing LLC., 2025.

[60] A. Qureshi, R. D. Ballard, and H. S. Nelson, "Obstructive sleep apnea," *Journal of Allergy and Clinical Immunology,* vol. 112, no. 4, pp. 643-651, 2003, doi: 10.1016/j.jaci.2003.08.031.

[61] D. L. Morrison, S. H. Launois, S. Isono, T. R. Feroah, W. A. Whitelaw, and J. E. Remmers, "Pharyngeal narrowing and closing pressures in patients with obstructive sleep apnea," (in eng), *Am Rev Respir Dis,* vol. 148, no. 3, pp. 606-11, Sep 1993, doi: 10.1164/ajrccm/148.3.606.

[62] D. W. Hudgel and C. Hendricks, "Palate and hypopharynx--sites of inspiratory narrowing of the upper airway during sleep," (in eng), *Am Rev Respir Dis,* vol. 138, no. 6, pp. 1542-7, Dec 1988, doi: 10.1164/ajrccm/138.6.1542.

[63] A. D. Lucey *et al.*, "Nasal Septum Upper Esophageal Sphincte r Nasopharynx Velopharynx Oropharynx Hypopharynx Catheter," 2014.

[64] E. Morphologia, "Histology: A Text and Atlas: With Correlated Cell and Molecular Biology. Eighth Edition, 2018 Authors: Wojciech Pawlina; Michael H. Ross," *Morphologia,* vol. 13, pp. 76-89, 12/27 2019, doi: 10.26641/1997-9665.2019.4.76-89.

[65] C. Maspero, L. Giannini, G. Galbiati, G. Rosso, and G. Farronato, "Obstructive sleep apnea syndrome: a literature review," (in eng ita), *Minerva Stomatol,* vol. 64, no. 2, pp. 97-109, Apr 2015.

[66] G. R. Geovanini *et al.*, "Association between Obstructive Sleep Apnea and Cardiovascular Risk Factors: Variation by Age, Sex, and Race. The Multi-Ethnic Study of Atherosclerosis," (in eng), *Ann Am Thorac Soc,* vol. 15, no. 8, pp. 970-977, Aug 2018, doi: 10.1513/AnnalsATS.201802-121OC.

[67] "Global surveillance, prevention and control of chronic respiratory diseases." WHO. https://www.who.int/publications/i/item/global-surveillance-prevention-and-control-of-chronic-respiratory-diseases (accessed Oct. 23rd, 2025).

[68] N. F. Watson, "Health Care Savings: The Economic Value of Diagnostic and Therapeutic Care for Obstructive Sleep Apnea," (in eng), *J Clin Sleep Med,* vol. 12, no. 8, pp. 1075-7, Aug 15 2016, doi: 10.5664/jcsm.6034.

[69] P. Balagny *et al.*, "Prevalence, treatment and determinants of obstructive sleep apnoea and its symptoms in a population-based French cohort," (in eng), *ERJ Open Res,* vol. 9, no. 3, May 2023, doi: 10.1183/23120541.00053-2023.

[70] E. O. Bixler *et al.*, "Prevalence of sleep-disordered breathing in women: effects of gender," (in eng), *Am J Respir Crit Care Med,* vol. 163, no. 3 Pt 1, pp. 608-13, Mar 2001, doi: 10.1164/ajrccm.163.3.9911064.

[71] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," (in eng), *N Engl J Med,* vol. 328, no. 17, pp. 1230-5, Apr 29 1993, doi: 10.1056/nejm199304293281704.

[72] M. S. Ip, B. Lam, L. C. Tang, I. J. Lauder, T. Y. Ip, and W. K. Lam, "A community study of sleep-disordered breathing in middle-aged Chinese women in Hong Kong: prevalence and gender differences," (in eng), *Chest,* vol. 125, no. 1, pp. 127-34, Jan 2004, doi: 10.1378/chest.125.1.127.

[73] M. K. Reeves-Hoché, D. W. Hudgel, R. Meck, R. Witteman, A. Ross, and C. W. Zwillich, "Continuous versus bilevel positive airway pressure for obstructive sleep apnea," (in eng), *Am J Respir Crit Care Med,* vol. 151, no. 2 Pt 1, pp. 443-9, Feb 1995, doi: 10.1164/ajrccm.151.2.7842204.

[74] M. T. Bianchi and B. Goparaju, "Potential Underestimation of Sleep Apnea Severity by At-Home Kits: Rescoring In-Laboratory Polysomnography Without Sleep Staging," (in eng), *J Clin Sleep Med,* vol. 13, no. 4, pp. 551-555, Apr 15 2017, doi: 10.5664/jcsm.6540.

[75] S. P. Gunta *et al.*, "Obstructive Sleep Apnea and Cardiovascular Diseases: Sad Realities and Untold Truths regarding Care of Patients in 2022," (in eng), *Cardiovasc Ther,* vol. 2022, p. 6006127, 2022, doi: 10.1155/2022/6006127.

[76] P. Chazal, C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley, "Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea," *IEEE transactions on bio-medical engineering,* vol. 50, pp. 686-96, 07/01 2003, doi: 10.1109/TBME.2003.812203.

[77] T. Penzel, J. McNames, A. Murray, P. de Chazal, G. Moody, and B. Raymond, "Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings," (in eng), *Med Biol Eng Comput,* vol. 40, no. 4, pp. 402-7, Jul 2002, doi: 10.1007/bf02345072.

[78] M. Chyad, S. Gharghan, H. Hamood, A. Altayyar, S. Zubaidi, and H. Ridha, "Hybridization of soft-computing algorithms with neural network for prediction obstructive sleep apnea using biomedical sensor measurements," *Neural Computing and Applications,* vol. 34, pp. 1-25, 06/01 2022, doi: 10.1007/s00521-022-06919-w.

[79] L. C. Markun and A. Sampat, "Clinician-Focused Overview and Developments in Polysomnography," (in eng), *Curr Sleep Med Rep,* vol. 6, no. 4, pp. 309-321, 2020, doi: 10.1007/s40675-020-00197-5.

[80] "The 2007 AASM Scoring Manual vs. the AASM Scoring Manual v2.0 " The American Academy of Sleep Medicine https://aasm.org/wp-content/uploads/2017/11/Summary-of-Updates-in-v2.0-FINAL.pdf (accessed Oct. 23rd, 2025).

[81] C. Guilleminault and V. C. Abad, "Obstructive sleep apnea syndromes," (in eng), *Med Clin North Am,* vol. 88, no. 3, pp. 611-30, viii, May 2004, doi: 10.1016/j.mcna.2004.01.002.

[82] P. M. Macey, M. A. Woo, R. Kumar, R. L. Cross, and R. M. Harper, "Relationship between obstructive sleep apnea severity and sleep, depression and anxiety symptoms in newly-diagnosed patients," (in eng), *PLoS One,* vol. 5, no. 4, p. e10211, Apr 16 2010, doi: 10.1371/journal.pone.0010211.

[83] W. Randerath *et al.*, "Challenges and perspectives in obstructive sleep apnoea," *European Respiratory Journal,* vol. 52, no. 3, p. 1702616, 2018, doi: 10.1183/13993003.02616-2017.

[84] W. Whitelaw and K. Burgess, "Diagnosis of sleep apnoea: Some critical issues," *The Indian journal of medical research,* vol. 131, pp. 217-29, 02/01 2010.

[85] K. Ramar and C. Guilleminault, "Sleep apnea (central and obstructive)," in *Sleep Medicine*, H. R. Smith, C. L. Comella, and B. Högl Eds., (Cambridge Clinical Guides. Cambridge: Cambridge University Press, 2008, pp. 129-156.

[86] A. Abbasi *et al.*, "A comprehensive review of obstructive sleep apnea," (in eng), *Sleep Sci,* vol. 14, no. 2, pp. 142-154, Apr-Jun 2021, doi: 10.5935/1984-0063.20200056.

[87] M. El Shayeb, L. A. Topfer, T. Stafinski, L. Pawluk, and D. Menon, "Diagnostic accuracy of level 3 portable sleep tests versus level 1 polysomnography for sleep-disordered breathing: a systematic review and meta-analysis," (in eng), *Cmaj,* vol. 186, no. 1, pp. E25-51, Jan 7 2014, doi: 10.1503/cmaj.130952.

[88] D. E. Jonas *et al.*, "Screening for Obstructive Sleep Apnea in Adults: Evidence Report and Systematic Review for the US Preventive Services Task Force," (in eng), *Jama,* vol. 317, no. 4, pp. 415-433, Jan 24 2017, doi: 10.1001/jama.2016.19635.

[89] A. Guerrero *et al.*, "Management of sleep apnea without high pretest probability or with comorbidities by three nights of portable sleep monitoring," (in eng), *Sleep,* vol. 37, no. 8, pp. 1363-73, Aug 1 2014, doi: 10.5665/sleep.3932.

[90] E. J. Pereira, H. S. Driver, S. C. Stewart, and M. F. Fitzpatrick, "Comparing a combination of validated questionnaires and level III portable monitor with polysomnography to diagnose and exclude sleep apnea," (in eng), *J Clin Sleep Med,* vol. 9, no. 12, pp. 1259-66, Dec 15 2013, doi: 10.5664/jcsm.3264.

[91] M. R. Zeidler, V. Santiago, J. M. Dzierzewski, M. N. Mitchell, S. Santiago, and J. L. Martin, "Predictors of Obstructive Sleep Apnea on Polysomnography after a Technically Inadequate or Normal Home Sleep Test," (in eng), *J Clin Sleep Med,* vol. 11, no. 11, pp. 1313-8, Nov 15 2015, doi: 10.5664/jcsm.5194.

[92] T. Kirby, "Colin Sullivan: inventive pioneer of sleep medicine," (in eng), *Lancet,* vol. 377, no. 9776, p. 1485, Apr 30 2011, doi: 10.1016/s0140-6736(11)60589-8.

[93] Y. Donchin and F. J. Seagull, "The hostile environment of the intensive care unit," (in eng), *Curr Opin Crit Care,* vol. 8, no. 4, pp. 316-20, Aug 2002, doi: 10.1097/00075198-200208000-00008.

[94] A. Roebuck *et al.*, "A review of signals used in sleep analysis," (in eng), *Physiol Meas,* vol. 35, no. 1, pp. R1-57, Jan 2014, doi: 10.1088/0967-3334/35/1/r1.

[95] J. Haba-Rubio, E. Sforza, T. Weiss, C. Schröder, and J. Krieger, "Effect of CPAP treatment on inspiratory arousal threshold during NREM sleep in OSAS," *Sleep and Breathing,* vol. 9, no. 1, pp. 12-19, 2005/03/01 2005, doi: 10.1007/s11325-005-0002-5.

[96] L. R. Young, Z. H. Taxin, R. G. Norman, J. A. Walsleben, D. M. Rapoport, and I. Ayappa, "Response to CPAP withdrawal in patients with mild versus severe obstructive sleep apnea/hypopnea syndrome," (in eng), *Sleep,* vol. 36, no. 3, pp. 405-12, Mar 1 2013, doi: 10.5665/sleep.2460.

[97] S. J. Redmond, P. de Chazal, C. O'Brien, S. Ryan, W. T. McNicholas, and C. Heneghan, "Sleep staging using cardiorespiratory signals," *Somnologie - Schlafforschung und Schlafmedizin,* vol. 11, no. 4, pp. 245-256, 2007/12/01 2007, doi: 10.1007/s11818-007-0314-8.

[98] N. Canessa *et al.*, "Obstructive sleep apnea: brain structural changes and neurocognitive function before and after treatment," (in eng), *Am J Respir Crit Care Med,* vol. 183, no. 10, pp. 1419-26, May 15 2011, doi: 10.1164/rccm.201005-0693OC.

[99] P. E. Brander, M. Soirinsuo, and P. Lohela, "Nasopharyngeal symptoms in patients with obstructive sleep apnea syndrome. Effect of nasal CPAP treatment," (in eng), *Respiration,* vol. 66, no. 2, pp. 128-35, 1999, doi: 10.1159/000029354.

[100] N. B. Kribbs *et al.*, "Objective measurement of patterns of nasal CPAP use by patients with obstructive sleep apnea," (in eng), *Am Rev Respir Dis,* vol. 147, no. 4, pp. 887-95, Apr 1993, doi: 10.1164/ajrccm/147.4.887.

[101] J.-L. Pépin *et al.*, "Effective compliance during the first 3 months of continuous positive airway pressure. A European prospective study of 121 patients," *American journal of respiratory and critical care medicine,* vol. 160 4, pp. 1124-9, 1999.

[102] P. R. Genta *et al.*, "The Importance of Mask Selection on Continuous Positive Airway Pressure Outcomes for Obstructive Sleep Apnea. An Official American Thoracic Society Workshop Report," (in eng), *Ann Am Thorac Soc,* vol. 17, no. 10, pp. 1177-1185, Oct 2020, doi: 10.1513/AnnalsATS.202007-864ST.

[103] R. Mihai, M. Vandeleur, S. Pecoraro, M. J. Davey, and G. M. Nixon, "Autotitrating CPAP as a Tool for CPAP Initiation for Children," (in eng), *J Clin Sleep Med,* vol. 13, no. 5, pp. 713-719, May 15 2017, doi: 10.5664/jcsm.6590.

[104] S. P. Patil, I. A. Ayappa, S. M. Caples, R. J. Kimoff, S. R. Patel, and C. G. Harrod, "Treatment of Adult Obstructive Sleep Apnea with Positive Airway Pressure: An American Academy of Sleep Medicine Clinical Practice Guideline," (in eng), *J Clin Sleep Med,* vol. 15, no. 2, pp. 335-343, Feb 15 2019, doi: 10.5664/jcsm.7640.

[105] C. Perin and P. R. Genta, "Less may be more: CPAP vs. APAP in the treatment of obstructive sleep apnea," (in eng por), *J Bras Pneumol,* vol. 47, no. 6, p. e20210455, Dec 15 2021, doi: 10.36416/1806-3756/e20210455.

[106] L. M. Donovan, S. Boeder, A. Malhotra, and S. R. Patel, "New developments in the use of positive airway pressure for obstructive sleep apnea," (in eng), *J Thorac Dis,* vol. 7, no. 8, pp. 1323-42, Aug 2015, doi: 10.3978/j.issn.2072-1439.2015.07.30.

[107] A. Malhotra, C. R. Heilmann, K. K. Banerjee, J. P. Dunn, M. C. Bunck, and J. Bednarik, "Weight reduction and the impact on apnea-hypopnea index: A systematic meta-analysis," *Sleep Medicine,* vol. 121, pp. 26-31, 2024/09/01/ 2024, doi: 10.1016/j.sleep.2024.06.014.

[108] O. Baser, Y. Lu, S. Chen, and E. Baser, "Tirzepatide and Semaglutide for the Treatment of Obstructive Sleep Apnea and Obesity: A Retrospective Analysis," *Medical Research Archives; Vol 13 No 1 (2025): Vol.13 issue 1 January 2025,* 2025, doi: 10.18103/mra.v13i1.6236.

[109] J. Ngiam, K. Sutherland, R. Balasubramaniam, M. Marklund, F. Almeida, and P. Cistulli, "Oral Appliance Therapy for Sleep-Disordered Breathing," in *Contemporary Oral Medicine: A Comprehensive Approach to Clinical Practice*, C. S. Farah, R. Balasubramaniam, and M. J. McCullough Eds. Cham: Springer International Publishing, 2019, pp. 2303-2331.

[110] W. J. Randerath, M. Heise, R. Hinz, and K. H. Ruehle, "An individually adjustable oral appliance vs continuous positive airway pressure in mild-to-moderate obstructive sleep apnea syndrome," (in eng), *Chest,* vol. 122, no. 2, pp. 569-75, Aug 2002, doi: 10.1378/chest.122.2.569.

[111] H. M. Engleman *et al.*, "Randomized crossover trial of two treatments for sleep apnea/hypopnea syndrome: continuous positive airway pressure and mandibular repositioning splint," (in eng), *Am J Respir Crit Care Med,* vol. 166, no. 6, pp. 855-9, Sep 15 2002, doi: 10.1164/rccm.2109023.

[112] N. T. Phan, B. Wallwork, and B. Panizza, "Surgery for adult patients with obstructive sleep apnoea: A review for general practitioners," (in eng), *Aust Fam Physician,* vol. 45, no. 8, pp. 574-8, Aug 2016.

[113] N. R. Lee, C. D. Givens, Jr., J. Wilson, and R. B. Robins, "Staged surgical treatment of obstructive sleep apnea syndrome: a review of 35 patients," (in eng), *J Oral Maxillofac Surg,* vol. 57, no. 4, pp. 382-5, Apr 1999, doi: 10.1016/s0278-2391(99)90272-0.

[114] K. Li, R. Riley, N. Powell, R. Troell, and C. Guilleminault, "Overview of Phase II Surgery for Obstructive Sleep Apnea Syndrome," *Ear, Nose & Throat Journal,* vol. 78, pp. 851-857, 11/01 1999, doi: 10.1177/014556139907801109.

[115] A. Yoshihisa and Y. Takeishi, "Sleep Disordered Breathing and Cardiovascular Diseases," (in eng), *J Atheroscler Thromb,* vol. 26, no. 4, pp. 315-327, Apr 1 2019, doi: 10.5551/jat.RV17032.

[116] D. S. Martin and M. P. Grocott, "Oxygen therapy in critical illness: precise control of arterial oxygenation and permissive hypoxemia," (in eng), *Crit Care Med,* vol. 41, no. 2, pp. 423-32, Feb 2013, doi: 10.1097/CCM.0b013e31826a44f6.

[117] G. Weiss and L. T. Goodnough, "Anemia of chronic disease," (in eng), *N Engl J Med,* vol. 352, no. 10, pp. 1011-23, Mar 10 2005, doi: 10.1056/NEJMra041809.

[118] J. L. Vincent and D. De Backer, "Oxygen transport-the oxygen delivery controversy," (in eng), *Intensive Care Med,* vol. 30, no. 11, pp. 1990-6, Nov 2004, doi: 10.1007/s00134-004-2384-4.

[119] P. E. Peppard, T. Young, M. Palta, and J. Skatrud, "Prospective study of the association between sleep-disordered breathing and hypertension," (in eng), *N Engl J Med,* vol. 342, no. 19, pp. 1378-84, May 11 2000, doi: 10.1056/nejm200005113421901.

[120] F. J. Nieto *et al.*, "Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. Sleep Heart Health Study," (in eng), *Jama,* vol. 283, no. 14, pp. 1829-36, Apr 12 2000, doi: 10.1001/jama.283.14.1829.

[121] E. O. Bixler *et al.*, "Association of hypertension and sleep-disordered breathing," (in eng), *Arch Intern Med,* vol. 160, no. 15, pp. 2289-95, Aug 14-28 2000, doi: 10.1001/archinte.160.15.2289.

[122] J. M. Marin *et al.*, "Association between treated and untreated obstructive sleep apnea and risk of hypertension," (in eng), *Jama,* vol. 307, no. 20, pp. 2169-76, May 23 2012, doi: 10.1001/jama.2012.3418.

[123] C. Prinz, T. Bitter, C. Piper, D. Horstkotte, L. Faber, and O. Oldenburg, "Sleep apnea is common in patients with coronary artery disease," (in eng), *Wien Med Wochenschr,* vol. 160, no. 13-14, pp. 349-55, Jul 2010, doi: 10.1007/s10354-009-0737-x.

[124] H. Schäfer, U. Koehler, S. Ewig, E. Hasper, S. Tasci, and B. Lüderitz, "Obstructive sleep apnea as a risk marker in coronary artery disease," (in eng), *Cardiology,* vol. 92, no. 2, pp. 79-84, 1999, doi: 10.1159/000006952.

[125] Y. Peker, J. Hedner, J. Norum, H. Kraiczi, and J. Carlson, "Increased incidence of cardiovascular disease in middle-aged men with obstructive sleep apnea: a 7-year follow-up," (in eng), *Am J Respir Crit Care Med,* vol. 166, no. 2, pp. 159-65, Jul 15 2002, doi: 10.1164/rccm.2105124.

[126] D. D. Sin, F. Fitzgerald, J. D. Parker, G. Newton, J. S. Floras, and T. D. Bradley, "Risk factors for central and obstructive sleep apnea in 450 men and women with congestive heart failure," (in eng), *Am J Respir Crit Care Med,* vol. 160, no. 4, pp. 1101-6, Oct 1999, doi: 10.1164/ajrccm.160.4.9903020.

[127] A.-L. Pintilie *et al.*, "Sleep Apnea: The Slept-Upon Cardiovascular Risk Factor," *Biomedicines,* vol. 13, no. 10, p. 2529, 2025. [Online]. Available: https://www.mdpi.com/2227-9059/13/10/2529.

[128] T. Bitter, L. Faber, D. Hering, C. Langer, D. Horstkotte, and O. Oldenburg, "Sleep-disordered breathing in heart failure with normal left ventricular ejection fraction," (in eng), *Eur J Heart Fail,* vol. 11, no. 6, pp. 602-8, Jun 2009, doi: 10.1093/eurjhf/hfp057.

[129] A. S. Gami *et al.*, "Association of atrial fibrillation and obstructive sleep apnea," (in eng), *Circulation,* vol. 110, no. 4, pp. 364-7, Jul 27 2004, doi: 10.1161/01.Cir.0000136587.68725.8e.

[130] N. M. Punjabi *et al.*, "Sleep-disordered breathing and mortality: a prospective cohort study," (in eng), *PLoS Med,* vol. 6, no. 8, p. e1000132, Aug 2009, doi: 10.1371/journal.pmed.1000132.

[131] T. Leppänen, A. Kulkas, A. Oksenberg, B. Duce, E. Mervaala, and J. Töyräs, "Differences in arousal probability and duration after apnea and hypopnea events in adult obstructive sleep apnea patients," (in eng), *Physiol Meas,* vol. 39, no. 11, p. 114004, Nov 6 2018, doi: 10.1088/1361-6579/aae42c.

[132] A. Azarbarzin *et al.*, "The Sleep Apnea-Specific Hypoxic Burden Predicts Incident Heart Failure," (in eng), *Chest,* vol. 158, no. 2, pp. 739-750, Aug 2020, doi: 10.1016/j.chest.2020.03.053.

[133] A. Muraja-Murro *et al.*, "Adjustment of apnea-hypopnea index with severity of obstruction events enhances detection of sleep apnea patients with the highest risk of severe health consequences," (in eng), *Sleep Breath,* vol. 18, no. 3, pp. 641-7, Sep 2014, doi: 10.1007/s11325-013-0927-z.

[134] B. Vanessa *et al.*, "Sleep apnoea and endothelial dysfunction: An individual patient data meta-analysis," *Sleep Medicine Reviews,* vol. 52, p. 101309, 03/01 2020, doi: 10.1016/j.smrv.2020.101309.

[135] Z. Huang *et al.*, "Implication of prolonged nocturnal hypoxemia and obstructive sleep apnea for pulmonary hemodynamics in patients being evaluated for pulmonary hypertension: a retrospective study," (in eng), *J Clin Sleep Med,* vol. 19, no. 2, pp. 213-223, Feb 1 2023, doi: 10.5664/jcsm.10286.

[136] J. Theorell-Haglöw, C. Berne, C. Janson, and E. Lindberg, "The role of obstructive sleep apnea in metabolic syndrome: a population-based study in women," (in eng), *Sleep Med,* vol. 12, no. 4, pp. 329-34, Apr 2011, doi: 10.1016/j.sleep.2010.06.014.

[137] R. Deo *et al.*, "Electrocardiographic Measures and Prediction of Cardiovascular and Noncardiovascular Death in CKD," (in eng), *J Am Soc Nephrol,* vol. 27, no. 2, pp. 559-69, Feb 2016, doi: 10.1681/asn.2014101045.

[138] B. Lechat *et al.*, "A Novel EEG Derived Measure of Disrupted Delta Wave Activity during Sleep Predicts All-Cause Mortality Risk," *Annals of the American Thoracic Society,* vol. 19, 10/21 2021, doi: 10.1513/AnnalsATS.202103-315OC.

[139] J. Pan, J. Wu, J. Liu, J. Wu, and F. Wang, "A Systematic Review of Sleep in Patients with Disorders of Consciousness: From Diagnosis to Prognosis," *Brain Sciences,* vol. 11, p. 1072, 08/16 2021, doi: 10.3390/brainsci11081072.

[140] S. Ryan and W. T. McNicholas, "Intermittent hypoxia and activation of inflammatory molecular pathways in OSAS," (in eng), *Arch Physiol Biochem,* vol. 114, no. 4, pp. 261-6, Oct 2008, doi: 10.1080/13813450802307337.

[141] T. Kendzerska, A. S. Gershon, G. Hawker, R. S. Leung, and G. Tomlinson, "Obstructive sleep apnea and risk of cardiovascular events and all-cause mortality: a decade-long historical cohort study," (in eng), *PLoS Med,* vol. 11, no. 2, p. e1001599, Feb 2014, doi: 10.1371/journal.pmed.1001599.

[142] O. Oldenburg *et al.*, "Nocturnal hypoxaemia is associated with increased mortality in stable heart failure patients," (in eng), *Eur Heart J,* vol. 37, no. 21, pp. 1695-703, Jun 1 2016, doi: 10.1093/eurheartj/ehv624.

[143] K. L. Stone *et al.*, "Sleep Disordered Breathing and Risk of Stroke in Older Community-Dwelling Men," (in eng), *Sleep,* vol. 39, no. 3, pp. 531-40, Mar 1 2016, doi: 10.5665/sleep.5520.

[144] T. Kendzerska *et al.*, "Cardiovascular consequences of obstructive sleep apnea in women: a historical cohort study," (in eng), *Sleep Med,* vol. 68, pp. 71-79, Apr 2020, doi: 10.1016/j.sleep.2019.08.021.

[145] P. H. Xu, D. Y. T. Fong, M. M. S. Lui, D. C. L. Lam, and M. S. M. Ip, "Cardiovascular outcomes in obstructive sleep apnoea and implications of clinical phenotyping on effect of CPAP treatment," (in eng), *Thorax,* vol. 78, no. 1, pp. 76-84, Jan 2023, doi: 10.1136/thoraxjnl-2021-217714.

[146] M. F. Damiani *et al.*, "Obstructive Sleep Apnea, Hypertension, and Their Additive Effects on Atherosclerosis," (in eng), *Biochem Res Int,* vol. 2015, p. 984193, 2015, doi: 10.1155/2015/984193.

[147] F. Frangopoulos, I. Nicolaou, S. Zannetos, N. T. Economou, T. Adamide, and G. Trakada, "Association between Respiratory Sleep Indices and Cardiovascular Disease in Sleep Apnea-A Community-Based Study in Cyprus," (in eng), *J Clin Med,* vol. 9, no. 8, Aug 1 2020, doi: 10.3390/jcm9082475.

[148] K. Sutherland *et al.*, "Comparative associations of oximetry patterns in Obstructive Sleep Apnea with incident cardiovascular disease," *Sleep,* vol. 45, 07/27 2022, doi: 10.1093/sleep/zsac179.

[149] W. Cao, J. Luo, and Y. Xiao, "A Review of Current Tools Used for Evaluating the Severity of Obstructive Sleep Apnea," (in eng), *Nat Sci Sleep,* vol. 12, pp. 1023-1031, 2020, doi: 10.2147/nss.S275252.

[150] A. Polytarchou *et al.*, "Nocturnal oximetry parameters as predictors of sleep apnea severity in resource-limited settings," (in eng), *J Sleep Res,* vol. 32, no. 1, p. e13638, Feb 2023, doi: 10.1111/jsr.13638.

[151] E. Borsini and C. A. Nigro, "Proposal of a diagnostic algorithm based on the use of pulse oximetry in obstructive sleep apnea," (in eng), *Sleep Breath,* vol. 27, no. 5, pp. 1677-1686, Oct 2023, doi: 10.1007/s11325-022-02757-1.

[152] T. Karhu, S. Myllymaa, S. Nikkonen, D. R. Mazzotti, J. Töyräs, and T. Leppänen, "Longer and Deeper Desaturations Are Associated With the Worsening of Mild Sleep Apnea: The Sleep Heart Health Study," (in eng), *Front Neurosci,* vol. 15, p. 657126, 2021, doi: 10.3389/fnins.2021.657126.

[153] Y. Ng *et al.*, "Oxygen Desaturation Index Differs Significantly Between Types of Sleep Software," (in eng), *J Clin Sleep Med,* vol. 13, no. 4, pp. 599-605, Apr 15 2017, doi: 10.5664/jcsm.6552.

[154] D. Temirbekov, S. Güneş, Z. M. Yazıcı, and İ. Sayın, "The Ignored Parameter in the Diagnosis of Obstructive Sleep Apnea Syndrome: The Oxygen Desaturation Index," (in eng), *Turk Arch Otorhinolaryngol,* vol. 56, no. 1, pp. 1-6, Mar 2018, doi: 10.5152/tao.2018.3025.

[155] F. Chung, P. Liao, H. Elsaid, S. Islam, C. M. Shapiro, and Y. Sun, "Oxygen desaturation index from nocturnal oximetry: a sensitive and specific tool to detect sleep-disordered breathing in surgical patients," (in eng), *Anesth Analg,* vol. 114, no. 5, pp. 993-1000, May 2012, doi: 10.1213/ANE.0b013e318248f4f5.

[156] L. Q. N. Liew *et al.*, "Nocturnal Oxygen Desaturation Index Correlates with Respiratory Depression in Post-Surgical Patients Receiving Opioids - A Post-Hoc Analysis from the Prediction of Opioid-Induced Respiratory Depression in Patients Monitored by Capnography (PRODIGY) Study," (in eng), *Nat Sci Sleep,* vol. 14, pp. 805-817, 2022, doi: 10.2147/nss.S351840.

[157] D. Alvarez, R. Hornero, M. García, F. del Campo, and C. Zamarrón, "Improving diagnostic ability of blood oxygen saturation from overnight pulse oximetry in obstructive sleep apnea detection by means of central tendency measure," (in eng), *Artif Intell Med,* vol. 41, no. 1, pp. 13-24, Sep 2007, doi: 10.1016/j.artmed.2007.06.002.

[158] S. Gyulay, L. G. Olson, M. J. Hensley, M. T. King, K. M. Allen, and N. A. Saunders, "A comparison of clinical assessment and home oximetry in the diagnosis of obstructive sleep apnea," (in eng), *Am Rev Respir Dis,* vol. 147, no. 1, pp. 50-3, Jan 1993, doi: 10.1164/ajrccm/147.1.50.

[159] L. W. Hang *et al.*, "Validation of overnight oximetry to diagnose patients with moderate to severe obstructive sleep apnea," (in eng), *BMC Pulm Med,* vol. 15, p. 24, Mar 20 2015, doi: 10.1186/s12890-015-0017-z.

[160] N. H. Rashid, S. Zaghi, M. Scapuccin, M. Camacho, V. Certal, and R. Capasso, "The Value of Oxygen Desaturation Index for Diagnosing Obstructive Sleep Apnea: A Systematic Review," (in eng), *Laryngoscope,* vol. 131, no. 2, pp. 440-447, Feb 2021, doi: 10.1002/lary.28663.

[161] C. L. Chai-Coetzer *et al.*, "Predictors of long-term adherence to continuous positive airway pressure therapy in patients with obstructive sleep apnea and cardiovascular disease in the SAVE study," (in eng), *Sleep,* vol. 36, no. 12, pp. 1929-37, Dec 1 2013, doi: 10.5665/sleep.3232.

[162] L. Varghese, G. Rebekah, P. N, A. Oliver, and R. Kurien, "Oxygen desaturation index as alternative parameter in screening patients with severe obstructive sleep apnea," (in eng), *Sleep Sci,* vol. 15, no. Spec 1, pp. 224-228, Jan-Mar 2022, doi: 10.5935/1984-0063.20200119.

[163] N. M. Punjabi, A. B. Newman, T. B. Young, H. E. Resnick, and M. H. Sanders, "Sleep-disordered breathing and cardiovascular disease: an outcome-based definition of hypopneas," (in eng), *Am J Respir Crit Care Med,* vol. 177, no. 10, pp. 1150-5, May 15 2008, doi: 10.1164/rccm.200712-1884OC.

[164] I. T. Ling, A. L. James, and D. R. Hillman, "Interrelationships between body mass, oxygen desaturation, and apnea-hypopnea indices in a sleep clinic population," (in eng), *Sleep,* vol. 35, no. 1, pp. 89-96, Jan 1 2012, doi: 10.5665/sleep.1592.

[165] L. Mo *et al.*, "Severe obstructive sleep apnea is associated with significant coronary artery plaque burden independent of traditional cardiovascular risk factors," (in eng), *Int J Cardiovasc Imaging,* vol. 36, no. 2, pp. 347-355, Feb 2020, doi: 10.1007/s10554-019-01710-w.

[166] A. Kulkas, P. Tiihonen, P. Julkunen, E. Mervaala, and J. Töyräs, "Novel parameters indicate significant differences in severity of obstructive sleep apnea with patients

having similar apnea-hypopnea index," (in eng), *Med Biol Eng Comput,* vol. 51, no. 6, pp. 697-708, Jun 2013, doi: 10.1007/s11517-013-1039-4.

[167] D. Linz *et al.*, "Nocturnal hypoxemic burden is associated with epicardial fat volume in patients with acute myocardial infarction," (in eng), *Sleep Breath,* vol. 22, no. 3, pp. 703-711, Sep 2018, doi: 10.1007/s11325-017-1616-0.

[168] M. Bonnet *et al.*, "EEG arousals: Scoring rules and examples. A preliminary report from the Sleep Disorders Atlas Task Force of the American Sleep Disorder Association," *Sleep,* vol. 15, pp. 173-184, 01/01 1992.

[169] A. Kales, A. Rechtschaffen, L. A. B. I. S. University of California, and N. N. I. Network, *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Allan Rechtschaffen and Anthony Kales, editors* (National Institutes of Health publication, no. 204). Bethesda, Md: U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968.

[170] "Sleep Heart Health Study: Overview of Scoring Manual." National Sleep Research Resource. https://sleepdata.org/datasets/shhs/pages/mop/6-610-mop-overview-of-scoring.md (accessed Oct. 23rd, 2025).

[171] W. Trzepizur *et al.*, "Sleep Apnea-Specific Hypoxic Burden, Symptom Subtypes, and Risk of Cardiovascular Events and All-Cause Mortality," (in eng), *Am J Respir Crit Care Med,* vol. 205, no. 1, pp. 108-117, Jan 1 2022, doi: 10.1164/rccm.202105-1274OC.

[172] R. Mehra and A. Azarbarzin, "Sleep Apnea-Specific Hypoxic Burden and Not the Sleepy Phenotype as a Novel Measure of Cardiovascular and Mortality Risk in a Clinical Cohort," (in eng), *Am J Respir Crit Care Med,* vol. 205, no. 1, pp. 12-13, Jan 1 2022, doi: 10.1164/rccm.202110-2371ED.

[173] M. A. Martinez-Garcia, M. Sánchez-de-la-Torre, D. P. White, and A. Azarbarzin, "Hypoxic Burden in Obstructive Sleep Apnea: Present and Future," (in eng spa), *Arch Bronconeumol,* vol. 59, no. 1, pp. 36-43, Jan 2023, doi: 10.1016/j.arbres.2022.08.005.

[174] M. Blanchard *et al.*, "Hypoxic burden and heart rate variability predict stroke incidence in sleep apnoea," (in eng), *Eur Respir J,* vol. 57, no. 3, Mar 2021, doi: 10.1183/13993003.04022-2020.

[175] J. S. Kim *et al.*, "Association of novel measures of sleep disturbances with blood pressure: the Multi-Ethnic Study of Atherosclerosis," (in eng), *Thorax,* vol. 75, no. 1, pp. 57-63, Jan 2020, doi: 10.1136/thoraxjnl-2019-213533.

[176] T. Karhu, T. Leppänen, J. Töyräs, A. Oksenberg, S. Myllymaa, and S. Nikkonen, "ABOSA - Freely available automatic blood oxygen saturation signal analysis software: Structure and validation," (in eng), *Comput Methods Programs Biomed,* vol. 226, p. 107120, Nov 2022, doi: 10.1016/j.cmpb.2022.107120.

[177] J. C. Jun, "Dying with OSA, or from It: A Cautionary Note about Novel Hypoxia Metrics," (in eng), *Am J Respir Crit Care Med,* vol. 206, no. 12, pp. 1563-1564, Dec 15 2022, doi: 10.1164/rccm.202206-1052LE.

[178] M. Rissanen *et al.*, "Obstructive Sleep Apnea Patients With Atrial Arrhythmias Suffer From Prolonged Recovery From Desaturations," (in eng), *IEEE Trans Biomed Eng,* vol. 70, no. 7, pp. 2122-2130, Jul 2023, doi: 10.1109/tbme.2023.3236680.

[179] Y. Liu, A. Abdul Ghafoor, M. Hajipour, and N. Ayas, "Role of precision medicine in obstructive sleep apnoea," (in eng), *BMJ Med,* vol. 2, no. 1, p. e000218, 2023, doi: 10.1136/bmjmed-2022-000218.

[180] P. Pahari *et al.*, "Obstructive sleep apnea-related intermittent hypoxaemia is associated with impaired vigilance," (in eng), *J Sleep Res,* vol. 32, no. 3, p. e13803, Jun 2023, doi: 10.1111/jsr.13803.

[181] S. Kainulainen *et al.*, "Severity of Desaturations Reflects OSA-Related Daytime Sleepiness Better Than AHI," (in eng), *J Clin Sleep Med,* vol. 15, no. 8, pp. 1135-1142, Aug 15 2019, doi: 10.5664/jcsm.7806.

[182] S. Kainulainen, "Pulse Oximetry-Derived Biomarkers for Severity Assessment of Obstructive Sleep Apnea: Associating Parametric and Frequency-Domain Features of SPO2 and PPG Signals with Daytime Sleepiness and Impaired Vigilance. ," PhD, University of Eastern Finlan, Kuopio, Finland, Volume 387, 2020.

[183] S. Kainulainen *et al.*, "Severe desaturations increase psychomotor vigilance task-based median reaction time and number of lapses in obstructive sleep apnoea patients," (in eng), *Eur Respir J,* vol. 55, no. 4, Apr 2020, doi: 10.1183/13993003.01849-2019.

[184] S. Kainulainen *et al.*, "Power spectral densities of nocturnal pulse oximetry signals differ in OSA patients with and without daytime sleepiness," (in eng), *Sleep Med,* vol. 73, pp. 231-237, Sep 2020, doi: 10.1016/j.sleep.2020.07.015.

[185] S. Hietakoste *et al.*, "Obstructive sleep apnoea-related respiratory events and desaturation severity are associated with the cardiac response," (in eng), *ERJ Open Res,* vol. 8, no. 4, Oct 2022, doi: 10.1183/23120541.00121-2022.

[186] T. Karhu *et al.*, "Diabetes and cardiovascular diseases are associated with the worsening of intermittent hypoxaemia," (in eng), *J Sleep Res,* vol. 31, no. 1, p. e13441, Feb 2022, doi: 10.1111/jsr.13441.

[187] F. D. Sigurdardottir *et al.*, "Novel oxygen desaturation parameters are associated with cardiac troponin I: Data from the Akershus Sleep Apnea Project," (in eng), *J Sleep Res,* vol. 31, no. 5, p. e13581, Oct 2022, doi: 10.1111/jsr.13581.

[188] W. Cao, J. Luo, R. Huang, and Y. Xiao, "Implication of a novel measure of obstructive sleep apnea severity for cardiovascular morbidity," (in eng), *Sleep Med,* vol. 103, pp. 204-210, Mar 2023, doi: 10.1016/j.sleep.2023.02.001.

[189] T. Leppänen, J. Töyräs, E. Mervaala, T. Penzel, and A. Kulkas, "Severity of individual obstruction events increases with age in patients with obstructive sleep apnea," (in eng), *Sleep Med,* vol. 37, pp. 32-37, Sep 2017, doi: 10.1016/j.sleep.2017.06.004.

[190] S. Thanaviratananich, H. Cheng, N. Chirakalwasan, and S. Reutrakul, "Association between nocturnal hypoxemic burden and glucose metabolism," (in eng), *Sleep Breath,* vol. 26, no. 3, pp. 1465-1470, Sep 2022, doi: 10.1007/s11325-021-02464-3.

[191] S. Khoshkish *et al.*, "The association between different features of sleep-disordered breathing and blood pressure: A cross-sectional study," (in eng), *J Clin Hypertens (Greenwich),* vol. 20, no. 3, pp. 575-581, Mar 2018, doi: 10.1111/jch.13202.

[192] D. Álvarez, G. C. Gutiérrez-Tobal, F. Vaquerizo-Villar, F. Moreno, F. del Campo, and R. Hornero, "Oximetry Indices in the Management of Sleep Apnea: From Overnight Minimum Saturation to the Novel Hypoxemia Measures," in *Advances in the Diagnosis and Treatment of Sleep Apnea : Filling the Gap Between Physicians and Engineers*, T. Penzel and R. Hornero Eds. Cham: Springer International Publishing, 2022, pp. 219-239.

[193] S. Reutrakul and B. Mokhlesi, "Obstructive Sleep Apnea and Diabetes: A State of the Art Review," (in eng), *Chest,* vol. 152, no. 5, pp. 1070-1086, Nov 2017, doi: 10.1016/j.chest.2017.05.009.

[194] V. K. Kapur, H. E. Resnick, and D. J. Gottlieb, "Sleep disordered breathing and hypertension: does self-reported sleepiness modify the association?," (in eng), *Sleep,* vol. 31, no. 8, pp. 1127-32, Aug 2008.

[195] A. Cai, L. Wang, and Y. Zhou, "Hypertension and obstructive sleep apnea," (in eng), *Hypertens Res,* vol. 39, no. 6, pp. 391-5, Jun 2016, doi: 10.1038/hr.2016.11.

[196] J. Doumit and B. Prasad, "Sleep Apnea in Type 2 Diabetes," (in eng), *Diabetes Spectr,* vol. 29, no. 1, pp. 14-9, Feb 2016, doi: 10.2337/diaspect.29.1.14.

[197] G. Labarca, J. Campos, K. Thibaut, J. Dreyse, and J. Jorquera, "Do T90 and SaO(2) nadir identify a different phenotype in obstructive sleep apnea?," (in eng), *Sleep Breath,* vol. 23, no. 3, pp. 1007-1010, Sep 2019, doi: 10.1007/s11325-019-01860-0.

[198] S. M. Hassan *et al.*, "Polysomnography-derived Hypoxemic Markers Associated With Pulmonary Hypertension in Obstructive Sleep Apnea," presented at the American Thoracic Society 2023 International Conference, Washington, DC, USA, 2023.

[199] D. Linz *et al.*, "Low Prognostic Value of Novel Nocturnal Metrics in Patients With OSA and High Cardiovascular Event Risk: Post Hoc Analyses of the SAVE Study," (in eng), *Chest,* vol. 158, no. 6, pp. 2621-2631, Dec 2020, doi: 10.1016/j.chest.2020.06.072.

[200] W. T. Wu *et al.*, "Utility of overnight pulse oximeter as a screening tool for sleep apnea to assess the 8-year risk of cardiovascular disease: Data from a large-scale bus driver cohort study," (in eng), *Int J Cardiol,* vol. 225, pp. 206-212, Dec 15 2016, doi: 10.1016/j.ijcard.2016.09.110.

[201] A. Leino *et al.*, "Acute stroke and TIA patients have specific polygraphic features of obstructive sleep apnea," (in eng), *Sleep Breath,* vol. 24, no. 4, pp. 1495-1505, Dec 2020, doi: 10.1007/s11325-019-02010-2.

[202] T. Young *et al.*, "Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort," (in eng), *Sleep,* vol. 31, no. 8, pp. 1071-8, Aug 2008.

[203] F. P. Cappuccio, D. Cooper, L. D'Elia, P. Strazzullo, and M. A. Miller, "Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies," (in eng), *Eur Heart J,* vol. 32, no. 12, pp. 1484-92, Jun 2011, doi: 10.1093/eurheartj/ehr007.

[204] S. I. Gunnarsson *et al.*, "Minimal nocturnal oxygen saturation predicts future subclinical carotid atherosclerosis: the Wisconsin sleep cohort," (in eng), *J Sleep Res,* vol. 24, no. 6, pp. 680-6, Dec 2015, doi: 10.1111/jsr.12321.

[205] J. A. Damen *et al.*, "Prediction models for cardiovascular disease risk in the general population: systematic review," (in eng), *Bmj,* vol. 353, p. i2416, May 16 2016, doi: 10.1136/bmj.i2416.

[206] R. Polman, J. R. Hurst, O. F. Uysal, S. Mandal, D. Linz, and S. Simons, "Cardiovascular disease and risk in COPD: a state of the art review," (in eng), *Expert Rev Cardiovasc Ther,* vol. 22, no. 4-5, pp. 177-191, Apr-May 2024, doi: 10.1080/14779072.2024.2333786.

[207] A. Løkke *et al.*, "Exacerbations Predict Severe Cardiovascular Events in Patients with COPD and Stable Cardiovascular Disease-A Nationwide, Population-Based Cohort Study," (in eng), *Int J Chron Obstruct Pulmon Dis,* vol. 18, pp. 419-429, 2023, doi: 10.2147/copd.S396790.

[208] A. Eisen *et al.*, "Angina and Future Cardiovascular Events in Stable Patients With Coronary Artery Disease: Insights From the Reduction of Atherothrombosis for Continued Health (REACH) Registry," (in eng), *J Am Heart Assoc,* vol. 5, no. 10, Sep 28 2016, doi: 10.1161/jaha.116.004080.

[209] M. H. Katz and W. W. Hauck, "Proportional hazards (Cox) regression," *Journal of General Internal Medicine,* vol. 8, no. 12, pp. 702-711, 1993/12/01 1993, doi: 10.1007/BF02598295.

[210] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part I: basic concepts and first analyses," (in eng), *Br J Cancer,* vol. 89, no. 2, pp. 232-8, Jul 21 2003, doi: 10.1038/sj.bjc.6601118.

[211] R. Singh and K. Mukhopadhyay, "Survival analysis in clinical trials: Basics and must know areas," (in eng), *Perspect Clin Res,* vol. 2, no. 4, pp. 145-8, Oct 2011, doi: 10.4103/2229-3485.86872.

[212] S. V. Deo, V. Deo, and V. Sundaram, "Survival analysis-part 1," (in eng), *Indian J Thorac Cardiovasc Surg,* vol. 36, no. 6, pp. 668-672, Nov 2020, doi: 10.1007/s12055-020-01049-1.
[213] D. G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition*. Springer New York, 2011.
[214] C. Starbuck, "Linear Model Extensions," in *The Fundamentals of People Analytics: With Applications in R*, C. Starbuck Ed. Cham: Springer International Publishing, 2023, pp. 207-221.
[215] D. R. Cox, *Analysis of survival data*. Chapman and Hall/CRC, 1984.
[216] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 34, no. 2, pp. 187-202, 1972.
[217] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, "Survival analysis part II: multivariate data analysis–an introduction to concepts and methods," *British journal of cancer,* vol. 89, no. 3, pp. 431-436, 2003.
[218] S. Abd ElHafeez, G. D'Arrigo, D. Leonardis, M. Fusaro, G. Tripepi, and S. Roumeliotis, "Methods to Analyze Time-to-Event Data: The Cox Regression Analysis," (in eng), *Oxid Med Cell Longev,* vol. 2021, p. 1302811, 2021, doi: 10.1155/2021/1302811.
[219] M. Stevenson, *An Introduction to Survival Analysis*. EpiCentre: IVABS, Massey University, 2007.
[220] J. M. Bland and D. G. Altman, "The logrank test," (in eng), *Bmj,* vol. 328, no. 7447, p. 1073, May 1 2004, doi: 10.1136/bmj.328.7447.1073.
[221] M. Babińska, J. Chudek, E. Chełmecka, M. Janik, K. Klimek, and A. Owczarek, "Limitations of Cox Proportional Hazards Analysis in Mortality Prediction of Patients with Acute Coronary Syndrome," *Studies in Logic, Grammar and Rhetoric,* vol. 43, 12/01 2015, doi: 10.1515/slgr-2015-0040.
[222] S. Branders, B. Frénay, and P. Dupont, *Survival Analysis with Cox Regression and Random Non-linear Projections*. 2015.
[223] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap Eds. Cham: Springer International Publishing, 2020, pp. 3-21.
[224] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," (in eng), *Behav Ther,* vol. 51, no. 5, pp. 675-687, Sep 2020, doi: 10.1016/j.beth.2020.05.002.
[225] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B,* vol. 4, pp. 51-62, 12/15 2017, doi: 10.20544/HORIZONS.B.04.1.17.P05.
[226] P. Pooja and C. Parul, "A Comprehensive Review of Various Machine Learning Techniques," in *Explainable Machine Learning Models and Architectures*: Wiley, 2023, pp. 1-10.
[227] E. K. Tang, P. N. Suganthan, X. Yao, and A. K. Qin, "Linear dimensionality reduction using relevance weighted LDA," *Pattern Recognition,* vol. 38, no. 4, pp. 485-493, 2005/04/01/ 2005, doi: 10.1016/j.patcog.2004.09.005.
[228] I. El-Feghi, M. A. Sid-Ahmed, and M. Ahmadi, "Automatic localization of craniofacial landmarks for assisted cephalometry," *Pattern Recognition,* vol. 37, no. 3, pp. 609-621, 2004/03/01/ 2004, doi: 10.1016/j.patcog.2003.09.002.
[229] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recognition,* vol. 37, no. 5, pp. 857-874, 2004/05/01/ 2004, doi: 10.1016/j.patcog.2003.12.002.

[230] Ş. Büyüköztürk and Ö. Çokluk Bökeoğlu, "Discriminant Function Analysis: Concept and Application," *Egitim Arastirmalari - Eurasian Journal of Educational Research,* vol. 8, pp. 73-92, 01/01 2008.
[231] J. Hansen, "Using SPSS for Windows and Macintosh: Analyzing and Understanding Data," *The American Statistician,* vol. 59, no. 1, pp. 113-113, 2005/02/01 2005, doi: 10.1198/tas.2005.s139.
[232] P. A. Lachenbruch, *Discriminant Analysis*. Hafner Press, 1975.
[233] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data — with application to face recognition," *Pattern Recognition,* vol. 34, no. 10, pp. 2067-2070, 2001/10/01/ 2001, doi: 10.1016/S0031-3203(00)00162-X.
[234] R. P. W. Duin and M. Loog, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, no. 6, pp. 732-739, 2004, doi: 10.1109/TPAMI.2004.13.
[235] T. Hastie and R. Tibshirani, "Discriminant Analysis by Gaussian Mixtures," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 58, no. 1, pp. 155-176, 1996. [Online]. Available: http://www.jstor.org/stable/2346171.
[236] J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, no. 7, pp. 763-767, 1996, doi: 10.1109/34.506799.
[237] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association,* vol. 84, no. 405, pp. 165-175, 1989, doi: 10.2307/2289860.
[238] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes," presented at the Proceedings of the 15th International Conference on Neural Information Processing Systems: Natural and Synthetic, Vancouver, British Columbia, Canada, 2001.
[239] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," (in eng), *Cancer Genomics Proteomics,* vol. 15, no. 1, pp. 41-51, Jan-Feb 2018, doi: 10.21873/cgp.20063.
[240] M. Stitson, J. Weston, A. Gammerman, V. Vovk, and V. Vapnik, "Theory of Support Vector Machines," 01/01 1996.
[241] A. Shmilovici, "Support Vector Machines," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach Eds. Boston, MA: Springer US, 2005, pp. 257-276.
[242] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information,* vol. 15, p. 235, 04/19 2024, doi: 10.3390/info15040235.
[243] Y.-x. Hu, J. N. K. Liu, and L.-w. Jia, "Sensitivity and Generalization of SVM with Weighted and Reduced Features," in *Reliable Knowledge Discovery*, Boston, MA, H. Dai, J. N. K. Liu, and E. Smirnov, Eds., 2012// 2012: Springer US, pp. 161-182.
[244] C.-w. Hsu, C.-c. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin," 11/29 2003.
[245] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," (in English), *Frontiers of Computer Science,* vol. 14, no. 2, pp. 241-258, Apr 2020 2024-08-27 2020, doi: 10.1007/s11704-019-8208-z.
[246] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, Berlin, Heidelberg, 2000// 2000: Springer Berlin Heidelberg, pp. 1-15.
[247] D. Nguyen *et al.*, "Ensemble learning using traditional machine learning and deep neural network for diagnosis of Alzheimer's disease," (in eng), *IBRO Neurosci Rep,* vol. 13, pp. 255-263, Dec 2022, doi: 10.1016/j.ibneur.2022.08.010.

[248] S. M. Ganie and M. B. Malik, "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators," *Healthcare Analytics,* vol. 2, p. 100092, 2022/11/01/ 2022, doi: 10.1016/j.health.2022.100092.

[249] K. Liu *et al.*, "Building an ensemble learning model for gastric cancer cell line classification via rapid raman spectroscopy," *Computational and Structural Biotechnology Journal,* vol. 21, pp. 802-811, 2023/01/01/ 2023, doi: 10.1016/j.csbj.2022.12.050.

[250] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, no. 2, pp. 123-140, 1996/08/01 1996, doi: 10.1023/A:1018054314350.

[251] R. E. Schapire, "The Boosting Approach to Machine Learning: An Overview," in *Nonlinear Estimation and Classification*, D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu Eds. New York, NY: Springer New York, 2003, pp. 149-171.

[252] K. Fawagreh, M. M. Gaber, and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering,* vol. 2, no. 1, pp. 602-609, 2014/12/01 2014, doi: 10.1080/21642583.2014.956265.

[253] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *International Journal of Data Science and Analytics,* vol. 20, no. 2, pp. 197-211, 2025/08/01 2025, doi: 10.1007/s41060-024-00509-w.

[254] H. Tin Kam, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 14-16 Aug. 1995 1995, vol. 1, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

[255] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," *Neural Computation,* vol. 9, no. 7, pp. 1545-1588, 1997, doi: 10.1162/neco.1997.9.7.1545.

[256] H. Tin Kam, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 832-844, 1998, doi: 10.1109/34.709601.

[257] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees," CRC Press, 1984, ch. SPLITTING RULES.

[258] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley interdisciplinary reviews. Data mining and knowledge discovery,* vol. 9, no. 3, pp. e1301-n/a, 2019, doi: 10.1002/widm.1301.

[259] S. Bernard, L. Heutte, and S. Adam, "Influence of Hyperparameters on Random Forest Accuracy," in *Multiple Classifier Systems*, Berlin, Heidelberg, J. A. Benediktsson, J. Kittler, and F. Roli, Eds., 2009// 2009: Springer Berlin Heidelberg, pp. 171-180.

[260] G. Martínez-Muñoz and A. Suárez, "Out-of-bag estimation of the optimal sample size in bagging," *Pattern Recognition,* vol. 43, no. 1, pp. 143-152, 2010/01/01/ 2010, doi: 10.1016/j.patcog.2009.05.010.

[261] S. Janitza, H. Binder, and A. L. Boulesteix, "Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications," (in eng), *Biom J,* vol. 58, no. 3, pp. 447-73, May 2016, doi: 10.1002/bimj.201400246.

[262] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," (in eng), *BMC Bioinformatics,* vol. 8, p. 25, Jan 25 2007, doi: 10.1186/1471-2105-8-25.

[263] M. Segal, "Machine Learning Benchmarks and Random Forest Regression," *Technical Report, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco,* 05/14 2003.

[264] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How Many Trees in a Random Forest?," in *Machine Learning and Data Mining in Pattern Recognition*, Berlin, Heidelberg, P. Perner, Ed., 2012// 2012: Springer Berlin Heidelberg, pp. 154-168.
[265] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest?," *Journal of Machine Learning Research,* vol. 18, 05/16 2017, doi: 10.48550/arXiv.1705.05654.
[266] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016.
[267] A. Moore and M. Bell, "XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study," (in eng), *Clin Med Insights Cardiol,* vol. 16, p. 11795468221133611, 2022, doi: 10.1177/11795468221133611.
[268] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling Up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2012.
[269] S. Tyree, K. Weinberger, K. Agrawal, and J. Paykin, "Parallel Boosted Regression Trees for Web Search Ranking," presented at the Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011.
[270] P. Li, C. J. C. Burges, and Q. Wu, "McRank: learning to rank using multiple classification and gradient boosting," presented at the Proceedings of the 21st International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2007.
[271] M. Feurer and F. Hutter, "Hyperparameter Optimization," in *Automated Machine Learning: Methods, Systems, Challenges*, F. Hutter, L. Kotthoff, and J. Vanschoren Eds. Cham: Springer International Publishing, 2019, pp. 3-33.
[272] M. Zhu *et al.*, "Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data," *IEEE Access,* vol. 6, pp. 4641-4652, 2018, doi: 10.1109/ACCESS.2018.2789428.
[273] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 1-3 March 2018 2018, pp. 1-11, doi: 10.1109/ICCTCT.2018.8551020.
[274] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res. (JAIR),* vol. 16, pp. 321-357, 06/01 2002, doi: 10.1613/jair.953.
[275] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics,* vol. 14, no. 1, p. 106, 2013/03/22 2013, doi: 10.1186/1471-2105-14-106.
[276] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," *Computing and Informatics,* vol. 34, pp. 1017-1037, 01/01 2015.
[277] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.
[278] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Chapter 5 - Diagnosing of disease using machine learning," in *Machine Learning and the Internet of Medical Things in Healthcare*, K. K. Singh, M. Elhoseny, A. Singh, and A. A. Elngar Eds.: Academic Press, 2021, pp. 89-111.
[279] F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians," (in eng), *Korean J Anesthesiol,* vol. 75, no. 1, pp. 25-36, Feb 2022, doi: 10.4097/kja.21209.

[280] T. T. Wong and P. Y. Yeh, "Reliable Accuracy Estimates from k-Fold Cross Validation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 32, no. 8, pp. 1586-1594, 2020, doi: 10.1109/TKDE.2019.2912815.
[281] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Mach. Learn. Res.,* vol. 7, pp. 1-30, 2006.
[282] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin,* vol. 1, no. 6, pp. 80-83, 1945, doi: 10.2307/3001968.
[283] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017.
[284] S. Lundberg. "Welcome to the SHAP documentation." https://shap.readthedocs.io/en/latest/index.html (accessed OCT, 2025).
[285] N. A. EISEMAN, M. B. WESTOVER, J. E. MIETUS, R. J. THOMAS, and M. T. BIANCHI, "Classification algorithms for predicting sleepiness and sleep apnea severity," *Journal of Sleep Research,* vol. 21, no. 1, pp. 101-112, 2012, doi: 10.1111/j.1365-2869.2011.00935.x.
[286] L. Taranto-Montemurro *et al.*, "The Combination of Atomoxetine and Oxybutynin Greatly Reduces Obstructive Sleep Apnea Severity. A Randomized, Placebo-controlled, Double-Blind Crossover Trial," (in eng), *Am J Respir Crit Care Med,* vol. 199, no. 10, pp. 1267-1276, May 15 2019, doi: 10.1164/rccm.201808-1493OC.
[287] S. Singh, S. Z. Khan, D. Singh, S. Verma, and A. Talwar, "The uses of overnight pulse oximetry," (in eng), *Lung India,* vol. 37, no. 2, pp. 151-157, Mar-Apr 2020, doi: 10.4103/lungindia.lungindia_302_19.
[288] A. Malhotra *et al.*, "Metrics of sleep apnea severity: beyond the apnea-hypopnea index," (in eng), *Sleep,* vol. 44, no. 7, Jul 9 2021, doi: 10.1093/sleep/zsab030.
[289] J. F. Garvey, C. T. Taylor, and W. T. McNicholas, "Cardiovascular disease in obstructive sleep apnoea syndrome: the role of intermittent hypoxia and inflammation," (in eng), *Eur Respir J,* vol. 33, no. 5, pp. 1195-205, May 2009, doi: 10.1183/09031936.00111208.
[290] D. R. Mazzotti *et al.*, "0593 Hypoxemia During Sleep Disordered Breathing and Cardiovascular Disease: A Comparison of Different Oxygen Desaturation Measures," *Sleep,* vol. 43, no. Supplement_1, pp. A227-A227, 2020, doi: 10.1093/sleep/zsaa056.590.
[291] N. Esmaeili *et al.*, "Hypoxic Burden Based on Automatically Identified Desaturations Is Associated with Adverse Health Outcomes," (in eng), *Ann Am Thorac Soc,* vol. 20, no. 11, pp. 1633-1641, Nov 2023, doi: 10.1513/AnnalsATS.202303-248OC.
[292] S. He, K. Cook, K. Sutherland, Y. S. Bin, P. A. Cistulli, and P. de Chazal, "A comparison of hypoxic burden algorithms using three different methods for calculating baseline oxygen saturation for predicting cardiovascular death in the Sleep Heart Health Study," (in eng), *Annu Int Conf IEEE Eng Med Biol Soc,* vol. 2023, pp. 1-4, Jul 2023, doi: 10.1109/embc40787.2023.10340410.
[293] D. A. Dean, 2nd *et al.*, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," (in eng), *Sleep,* vol. 39, no. 5, pp. 1151-64, May 1 2016, doi: 10.5665/sleep.5774.
[294] S. F. Quan *et al.*, "The Sleep Heart Health Study: design, rationale, and methods," (in eng), *Sleep,* vol. 20, no. 12, pp. 1077-85, Dec 1997.
[295] S. Redline *et al.*, "Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group," (in eng), *Sleep,* vol. 21, no. 7, pp. 759-67, Nov 1 1998.

[296] P. I. Terrill, C. Dakin, B. A. Edwards, S. J. Wilson, and J. E. MacLean, "A graphical method for comparing nocturnal oxygen saturation profiles in individuals and populations: Application to healthy infants and preterm neonates," *Pediatric Pulmonology,* vol. 53, no. 5, pp. 645-655, 2018, doi: 10.1002/ppul.23987.

[297] Y. M. Huang *et al.*, "Sleep duration and risk of cardio-cerebrovascular disease: A dose-response meta-analysis of cohort studies comprising 3.8 million participants," (in eng), *Front Cardiovasc Med,* vol. 9, p. 907990, 2022, doi: 10.3389/fcvm.2022.907990.

[298] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics,* vol. 2, p. 100060, 2022/11/01/ 2022, doi: 10.1016/j.health.2022.100060.

[299] J. Ingham and P. D. Macnaughton, "Measurement of pO2, pCO2, pH, pulse oximetry and capnography," *Anaesthesia & Intensive Care Medicine,* vol. 6, no. 12, pp. 413-415, 2005/12/01/ 2005, doi: 10.1383/anes.2005.6.12.413.

[300] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," presented at the Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, 2011.

[301] Z. Zhao and H. Liu, "Searching for interacting features in subset selection," *Intell. Data Anal.,* vol. 13, no. 2, pp. 207–228, 2009.

[302] L. o. Grinsztajn, E. Oyallon, and G. l. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," presented at the Proceedings of the 36th International Conference on Neural Information Processing Systems , articleno = 37 , numpages = 14, 2022.

[303] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion,* vol. 81, no. C, pp. 84–90 , numpages = 7, may 2022, doi: 10.1016/j.inffus.2021.11.011.

[304] P. Shah, M. Shukla, N. H. Dholakia, and H. Gupta, "Predicting cardiovascular risk with hybrid ensemble learning and explainable AI," *Scientific Reports,* vol. 15, no. 1, p. 17927, 2025/05/23 2025, doi: 10.1038/s41598-025-01650-7.

[305] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," (in eng), *Sci Rep,* vol. 14, no. 1, p. 23277, Oct 7 2024, doi: 10.1038/s41598-024-74656-2.

[306] S. Simon *et al.*, "The Impact of Time Horizon on Classification Accuracy: Application of Machine Learning to Prediction of Incident Coronary Heart Disease," *JMIR Cardio,* vol. 6, no. 2, 2022/01/01/ 2022, doi: 10.2196/38040.