INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Ensemble AI Solutions for Personalized Sleep Monitoring Using Wrist-worn Wearables

VASCO ANTÓNIO PORTILHO CARVALHO DA SILVA
Setembro de 2025

P.PORTO

Ensemble AI Solutions for Personalized Sleep Monitoring Using Wrist-worn Wearables

Vasco António Portilho Carvalho da Silva
Student No.: 1200750

Dissertation for the degree of
Master in Artificial Intelligence Engineering

Supervisor: Luís Manuel Silva Conceição, Assistant Professor, Institute of Engineering, Polytechnic of Porto

Evaluation Committee:

President:

Carlos Fernando da Silva Ramos, Full Professor, Institute of Engineering, Polytechnic of Porto

Members:

Rita da Silva Amaral, Assistant Researcher, Faculty of Medicine, University of Porto
Luís Manuel Silva Conceição, Assistant Professor, Institute of Engineering, Polytechnic of Porto

Porto, September 30, 2025

Abstract

Sleep disorders, including insomnia and sleep apnoea, affect a significant proportion of the global population and are closely linked to cardiovascular, metabolic, and mental health conditions. Accurate and long-term monitoring of sleep is therefore a public health priority, as early detection and personalised management can substantially improve quality of life and reduce healthcare costs.

This dissertation explores how wrist-worn wearable devices, combined with advanced machine learning and explainable artificial intelligence (XAI) techniques, can enhance the monitoring and analysis of sleep. While polysomnography (PSG) remains the clinical gold standard for sleep assessment, its cost, intrusiveness, and limited scalability restrict its long-term and widespread applicability. To address these limitations, this work proposes an integrated framework that leverages multimodal data, including photoplethysmography (PPG) and accelerometry, for automatic sleep stage classification and the detection of sleep apnoea.

The system incorporates ensemble machine learning models to generate high-quality, personalised insights into sleep quality. Furthermore, explainability is ensured through the application of XAI methods, namely SHAP and LIME, enabling healthcare professionals and end-users to understand and trust model predictions. Experimental validation was conducted using multiple publicly available datasets, demonstrating the system's robustness and generalisability across heterogeneous populations.

Ultimately, this research contributes to the development of transparent, non-invasive, and scalable sleep monitoring solutions. It lays the groundwork for real-world applications in personalised healthcare and the early detection of sleep disorders, promoting better clinical decision-making and long-term well-being.

Keywords: Artificial Intelligence, Machine Learning, Wearable Technology, Sleep Monitoring, Sleep Apnoea Detection, Explainable AI

Resumo

Os distúrbios do sono, incluindo a insónia e a apneia do sono, afetam uma parte significativa da população mundial e estão intimamente associados a condições cardiovasculares, metabólicas e de saúde mental. A monitorização precisa e de longo prazo do sono é, por isso, uma prioridade de saúde pública, uma vez que a deteção precoce e a gestão personalizada podem melhorar substancialmente a qualidade de vida e reduzir os custos dos cuidados de saúde.

Esta dissertação explora de que forma os dispositivos vestíveis de pulso, em combinação com técnicas avançadas de aprendizagem automática e inteligência artificial explicável (XAI), podem melhorar a monitorização e análise do sono. Embora a polissonografia (PSG) continue a ser o padrão clínico de referência para a avaliação do sono, o seu custo, carácter intrusivo e limitada escalabilidade restringem a sua aplicabilidade em contextos de longo prazo e em larga escala. Para colmatar estas limitações, este trabalho propõe um sistema integrado que utiliza dados multimodais, nomeadamente fotopletismografia (PPG) e acelerometria, para a classificação automática das fases do sono e deteção de apneia do sono.

O sistema recorre a modelos de aprendizagem automática em ensemble para gerar informações personalizadas e de elevada qualidade sobre a qualidade do sono. A explicabilidade é assegurada através da aplicação de métodos XAI, como o SHAP e o LIME, permitindo que profissionais de saúde e utilizadores compreendam e confiem nas previsões geradas pelos modelos. A validação experimental foi realizada com múltiplos conjuntos de dados públicos, demonstrando a robustez e a capacidade de generalização do sistema em populações heterogéneas.

Em última análise, esta investigação contribui para o desenvolvimento de soluções de monitorização do sono transparentes, não invasivas e escaláveis. Estabelece as bases para aplicações reais no contexto da saúde personalizada e na deteção precoce de distúrbios do sono, promovendo uma melhor tomada de decisão clínica e o bem-estar a longo prazo.

Palavras-chave: Artificial Intelligence, Machine Learning, Wearable Technology, Sleep Monitoring, Sleep Apnoea Detection, Explainable AI

Acknowledgement

The completion of this dissertation would not have been possible without the support, encouragement, and presence of many people and institutions, to whom I am deeply indebted and would like to express my heartfelt gratitude.

First and foremost, I would like to thank GECAD (Research Group on Intelligent Engineering and Decision Support), the research centre where I had the privilege to carry out this work. The opportunity to be part of such a stimulating environment, surrounded by knowledge, innovation, and collaboration, was fundamental to the success of this journey. I am equally grateful to ISEP (Instituto Superior de Engenharia do Porto), the institution that guided my academic path and provided me with the tools, education, and values that shaped not only this dissertation but also my personal and professional growth.

To my parents, whose unwavering dedication, encouragement, and unconditional support have been the foundation of every step I have taken. Their sacrifices, love, and belief in me have always been my greatest source of strength. To my girlfriend, for her companionship, patience, and constant motivation during the most demanding and uncertain moments of this journey, and for reminding me of the importance of balance and perspective.

To my family and friends, for their understanding, genuine friendship, and for always standing by me through good and difficult times. The countless conversations, words of encouragement, and moments of laughter provided the relief and balance needed to overcome challenges and to continue moving forward with determination.

To my colleagues at GECAD, whose collaborative spirit, knowledge sharing, and constant support enriched not only this research but also my experience as a researcher. The sense of community, teamwork, and openness made this endeavour both productive and truly enjoyable.

A very special thanks to my supervisor, Professor Luís Conceição, for his invaluable guidance, availability, patience, and insightful advice throughout the development of this dissertation. His trust, encouragement, and expertise provided the clarity and direction I needed at every stage, and for that I am sincerely grateful.

Finally, to all those who, directly or indirectly, contributed to this work, whether through words of encouragement, critical feedback, or simply by being present during this demanding but rewarding journey, I extend my deepest and most sincere gratitude.

This research work was developed under the project REMO (ITEA-2023-23005-REMO), funded by the European Regional Development Fund (ERDF) within the project number COMPETE2030-FEDER-01233000, and funded by National Funds through the Portuguese FCT—Fundação para a Ciência e a Tecnologia under the R&D Units Project Scope, UIDB/00760/2020 (https://doi.org/10.54499/UIDB/00760/2020).

Contents

List of Figures

List of Tables

List of Acronyms

AHI	Apnoea-Hypopnoea Index.
AI	Artificial Intelligence.
API	Application Programming Interface.
CNN	Convolutional Neural Network.
DL	Deep Learning.
ECG	Electrocardiogram.
F_1	F_1 Score.
GDPR	General Data Protection Regulation.
HR	Heart Rate.
HRV	Heart Rate Variability.
IBI	Inter-Beat Interval.
KNN	k-Nearest Neighbors.
LightGBM	Light Gradient Boosting Machine.
LSTM	Long Short-Term Memory.
MSDA-1DCNN	Multiscale Deep Neural Network - 1D Convolutional Neural Network.
MTL	Multitask Learning.
NREM	Non-Rapid Eye Movement.
OSA	Obstructive Sleep Apnoea.
PPG	Photoplethysmography.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
PSG	Polysomnography.
Q1	Quartile 1 (journal impact ranking).
Q2	Quartile 2 (journal impact ranking).
REM	Rapid Eye Movement.

REMO	Remote Monitoring for Active and Independent Ageing.
ROC	Receiver Operating Characteristic.
SMOTE	Synthetic Minority Over-sampling Technique.
SpO_2	Peripheral Oxygen Saturation.
SVM	Support Vector Machine.
TCN	Temporal Convolutional Network.
XAI	Explainable Artificial Intelligence.
XGBoost	Extreme Gradient Boosting.

Chapter 1

Introduction

This chapter lays the foundation for the dissertation by presenting the motivation, scope, and expected contributions of the research. It is structured into five main sections, each designed to progressively guide the reader from the broader context to the specific objectives of the work.

The first section, Context, discusses the growing role of wearable devices in health monitoring, with an emphasis on their application to sleep analysis. The second section, Problem Statement, defines the key challenges that underpin this research, including the limitations of current approaches and the gaps that remain in the state of the art. The third section, Objectives and Research Questions, outlines the overall aim of the thesis, breaking it down into specific objectives and formulating the central research question. The fourth section, Contributions, highlights the main scientific, methodological, and practical outcomes expected from this work. Finally, the section Document Structure explains how the remainder of the dissertation is organised, providing a roadmap for the chapters that follow.

1.1 Context

In recent years, wearable devices have emerged as powerful tools for monitoring health and well-being. Among these, wrist-worn wearables, such as smartwatches and fitness bands, are particularly notable for their continuous monitoring of daily activities, offering a non-invasive alternative for personalised health monitoring. Thanks to advances in sensing technologies and artificial intelligence (AI), these devices enable increasingly accurate and accessible analyses, transforming our understanding and management of essential factors for quality of life (Leclercq et al. 2022).

Sleep is a critical pillar of both physical and mental health, playing a central role in processes like physiological recovery, memory consolidation, mood regulation, and metabolic functioning. However, sleep disorders, such as insomnia and sleep apnoea, impose a significant socio-economic burden, increasing healthcare costs and reducing workplace productivity. Insomnia leads to absenteeism and presenteeism, while sleep apnoea is linked to long-term health issues like cardiovascular diseases (Skaer and Sclar 2010). While traditional sleep monitoring methods, such as polysomnography, are considered the gold standard, they have several limitations, such as high costs, invasiveness, and impracticality for large-scale and everyday applications (Borazio et al. 2014).

In order to better understand the impact and distribution of sleep disorders, recent studies provide detailed prevalence data, as summarized in Table 1.1.

Table 1.1: Prevalence of sleep disorders based on recent studies.

Sleep Disorder	Prevalence	Notes	Source
Insomnia	7.4%-15.8%	Higher in older adults and associated with psychological stress.	(Fund et al. 2019),(McArdle et al. 2022)
Sleep-disordered breathing (SDB)	24%-47.3%	Higher prevalence in men and older adults; linked to obesity and metabolic syndromes.	(McArdle et al. 2022)
Restless Legs Syndrome (RLS)	2.2%-3.7%	Reported as more common in middle-aged women and associated with distress.	(McArdle et al. 2022)
Excessive Daytime Sleepiness	13.4%-31%	Frequently reported among students and older adults; affects productivity and quality of life.	(Ahn et al. 2023), (Jahrami et al. 2019)
Obstructive Sleep Apnoea	14.41%-47.3%	Increasing prevalence globally due to rising obesity rates.	(Ahn et al. 2023), (McArdle et al. 2022)
Parasomnias (e.g., Nightmares)	12%-25.7%	Higher prevalence in younger populations; influenced by stress and environmental factors.	(Hua, Lyu, and Du 2022)

Wrist-worn wearables present an innovative solution to address these challenges. They facilitate continuous and real-time monitoring, collecting data that improve our understanding of factors influencing sleep quality. When integrated with machine learning algorithms and AI, these devices can provide a holistic view of sleep patterns and habits, allowing not only monitoring but also the identification of trends and offering personalised interventions to improve health and well-being (Roberts et al. 2020).

While the potential impact of wrist-worn wearables is widely acknowledged, significant gaps still exist. Many studies tend to focus solely on classifying sleep stages or measuring isolated parameters, often overlooking the need for more integrated analyses that consider multiple factors. In addition, there are ongoing challenges related to the precision and generalisability of analytical models and the adaptation of these solutions to diverse populations and contexts (S. Chen et al. 2015; Robbins et al. 2024).

Given the increasing prevalence of sleep-related issues and their association with various aspects of health, it is imperative to study the role of wrist-worn devices in sleep monitoring and quality improvement. These devices not only enable more accessible diagnostics and personalised interventions, but also present a unique opportunity to deepen our understanding of how daily habits, behavioural factors, and health conditions interact with sleep (Song, Chowdhury, et al. 2023).

1.2 Problem Statement

Despite significant advancements in wearable technologies and their integration with artificial intelligence (AI), there are still notable challenges in utilizing wrist-worn devices to derive

meaningful and actionable insights regarding sleep quality. Current methodologies mainly focus on classifying sleep stages or measuring isolated parameters, frequently overlooking the multifactorial nature of sleep, which is influenced by a complex interplay of daily habits, behavioural patterns, and underlying health conditions (Hamza et al. 2023).

While wrist-worn wearables offer an accessible and non-invasive alternative to traditional sleep monitoring methods, they still face limitations concerning accuracy and reliability (Mantua, Gravel, and Spencer 2016; Robbins et al. 2024). Traditional techniques remain widely regarded for precision in sleep analysis, delivering detailed and robust insights. However, their resource-intensive nature and the need for controlled environments render them impractical for long-term or large-scale applications (Y. J. Lee et al. 2024). This highlights an urgent need for wearables to not only complement traditional methods but also tackle the challenges of scalability and accessibility.

Furthermore, existing analytical models employed in wrist-worn wearables often struggle to integrate multiple data streams and generalize across diverse populations effectively. Many studies tend to focus on isolated metrics, such as heart rate or movement, thereby ignoring the broader interplay of factors that impact sleep quality. This oversight limits the potential of wearables to deliver actionable, high-quality insights capable of informing personalized interventions (Hamza et al. 2023; Moreno-Pino et al. 2022).

These challenges are exemplified by recent studies. For instance, an evaluation of the Fitbit Charge 4 demonstrated limitations in measuring sleep stages, particularly in accurately identifying deep sleep and REM sleep. While the device performed adequately in estimating total sleep time, its specificity for certain stages remains low, highlighting gaps in the precision of wearable sleep metrics (Schyvens et al. 2023). Similarly, a study leveraging deep learning with reflective photoplethysmography (rPPG) for the apnoea–hypopnea index (AHI) estimation achieved moderate correlation with gold-standard polysomnography ($r = 0.61$) and limited diagnostic reliability (Cohen's kappa = 0.51). This underscores the challenges in integrating and analysing complex physiological data for robust insights (Papini et al. 2020).

Another significant challenge is adapting these wearable solutions to various populations and contexts. Differences in lifestyle, health conditions, and environmental factors can influence the reliability and applicability of generated insights. While traditional methods excel in diagnostic precision, wearable technologies present an untapped opportunity to bridge this gap by providing accessible solutions tailored to real-world scenarios (Ojalvo, Pacheco, and Benedict 2023).

The present work aims to address these challenges by developing a robust framework to extract high-quality, integrated insights into sleep quality through wrist-worn wearables. By leveraging advanced machine learning models (M. P. Lee et al. 2024), integrating diverse data streams (Moreno-Pino et al. 2022), and incorporating contextual factors, this research seeks to merge the strengths of traditional precision with the accessibility offered by wearables. Improving the accuracy and integration of wearable data not only enhances sleep diagnosis but also enables more personalised, timely interventions. By combining multiple metrics it becomes possible to identify early signs of sleep disorders and offer personalized advice, leading to better management of conditions such as insomnia or sleep apnoea (Wongtaweesup et al. 2023). The ultimate goal is to enhance sleep monitoring and generate actionable insights that can guide personalized interventions, thereby promoting healthier sleep habits and improving overall well-being (Jenefa et al. 2023).

This thesis is carried out under the Remote Monitoring for Active and Independent Ageing (REMO) (Remote Monitoring for Active and Independent Ageing) project, which develops unobtrusive, continuous monitoring and personalized coaching solutions for older adults. It should be noted that this work represents an initial phase of the REMO project. A subsequent stage will involve the collection of real-world data from users in home settings. This will allow the proposed solution to be evaluated under more heterogeneous conditions, thereby enhancing its external validity and practical applicability.

1.3 Objectives and Research Questions

The main goal of this thesis is to develop a framework for leveraging wrist-worn wearable data and ensemble Artificial Intelligence (AI) techniques to support personalized sleep monitoring. With this broader goal in mind, the following specific objectives were established:

- **OB1:** Investigate the state-of-the-art methods for utilizing wrist-worn wearable device data in sleep monitoring, including data collection techniques, processing methodologies, and analytical frameworks.
- **OB2:** Identify and characterize specific sleep patterns, behaviours, and influencing factors that can be derived from wrist-worn wearable data.
- **OB3:** Develop and apply ensemble AI models capable of integrating multiple data streams from wrist-worn wearable devices to generate high-quality, personalized insights into sleep quality.
- **OB4:** Explore using explainable artificial intelligence (Explainable Artificial Intelligence (XAI)) techniques to interpret wrist-worn wearable device data, ensuring transparency and trustworthiness in the analysis.
- **OB5:** Validate and evaluate the developed framework using real-world wrist-worn wearable datasets, demonstrating its practical applicability and adaptability across diverse populations and contexts.

To guide the research conducted within the scope of this thesis and successfully accomplish the established objectives, the main research question was carefully formulated: *'How can wearable data and ensemble AI techniques support personalised sleep monitoring?'*.

1.4 Contributions

This thesis advances wrist-based, AI-supported sleep monitoring along four main axes:

1. **End-to-End Wearable Framework**
   A complete pipeline is designed that converts raw photoplethysmography (PPG) and tri-axial accelerometry into five-class sleep staging and multi-class apnoea detection, ready for deployment on commodity smart-bands.
2. **Interpretable Ensemble-Learning Models**
   Lightweight, tree-based ensembles (Random Forest, LightGBM, AdaBoost) are engineered and optimised to outperform baseline methods while remaining fast enough for edge inference. Integrated SHAP and LIME analyses expose the physiological basis of each decision, fostering clinical trust.

3. **Ethical-Regulatory Blueprint**
   A practical compliance strategy is articulated for high-risk healthcare AI under the European AI Act and GDPR, covering data minimisation, bias auditing, continual monitoring, and human-in-the-loop oversight.
4. **Open Sharing**
   All code, trained models, and example notebooks will be released openly.
5. **Dissemination**
   Results have already been presented in the peer-reviewed conference paper *"Lightweight Tree Ensembles with Optimized Features for Five-Class Sleep Apnea Stratification"* (EPIA 2025). The author has also co-authored of the forthcoming journal article *"Comprehensive Analysis of Machine Learning Models for Five-Class Sleep Stage Classification Using PPG Signals"*. Collectively, these contributions deliver an accurate, transparent, and regulation-ready foundation for large-scale, personalised sleep/health applications.

1.5 Document Structure

This dissertation is organised into **six chapters** and **two appendices**. The sequence is designed to lead the reader from the initial motivation through to the main findings and future work:

1. **Chapter 1 – Introduction**
   Presents the research motivation, problem statement, objectives, research question, and the expected contributions of the thesis.
2. **Chapter 2 – State of the Art**
   Provides a critical review of recent literature on wrist-based sleep monitoring, covering data-collection methods, signal-processing techniques, machine-learning models, and current research gaps.
3. **Chapter 3 – Methods, Materials, and Technological Challenges**
   Details the end-to-end pipeline developed in this work, including signal preprocessing, feature extraction, data preparation, model selection and optimisation, and the explainability (XAI) framework. The chapter also describes the architecture of the integrated backend–frontend system.
4. **Chapter 4 – Results and Ethical-Social Considerations**
   Reports the experimental results for (i) sleep-stage classification and (ii) multi-class detection of sleep-disordered breathing events. Global metrics, per-class analyses, confusion matrices, ROC curves, and both global and local XAI explanations are discussed. Additionally, examines the ethical, legal, and societal implications of using AI in healthcare.
5. **Chapter 5 – Conclusions**
   Summarises the main achievements, evaluates the extent to which the research objectives were met, and outlines directions for future work, such as large-scale validation, additional sensor integration, and on-device deployment.

Chapter 2

State-of-the-art

This chapter explores the state of the art in the fields relevant to this research, providing a robust foundation for understanding the technological and methodological context. It is divided into two sections, each addressing critical aspects necessary for developing this study.

The first section, Technological Foundations, delves into the core technologies, frameworks, and methodologies that underpin this research. It provides an overview of the essential tools and innovations driving progress in the field and establishes the groundwork for the proposed approaches.

The second section focuses on a systematic review of the literature, conducted using the PRISMA framework. This rigorous methodology ensures a transparent and unbiased evaluation of existing studies. The section details the inclusion and exclusion criteria, the quality assessment process, and the synthesis of findings to address key research questions.

Together, these sections provide a comprehensive overview of the field's current advancements, challenges, and gaps, offering insights that inform the direction and contributions of this study.

2.1 Technological Foundations

This section delves into the technological frameworks and tools underpinning this research. It highlights the essential concepts, methods, and platforms that enable the implementation of the proposed methodologies. The discussion focuses on the core technologies driving innovation in this domain and their practical implications.

2.1.1 Introduction to Wearable Technology

Wearable technology refers to intelligent electronic devices that can be worn on the body as accessories or integrated into clothing. These devices are designed to collect data, track activities, and provide real-time information to users (Lima 2023). Wearable technology has evolved rapidly in recent years, transitioning from simple pedometers to highly sophisticated multifunctional devices. Currently, a wide range of products is available, from smartwatches and smart glasses to clothing with embedded sensors (Cantanhede et al. 2018).

The benefits of wearable technology are numerous and span various domains. In healthcare, wearable devices enable continuous monitoring of vital signs, early detection of complications, and remote patient follow-up. For instance, smartwatches can detect cardiac arrhythmias and send immediate alerts in cases of risk (Chirieleison et al. 2024). In fitness and wellness, wearables like fitness bands and smartwatches help users track physical activity,

sleep, and other health habits, promoting a more active and healthy lifestyle (Hedrick et al. 2020). Additionally, wearable technology is revolutionizing diagnostic medicine by enabling continuous collection of physiological data, which can be analysed to identify patterns and predict potential health issues (Moraes et al. 2023). In rehabilitation, wearable devices are being used to monitor and stimulate specific exercises, as demonstrated in studies involving sensor-controlled games embedded in sandals for foot rehabilitation (De Assis et al. 2024). In medical research, wearables provide valuable data, enabling broader and more longitudinal studies on health and human behaviour (Gomes, Montanini, and Rocha Sobrinho 2024).

The relevance of wearable technology continues to grow, with applications expanding beyond health and fitness into areas such as workplace safety, entertainment, and social interaction. As this technology becomes more sophisticated and accessible, its impact on daily life and various industries is expected to increase significantly(Cantanhede et al. 2018).

Wearable technology encompasses a variety of devices designed to be worn on the body, some of which are represented in Figure 2.1. Key categories include headwear, such as

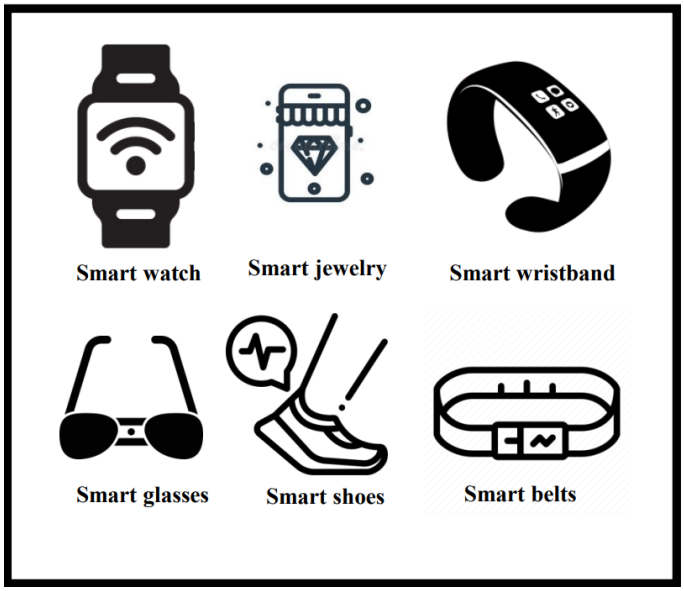

Figure 2.1: IoT-based Wearable products; Image from (Bello and Figetakis 2023).

smart glasses and AR/VR headsets, which provide immersive visual experiences or overlay information onto the real environment. Smart clothing integrates sensors to monitor vital signs, posture, and even regulate body temperature, as seen in shirts that measure heart rate(Tat et al. 2022). Smart jewellery, like rings and bracelets, tracks activities, monitors sleep, and supports contactless payments. Smart footwear includes shoes and insoles with embedded sensors for gait analysis, step counting, and energy generation through movement (Mousavi et al. 2023). Among wearables, wrist-worn devices are the most popular, offering a balance of functionality, convenience, and style. Smartwatches, like the Apple Watch and Samsung Galaxy Watch, provide notifications, health monitoring, GPS, and contactless payments, combining versatility and an extensive range of apps. Fitness bands, such as Fitbit and Xiaomi Mi Band, focus on activity tracking, sleep monitoring, and heart rate measurement, with minimalist designs and long battery life. Hybrid devices, like the Garmin Vivomove, blend traditional watch aesthetics with discreet smart features. Additionally,

health-focused devices offer advanced monitoring capabilities, including heart rate variability and body temperature, delivering more precise and comprehensive health data.

2.1.2 Sensors in Wearables

Wearable devices have revolutionized personal health monitoring by incorporating a variety of sensors that collect real-time data on various physiological parameters. These sensors play a crucial role in tracking health, activity, and well-being, providing users with valuable insights into their daily lives and overall health status.

Accelerometers and gyroscopes are fundamental components of many wearable devices, working in tandem to provide comprehensive motion tracking. Accelerometers measure linear acceleration and changes in position, while gyroscopes complement this data by measuring orientation and rotation. These sensors are essential for tracking physical activities such as step counting and exercise monitoring, analysing gait and body posture, and enhancing augmented reality applications (Calisti and Lattanzi 2024),(R. T. Li et al. 2016).

Heart rate monitoring has become a standard feature in many smartwatches and fitness trackers. These devices typically use optical heart rate sensors that continuously monitor a user's heart rate, providing real-time data on cardiovascular health. Some advanced wearables even offer electrocardiogram (ECG) capabilities for more detailed heart rhythm analysis (Dhar, Kumar, and Karmakar 2023).

Body temperature monitoring is another important feature of many wearable devices. The MLX90614 sensor, for example, can be used to measure body temperature when positioned on the wrist. This capability is particularly useful for detecting fevers, monitoring temperature changes during physical exercise, and assisting in early detection of potential health issues (Anggraini et al. 2023).

Blood oxygen saturation (SpO2) monitoring has gained significant attention, especially in recent years. Wearable devices often incorporate sensors like the MAX30100 to measure SpO2 levels. This feature is crucial for monitoring respiratory health, detecting potential breathing problems, and assisting in the management of conditions like sleep apnoea (Anggraini et al. 2023).

Many wearables include GPS functionality, which is particularly useful for outdoor activities. GPS sensors allow users to track location during runs, walks, or cycling sessions, measure distance travelled and pace, and enhance safety during outdoor workouts.

The field of wearable technology is rapidly evolving, with new types of sensors being integrated to expand monitoring capabilities. Electrodermal activity (EDA) sensors measure skin conductivity and are often used for stress monitoring and emotional state assessment. Blood pressure sensors are starting to appear in advanced smartwatches. Additionally, cutting-edge wearables are beginning to include biochemical sensors capable of measuring biomarkers like glucose, lactate, and cortisol, opening up new possibilities for continuous health monitoring (Dhar, Kumar, and Karmakar 2023), (Moses 2024).

The true power of these sensors lies in their ability to collect and analyse data in real-time. Modern wearables use sophisticated algorithms and artificial intelligence to process sensor data, providing users with actionable insights. This real-time processing enables immediate alerts for abnormal health readings, continuous tracking of physical activity and fitness progress, and early detection of potential health issues (Nur 2024).

A summary of the key sensors commonly found in wearable devices, along with their functions and benefits, is provided in Table 2.1.

Table 2.1: Sensors in Wearable Devices.

Sensor	**Function**	**Example**
Accelerometer	Measures acceleration and movement	Steps, exercise
Gyroscope	Measures rotation and orientation	Augmented reality (AR)
Heart Rate	Monitors heartbeats	Heart rate tracking
Temperature	Measures body temperature	Fever monitoring
SpO2	Measures blood oxygen saturation	Respiratory monitoring
GPS	Tracks location	Running, walking
EDA	Measures skin conductivity	Stress, emotions
Blood Pressure	Measures blood pressure	Advanced devices
Biochemical	Measures biomarkers (glucose, etc.)	Continuous monitoring

While wearable sensors have made significant strides, there are still challenges to overcome. Ensuring accuracy and reliability across various activities and conditions remains a priority. Data privacy and security are crucial, as these devices collect sensitive health information. Balancing sensor performance with energy efficiency is also key for long-term usability. Future developments in wearable sensor technology are focusing on integrating more advanced biochemical sensors for comprehensive health monitoring, developing non-invasive, flexible sensors that can provide accurate, continuous health information without disrupting daily activities, and enhancing AI and machine learning algorithms to improve data interpretation and predictive capabilities (Nur 2024), (Haval and Afzal 2024), (Moses 2024).

In conclusion, sensors in wearable devices are at the forefront of a healthcare revolution, enabling personalized, real-time health monitoring. As technology continues to advance, these devices are poised to play an increasingly important role in preventive healthcare and chronic disease management.

2.1.3 Artificial Intelligence Applied to Wearable Devices

The integration of Artificial Intelligence (AI) in wearable devices has revolutionized the way health and wellness data are collected and analysed. Technologies such as Machine Learning (ML) and Deep Learning (Deep Learning (DL)) play a crucial role in interpreting complex data from sensors, enabling more accurate and personalized insights.

Machine Learning (ML) refers to the ability of systems to learn and improve from data without relying on explicit programming. In wearable devices, ML is fundamental for applications such as detecting physical activities, monitoring sleep patterns, and predicting health risks. Popular techniques include Decision Trees, Support Vector Machines (SVM), and Random Forests. A notable example of ML application is the use of algorithms to detect daily activities and identify nearby objects. Proximity and sound sensors, for instance, have been used in wearable devices to monitor social distancing compliance during infectious disease outbreaks (Umutoni et al. 2023).

Deep Learning (DL), a subfield of ML, leverages artificial neural networks with multiple layers to learn complex data representations. In wearable devices, DL is applied to tasks such as advanced ECG signal analysis, recognition of complex movement patterns, and detection of cardiac anomalies. Frequently used models include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM). A recent study demonstrated the use of a combined CNN-LSTM architecture for Human Activity Recognition (HAR), significantly improving activity classification accuracy (C. Liu et al. 2024).

Explainable AI (XAI) is an approach aimed at making AI models more transparent and understandable, which is particularly critical in healthcare applications. XAI plays an essential role in increasing trust among users and healthcare professionals, facilitating the interpretation of complex results, and helping to identify potential biases in the models. A recent study explored the use of XAI techniques to analyse factors that significantly influence ML models in HAR, clearly demonstrating these impacts (C. Liu et al. 2024).

Despite advancements, there are significant challenges. Issues such as data privacy and security are increasingly critical as the use of personal information grows (Brutzman et al. 2024). Additionally, energy efficiency is a priority, requiring AI algorithms that minimize energy consumption to extend wearable device battery life (Bhat and Raychowdhury 2023). Another major challenge is personalization, as there is a growing need for AI models tailored to each user's unique characteristics (Kargarandehkordi, Slade, and Washington 2024). The integration of multiple sensors is also essential, allowing for more comprehensive and accurate analyses by combining data from various types of sensors (Brutzman et al. 2024). Enhancing real-time processing capabilities to provide immediate feedback and investing in clinical validation are necessary to ensure algorithms' efficacy and accuracy in real-world scenarios (Olawade et al. 2024).

With the continuous advancement of technology, the integration of AI into wearable devices promises to significantly transform health and wellness monitoring. These devices are expected to become increasingly seamless and integrated into daily life, offering deeper, more personalized, and impactful insights into users' health (Iyer, Gejji, and Pandya 2022).

2.2 Systematic review

This section provides a comprehensive exploration of the systematic review conducted to evaluate the current state-of-the-art in AI-supported wearable technologies for sleep analysis. It aims to establish the foundational knowledge required to address the research questions outlined in the study.

The section begins by outlining the systematic methodology employed, including adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, ensuring a rigorous and unbiased selection of pertinent literature. This is followed by a detailed description of the primary data sources utilised, emphasising their multidisciplinary scope and relevance. A clear set of inclusion and exclusion criteria is then presented, ensuring that the selected studies align with the objectives and scope of this review. All included studies underwent a thorough quality assessment to ensure scientific rigour, highlighting their methodological robustness and significance. The insights gained from the review are synthesised to address key aspects, such as data types, methodologies, and applications related to the research focus.

This structured approach lays the groundwork for the proposed study and identifies significant gaps and opportunities for future research in the field.

2.2.1 Research Methodology

A systematic review is a rigorous method for evaluating and synthesising all available research related to a specific research question, topic, or area of interest (T. Li, Saldanha, and Robinson 2022). This review adheres to the PRISMA guidelines (PRISMA Group 2024), with necessary adaptations to ensure a thorough examination of current research and to inform the development of innovative methods for sleep monitoring using wearable devices (Page et al. 2021). Following these guidelines, the review will proceed through the following steps: Conduct a systematic search of relevant studies detailing the data sources and search strategies used, including keywords. Screen and select studies based on predefined inclusion and exclusion criteria. Assess the quality of selected studies using standardised quality evaluation frameworks. Extract and analyse data systematically, focusing on key outcomes, methodologies, and findings. Finally, the evidence will be synthesised, the results will be interpreted, and the findings will be presented to address the research questions and comprehensively identify future research opportunities.

2.2.2 Research Questions

To conduct the systematic review, the following research questions were defined to guide the analysis of data and techniques applied to wearable devices in the context of sleep assessment and monitoring:

- **RQ1:** "What data types from wrist-worn wearable devices are most frequently utilized to assess and evaluate sleep?"
- **RQ2:** "What specific sleep patterns and behaviours can be inferred from data collected by wrist-worn wearable devices?"
- **RQ3:** "What key factors influencing sleep are identified and quantified through wrist-worn wearable device data?"
- **RQ4:** "How can explainable artificial intelligence (XAI) techniques be effectively applied to analyse and interpret wrist-worn wearable sleep data?"

These questions aim to guide the systematic review in identifying the most relevant approaches, methods, and technologies in the field of wearable devices applied to sleep analysis.

2.2.3 Data Sources

The data sources for this review were selected from four major academic databases: PubMed, Web of Science, IEEE Xplore, and Science Direct. These databases were chosen for their comprehensive coverage, relevance to the study objectives, and accessibility to high-quality research across multidisciplinary fields. Table 2.2 presents a brief description of each database.

PubMed was selected for its extensive collection of biomedical literature. As this review focuses on sleep analysis and wearable devices, PubMed provided access to studies addressing

health-related aspects, including physiological and clinical insights critical to understanding sleep patterns and disorders.

Web of Science was included due to its multidisciplinary nature. It offers a wide range of peer-reviewed articles across various fields. This database was particularly valuable for exploring the intersections between technology, health sciences, and behavioural studies relevant to wearable-based sleep research.

IEEE Xplore was chosen for its focus on engineering, computer science, and electronics. Given the technological nature of wearable devices and the increasing role of machine learning and AI in sleep analysis, IEEE Xplore offered access to cutting-edge research in these domains.

Science Direct was included for its robust collection of scientific and technical articles, particularly in the physical and health sciences. This database complemented the others by providing insights into the technical aspects of wearable technologies and their applications in sleep monitoring.

By leveraging these four databases, this review ensured a comprehensive and multidisciplinary approach, enabling the identification of high-quality and relevant studies to address the research objectives.

Table 2.2: Description of Databases Used

Database	**Description**
PubMed [1]	A database of biomedical literature, including research articles related to health and medicine.
Web of Science [2]	A multidisciplinary database providing access to scientific articles across various disciplines.
IEEE Xplore [3]	A database focusing on engineering, computer science, and electronics research.
Science Direct [4]	A database offering access to scientific and technical research articles, primarily in the physical sciences, health sciences, and social sciences.

2.2.4 Search Terms and Keywords

The identified keywords used in the search queries are summarised as follows:

- **Sleep-related terms:** *sleep*.
- **Wearable-related terms:** *wearable*, *wearables*, *fitness tracker*, *smart bands*.
- **AI-related terms:** *machine learning*, *deep learning*, *artificial intelligence*.

The search terms applied to each database are described below:

[1] pubmed.ncbi.nlm.nih.gov
[2] www.webofscience.com
[3] ieeexplore.ieee.org
[4] www.sciencedirect.com

- **PubMed:** ("sleep") AND ("wearable" OR "wearables" OR "fitness tracker" OR "smart bands") AND ("machine learning" OR "deep learning" OR "artificial intelligence").
- **Web of Science (WoS):** TS=("sleep" AND ("wearable" OR "wearables" OR "fitness tracker" OR "smart bands") AND ("machine learning" OR "deep learning" OR "artificial intelligence")).
- **ScienceDirect:** ("sleep") AND ("wearable" OR "wearables" OR "fitness tracker" OR "smart bands") AND ("machine learning" OR "deep learning" OR "artificial intelligence").
- **IEEE Xplore:** (("All Metadata":"sleep") AND ("All Metadata":"wearable" OR "All Metadata":"wearables" OR "All Metadata":"fitness tracker" OR "All Metadata": "smart bands") AND ("All Metadata":"machine learning" OR "deep learning" OR "All Metadata":"artificial intelligence")).

2.2.5 Inclusion and Exclusion Criteria

The selection of studies for this review was guided by a comprehensive set of inclusion and exclusion criteria, as outlined in Tables 2.3 and 2.4. These criteria ensured the relevance, quality, and alignment of the selected sources with the research objectives.

The inclusion criteria focused on identifying sources that addressed wearable data for sleep analysis (IC1), examined relationships between sleep patterns and other influencing factors (IC2), or explored AI techniques, including explainable AI(XAI), applied to wearable sleep data (IC3).

Conversely, the exclusion criteria were designed to eliminate sources that did not meet the scope or quality requirements of this review. Studies not written in English (EC1), published before 2020 (EC2), or lacking a focus on wrist-worn wearable data or AI techniques for sleep (EC3) were excluded. Additional exclusions included sources that did not report measurable metrics (EC4), duplicate records (EC5), systematic reviews or other review articles (EC6), and studies focusing on specific conditions outside the intended scope (EC7).

Table 2.3: Inclusion Criteria

Criteria	Description
IC1	The source addresses the use of wearable data for sleep analysis.
IC2	The source examines relationships between other factors and sleep patterns.
IC3	The source explores AI techniques applied to wearable sleep data, including explainable AI (XAI).

2.2.6 Quality Assessment

To ensure the reliability and validity of the findings, a rigorous quality assessment was conducted on all studies included in the final review. Each study was evaluated based on predefined criteria to assess the reported outcomes' methodological robustness, relevance, and clarity.

Table 2.4: Exclusion Criteria

Criteria	Description
EC1	Sources not written in English.
EC2	Sources published before 2020.
EC3	Sources that do not focus on using wrist-worn wearable data or AI techniques for sleep.
EC4	Sources that do not report measurable metrics.
EC5	Duplicated sources.
EC6	Sources classified as systematic reviews or other review articles.
EC7	Sources focusing on particular conditions.

The quality assessment process focused on several key aspects:

- **Clarity of Objectives:** Studies were assessed to ensure they presented clearly defined research questions and objectives that align with the scope of this review.
- **Methodological Rigour:** The methodological design was critically examined, focusing on data collection methods, sample size, and analytical approaches to ensure scientific robustness.
- **Relevance of Outcomes:** The reported results were evaluated based on their alignment with the inclusion criteria and contribution to addressing the research questions.
- **Transparency and Reproducibility:** Studies were reviewed for sufficient methodological detail and data presentation to allow replication or validation of findings.
- **Bibliometric Criterion:** All articles published in journals ranked as Quartile 1 (journal impact ranking) (Q1) and Quartile 2 (journal impact ranking) (Q2) were included. Articles from journals ranked as Q3 and Q4 were excluded unless they demonstrated significant scientific relevance and/or a high number of citations, warranting their inclusion.

The quality assessment protocol was meticulously implemented to include only those studies that satisfied rigorous methodological standards and demonstrated relevance to the research focus. This systematic approach ensured that the final compilation of studies provided robust, well-substantiated, and applicable insights regarding the research questions. Consequently, this process preserved the scientific integrity and credibility of the review.

2.2.7 Data Extraction and Synthesis

In total, 812 records were identified across multiple databases, including PubMed (229), Web of Science (412), IEEE Xplore (92), and Science Direct (79). After 243 duplicate records were removed, 569 unique sources proceeded to the abstract screening phase.

During this phase, sources were assessed based on the inclusion and exclusion criteria. A total of 485 records were excluded due to irrelevance or failure to meet the criteria. This left 84 reports sought for retrieval, all of which were successfully retrieved for further evaluation.

The 84 retrieved reports underwent an eligibility assessment. Based on a detailed application of the exclusion criteria, 29 were excluded. Ultimately, 55 studies were deemed relevant and included in the final review (see Appendix A for the table with the articles).

Figure 2.2 illustrates the selection process, highlighting each stage from identification to inclusion.

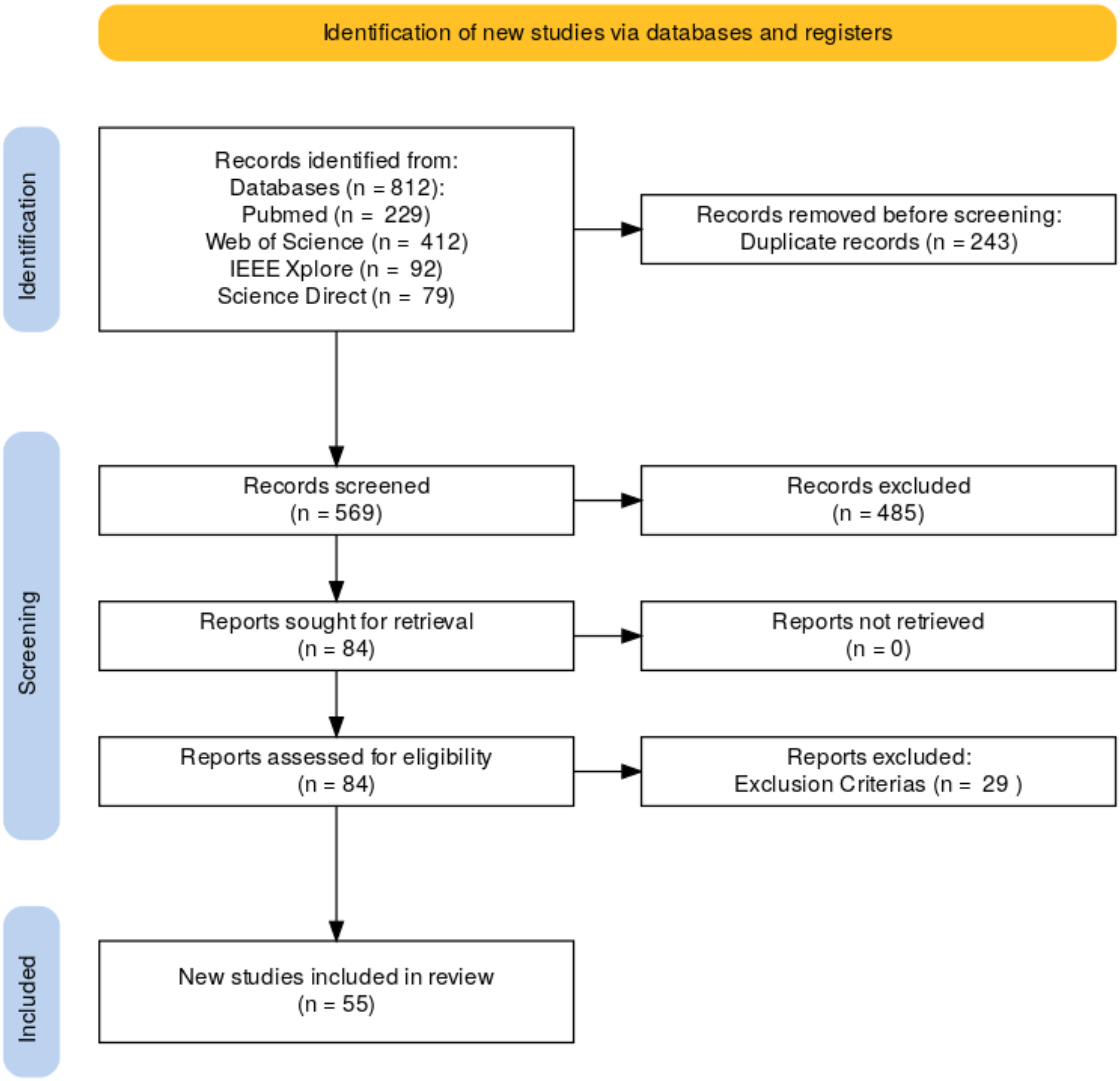

Figure 2.2: Flow diagram of the paper selection process for the systematic review

2.2.8 Research Questions' Answers

This section aims to address the research questions (RQ) based on the findings of the systematic review. Each subsection provides a detailed response to the respective research question, highlighting key insights and evidence from the reviewed literature.

RQ1: What types of data collected by wrist-worn wearable devices are most frequently utilised to assess and evaluate sleep?

Table 2.5 summarizes the primary data types collected by wrist-worn wearables and their frequency of use in sleep monitoring research. The analysis reveals three dominant data sources: Photoplethysmography (PPG), Electrocardiogram (ECG), and accelerometer data, which collectively form the foundation for modern sleep analysis in wearable devices.

Table 2.5: Summary of Data Types and Relevant Articles

Data Type	Articles
Electrocardiogram	(Ma et al. 2023) (Z. Wang et al. 2022) (Yang et al. 2022) (Qin and G. Liu 2022) (Ratheesh et al. 2023) (Sumitra et al. 2023) (Zhou, He, and K. Kang 2022) (Topalidis et al. 2023) (Park et al. 2024) (Iwasaki et al. 2021) (Liang and Chapa-Martell 2021) (Paul et al. 2024) (Chih et al. 2024) (Bahrami and Forouzanfar 2022) (Sharma et al. 2024) (A. John, Cardiff, and D. John 2021) (Das Turja et al. 2024) (X. Wang et al. 2020) (J.-W. Chen et al. 2023) (Shen et al. 2021)
Photoplethysmography	(Silva et al. 2024) (S. Wang et al. 2023) (Benedetti et al. 2022) (Q. Li et al. 2021) (Topalidis et al. 2023) (Song, SR, et al. 2023) (Nam et al. 2024) (Papini et al. 2020) (Liang and Chapa-Martell 2021) (Di Credico et al. 2024) (Chih et al. 2024) (Kotzen et al. 2023) (Motin et al. 2020) (Habib et al. 2023) (Khajehpiri et al. 2024) (Choksatchawathi et al. 2024) (Ye et al. 2021) (Olsen et al. 2024) (Osathitporn et al. 2023)
Accelerometer	(Silva et al. 2024)

Data Type	Articles
	(Benedetti et al. 2022) (Mahadevan et al. 2021) (Sundararajan et al. 2021) (Liang and Chapa-Martell 2021) (Kilic, Saylam, and Incel 2023) (Peraza et al. 2020) (S. Chakraborty, Aich, and Kim 2021)

Other types of data, such as skin temperature (Anusha et al. 2022), are used less frequently and primarily serve as auxiliary sources to complement the core data types like PPG, ECG, and accelerometer signals.

PPG is an optical technique used to measure blood volume changes in the microvascular bed of tissue. Wearable devices use light emitters (usually LEDs) and photodetectors to capture variations in light absorption caused by the pulsing blood flow. This data provides insights into heart rate, blood oxygen saturation (SpO2), and other cardiovascular metrics (Khajehpiri et al. 2024). PPG is widely utilised due to its non-invasive nature and compatibility with compact sensors in wearables.

ECG measures the electrical activity of the heart by detecting and amplifying electrical impulses generated during the cardiac cycle. It provides precise information on heart rate variability (HRV) (Z. Wang et al. 2022), arrhythmias, and sleep-related conditions like apnoea (Shen et al. 2021). While traditionally collected through clinical equipment, wearable technologies have adapted ECG for continuous monitoring in real-world environments, however, it remains a technique that is rarely implemented in the most commonly used consumer wearables.

Accelerometers detect movement and orientation by measuring acceleration forces. In sleep monitoring, this data helps identify sleep-wake states (Sundararajan et al. 2021), restless movements, and transitions between different stages of sleep (Song, SR, et al. 2023). Accelerometer data is particularly useful for detecting sleep patterns and disturbances, making it a foundational element in many wearable devices.

RQ2: What specific sleep patterns and behaviours can be inferred from data collected by wrist-worn wearable devices?

This section addresses RQ2 by exploring how data collected from wrist-worn wearable devices can provide insights into sleep patterns and behaviours. The focus is on three key areas: general sleep quality assessment, sleep stage classification, and the detection of sleep apnoea, highlighting the potential of these devices to enhance sleep monitoring and personalised healthcare.

Sleep Quality

The integration of wearable technology and machine learning has made remarkable advancements in sleep quality assessment. These devices provide detailed insights into various physiological and behavioural metrics captured during sleep. Non-invasive and continuous monitoring of critical data, such as heart rate, breathing patterns, movement, and blood pressure, plays a central role in evaluating sleep quality and its associated metrics.

For example, stress levels during sleep have been effectively monitored through heart rate and breathing fluctuations, with machine learning models achieving up to 98.41% accuracy, demonstrating their potential for early intervention systems (N et al. 2024). Similarly, Gradient Boosting models have successfully classified several sleep disturbances with 93.51% accuracy, highlighting the intricate relationship between lifestyle factors and sleep quality (Soni, Gupta, and Uppal 2023).

However, studies often face limitations due to the narrow range of sleep efficiency values measured (e.g., between 0.93 and 0.95), which reduces variability and may influence the high accuracy observed in predictive models. Additionally, personal differences, such as individual sleep patterns or temporal variations during exam periods and holidays, are often not accounted for, potentially limiting generalizability (Kilic, Saylam, and Incel 2023).

Monitoring transitions between sleep and wake states provides another layer of valuable information. Models such as the Deep Learning Sleep algorithm have achieved high accuracy and device-independent robustness in classifying sleep-wake segments. This robustness addresses data variability effectively, supporting both clinical and real-world applications (Peraza et al. 2020). Furthermore, accelerometer-based analyses using Random Forest models have improved sleep-wake detection and wear/non-wear period classification, offering a cost-effective alternative to traditional methods like polysomnography (Sundararajan et al. 2021). Advanced pipelines that integrate accelerometer, temperature, and light data have further enhanced the precision of sleep metrics, aligning them closely with polysomnography measures in both clinical and home settings (Mahadevan et al. 2021).

Insights into mental health are also increasingly drawn from wearable data. Balanced Random Forest models trained on sleep features and light sleep ratios have been instrumental in evaluating depression, anxiety, and positive mood states, achieving F1 scores of up to 0.776 for depression (Fukuda et al. 2020). Similarly, ensemble learning methods combining classifiers such as Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) have demonstrated 87% accuracy in detecting depression linked to sleep disturbances, further validating the connection between mental health and sleep quality (Hu et al. 2024).

Wearable data analysis has markedly improved the ability to evaluate sleep quality with precision. For instance, machine learning models optimised to extract meaningful patterns from sleep data have achieved an accuracy of 97.54%, making them highly effective for personalised healthcare applications (Hamza et al. 2023). CNN models have also proven capable of assessing sleep over time, achieving up to 97.30% accuracy in weekly evaluations and identifying consistent patterns in sleep quality across longer durations (Arora, P. Chakraborty, and Bhatia 2020). Additionally, Random Forest models have exhibited strong performance in distinguishing between sleep and wake states, achieving 96.04% accuracy, thus providing practical, cost-effective solutions for independent and reliable sleep monitoring in real-world scenarios (S. Chakraborty, Aich, and Kim 2021).

Cardiovascular insights also emerge from wearable sleep data, with models offering non-invasive blood pressure estimation using photoplethysmography (PPG) signals. These methods capture dynamic blood pressure variations during sleep, reinforcing the link between sleep quality and cardiovascular health (Khajehpiri et al. 2024; Motin et al. 2020). Predictive frameworks further extend these applications by identifying critical health events, such as in-sleep strokes, using accelerometer data, underscoring the broader potential of sleep data beyond quality metrics (S. Jeon, Y.-S. Lee, and Son 2023).

In conclusion, wearable devices combined with machine learning have revolutionised the understanding of sleep quality by enabling precise, real-time monitoring of physiological and behavioural metrics. From stress levels and wake-sleep transitions to mental health and cardiovascular insights, the data collected by these systems offer immense potential for advancing personalised healthcare and enhancing overall well-being.

Sleep Stage

The use of wrist-worn wearable devices for sleep stage classification has gained prominence due to their accessibility and scalability, especially with advancements in models leveraging photoplethysmography (PPG) signals and accelerometers combined with deep learning. Models such as SleepPPG-Net utilise temporal convolutional networks (Temporal Convolutional Network (TCN)) to capture contextual information from PPG data, achieving up to 84% accuracy and a kappa coefficient of 0.75. These models surpass traditional methods while facilitating adaptation to new populations through transfer learning approaches (Kotzen et al. 2023).

Similar potential is reflected in approaches like Res-U-Net, which processes PPG and accelerometer signals to predict sleep stages and detect obstructive sleep apnoea, combining pre-training on large datasets such as MESA with fine-tuning for specific applications, resulting in improved F1-scores and generalisation (Olsen et al. 2024).

Recent advancements in multitask learning (Multitask Learning (MTL)) further enhance the potential of wearable devices for sleep staging. By leveraging correlated tasks, MTL models improve the efficiency of learning while reducing the amount of required input data and computational resources. One such MTL model achieved a mean per night accuracy of 77.5% and a Cohen's kappa of 0.643 using low-frequency PPG data from a wrist-worn device, demonstrating its compatibility with wearable technology. This model integrates features such as heart rate (Heart Rate (HR)) and heart rate variability (HRV), effectively predicting sleep stages (wake, light, deep, and REM) with significantly fewer training parameters compared to traditional deep learning models (Chih et al. 2024).

Another significant development in PPG-based sleep stage classification is the use of convolutional neural networks (CNNs) to directly process PPG signals without requiring manual feature extraction. A study by Habib et al. introduced a CNN model that achieved classification accuracies of 94.4%, 94.2%, and 92.9% for two-stage (Wake-Sleep), three-stage (Wake-NREM-REM), and four-stage (Wake-Light, Deep, REM) sleep classification, respectively (Habib et al. 2023). This model utilises overlapped segmentation for data augmentation, addressing the challenge of limited data in sleep studies. By adopting a weighted loss function, the CNN ensures balanced classification across sleep stages, demonstrating superior performance compared to traditional machine learning approaches and other deep learning models. These findings highlight the suitability of CNN-based methods for PPG data, particularly in wearable applications where scalability and reliability are essential.

Recent advancements in transfer learning have demonstrated significant improvements in sleep stage classification using PPG signals. (Q. Li et al. 2021) proposed a deep convolutional neural network (CNN) model initially trained on ECG data from a large dataset (Sleep Heart Health Study) and refined using PPG and actigraphy signals through transfer learning. This approach achieved accuracies of up to 81.49% for binary classification and 68.62% for four-class sleep staging, alongside Cohen's Kappa values of 0.58 and 0.44, respectively (Q. Li et al. 2021). The model leveraged features such as heart rate variability (HRV), cardiorespiratory coupling (CRC), and signal quality metrics, demonstrating that PPG signals

can reliably replace ECG for sleep staging in many scenarios. These results highlight the potential of transfer learning to bridge data gaps between sensing modalities, enabling the use of wrist-worn wearables for accurate and scalable sleep staging.

Another innovative approach to sleep stage classification is the SLAMSS (Sequence-to-Sequence Long Short-Term Memory (LSTM) for Automated Mobile Sleep Staging) model, which utilises actigraphy signals and heart rate metrics derived from wearable devices. The model performs three-stage (wake, Non-Rapid Eye Movement (NREM), REM) and four-stage (wake, light, deep, REM) classification with accuracies of up to 79% and 72%, respectively, when applied to independent populations (Song, SR, et al. 2023). SLAMSS demonstrated robustness in handling low temporal resolution data from consumer-grade wearables, employing attention mechanisms to optimise performance in class-imbalanced scenarios, particularly for the under-represented deep sleep stage. Additionally, the model was validated on an Apple Watch dataset, highlighting its applicability to widely available commercial devices and promoting the feasibility of large-scale continuous sleep monitoring.

In the commercial sector, Samsung smartwatches have achieved 72.0% accuracy and a kappa coefficient of 0.58 in classifying deep sleep and REM stages. However, they still face limitations in identifying light sleep, particularly in individuals over 50, due to age-related changes and an increased apnoea-hypopnea index (Apnoea-Hypopnoea Index (AHI)) Silva et al. 2024. Signal variability and susceptibility to noise, common challenges in consumer-grade devices, are further compounded by lower data granularity, which hampers the accurate classification of minority stages such as deep sleep. Techniques like undersampling and more robust multi-class metrics have shown promise in addressing these limitations (Liang and Chapa-Martell 2021).

Light sleep classification remains one of the most significant challenges due to its overlapping with other stages and inherent class imbalance. However, solutions such as Synthetic Minority Over-sampling Technique (SMOTE) and weighted loss functions have demonstrated encouraging results (Sundararajan et al. 2021). Moreover, inter- and intra-individual variability, particularly in elderly populations or those with sleep apnoea, limits the generalizability of many models. Integrating additional signals, such as skin temperature and electrodermal activity, has notably improved accuracy, particularly in multimodal approaches that differentiate light and deep sleep (Anusha et al. 2022; Topalidis et al. 2023).

Future advancements include refining light and deep sleep classification, expanding multimodal data, and developing personalised models with real-time feedback, fostering greater user adherence and a broader impact on population-wide sleep health (Ma et al. 2023).

Sleep Apnoea

Sleep apnoea, particularly obstructive sleep apnoea (Obstructive Sleep Apnoea (OSA)), has been extensively studied due to its negative impact on sleep quality and vital physiological functions. While polysomnography (PSG) remains the gold standard for diagnosis, its high costs, complexity, and patient discomfort during clinical monitoring limit its widespread use.

As illustrated in Figure 2.3, polysomnography provides a comprehensive example of the traditional diagnostic approach. The image highlights the multiple sensors and equipment required, showcasing the discomfort and complexity often experienced by patients during this procedure. Wearable solutions that leverage physiological signals such as ECG, Photoplethysmography (PPG), and hybrid models are emerging as promising alternatives to continuous, home-based monitoring.

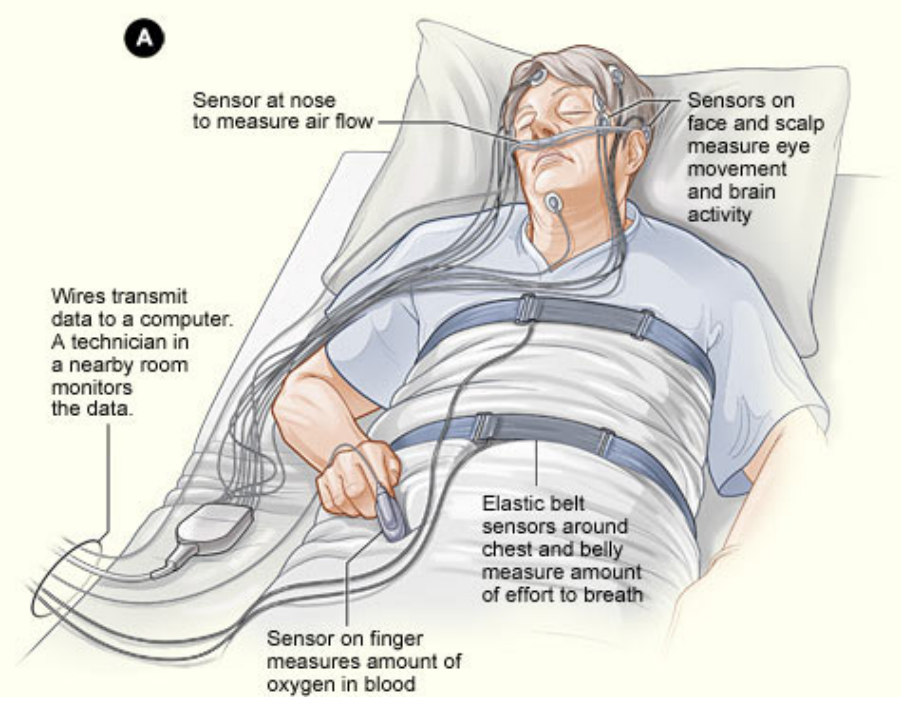

Figure 2.3: Polysomnography example; Image from (Y. Jeon, Heo, and S. J. Kang 2020).

Single-lead ECG data is a key signal that reveals heart rate variability (HRV) and R-R interval (RRI) patterns indicative of apnoea episodes. Recent advancements in deep learning have yielded exceptional results(Z. Wang et al. 2022). For instance, 1D-SEResGNet combines residual learning and attention mechanisms, achieving 90.3% accuracy and 87.6% sensitivity for OSA detection (Yang et al. 2022). Similarly, 1DCNN-RLM with BiGRU-TDM incorporates long-term temporal dependencies, reaching a detection accuracy 91.1% (Qin and G. Liu 2022). An LSTM-based forecasting model using R-R intervals and R-peak amplitudes achieved 94.95% accuracy, highlighting ECG's ability to capture temporal apnoea patterns (Bahrami and Forouzanfar 2022).

In addition, an AI-based scoring toolkit utilising a 1D Convolutional Neural Network (CNN) predicts the AHI, achieving F1 scores of up to 0.83, providing real-time predictions suitable for home-based applications (Ratheesh et al. 2023). A novel LSTM model using raw R-R interval (RRI) data demonstrated 100% sensitivity and specificity with a 60-second segmentation window. This approach showed strong correlations with Peripheral Oxygen Saturation (SpO_2) levels, arousal index, and apnoea events, offering robust performance even with limited clinical data (Iwasaki et al. 2021). FENet, a frequency extraction network, introduces dilated convolution layers to analyse RR-interval signals across multiple frequency bands. FENet excels in continuous and discontinuous OSA detection, achieving superior performance on large datasets such as BestAIR while reducing energy consumption for wearable devices (Ye et al. 2021).

Another study introduced a multiscale 1DCNN model (Multiscale Deep Neural Network - 1D Convolutional Neural Network (MSDA-1DCNN)), which uses RR intervals to detect OSA with 89.4% accuracy. The method balances high recognition rates and low computational load by leveraging multiscale convolution and attention mechanisms, making it suitable for wearable devices (Shen et al. 2021). A 10-layer convolutional neural network trained on

the Apnoea-ECG database achieved 97.8% accuracy after 50 epochs, leveraging spatial information from ECG recordings with batch normalisation and ReLU activations, proving highly effective for OSA detection (X. Wang et al. 2020).

In addition, ResNet and CSPNet models have been proposed to improve accuracy and parameter efficiency. Using SHHS1 data, models achieved 57.55% accuracy in four-level OSA classification, outperforming earlier methods reliant on segmented ECG signals (J.-W. Chen et al. 2023).

Another study introduced a 1D-CNN-based model for sleep apnoea detection using ECG signals from wearable IoT devices, achieving a remarkable second-by-second classification accuracy of 99.56% and sensitivity of 96.05% (A. John, Cardiff, and D. John 2021).

Photoplethysmography (PPG) is another non-invasive signal commonly available in wearable devices like smartwatches and rings. It monitors peripheral blood volume changes, which correlate with respiratory activity. ApSense, a deep learning algorithm optimised for PPG-based apnoea sensing, achieved state-of-the-art results using spatio-temporal blocks, BCNNs, and variational dropout LSTMs. ApSense demonstrated superior performance in balancing sensitivity and specificity across two benchmark datasets (MESA and HeartBEAT), achieving the highest AUROC in challenging high-variance data (Choksatchawathi et al. 2024). An intelligent ring system achieved detection accuracy of 86.73%, sensitivity of 87.42%, and specificity of 85.42% compared to Polysomnography (PSG). This low-cost device (USD 22) offers a practical, wearable solution for at-home OSA screening (S. Wang et al. 2023).

A 1D-CNN named SomnNET leverages peripheral oxygen saturation (SpO2) data for high-resolution, per-second apnoea detection. SomnNET achieved 97.08% accuracy, outperforming lower-resolution state-of-the-art methods. It demonstrated feasibility for smartwatch deployment with reduced complexity via pruning and binarisation techniques (A. John, Nundy, et al. 2021). In addition, a dual-stream temporal Res-U-Net, a deep transfer learning architecture, effectively classifies sleep stages and identifies sleep apnoea events from minimally preprocessed PPG signals. This method achieved high AHI alignment and robust generalisation across multiple datasets by leveraging auxiliary tasks like sleep stage prediction and employing transfer learning with datasets like MESA (Olsen et al. 2024).

Furthermore, (Benedetti et al. 2022; Papini et al. 2020) proposed an innovative AHI estimation method leveraging wrist-worn PPG. (Papini et al. 2020) deep learning model demonstrated strong agreement with PSG-derived AHI values (correlation = 0.61, estimation error = 3 ± 10 events/h). These results highlight the potential of low-cost, wearable solutions for OSA severity classification and screening, achieving robust performance even in clinically heterogeneous populations. Despite its advantages, PPG data remains susceptible to motion artefacts and signal noise, necessitating advanced denoising and preprocessing techniques.

Hybrid models integrating ECG and PPG signals aim to overcome individual signal limitations and improve detection accuracy. For example, BiGRU-based temporal models combined with ADASYN address imbalanced data issues, achieving a balanced trade-off between sensitivity and specificity for apnoea classification (Sumitra et al. 2023; Zhou, He, and K. Kang 2022). Hybrid models that combine scalograms and temporal features have also demonstrated enhanced performance in apnoea detection tasks, effectively improving classification robustness (Paul et al. 2024). Multi-frequency analysis models like FENet utilise both PPG and ECG signals, significantly improving detection accuracy while optimising energy consumption for wearable devices (Xianda Chen et al. 2021), (Ye et al. 2021). These approaches leverage

complementary features from ECG and PPG, improving detection robustness and scalability for real-world applications.

The development of ECG, PPG, and hybrid models, supported by advanced deep learning techniques, has paved the way for cost-effective and non-invasive sleep apnoea detection. Solutions like ApSense, SomnNET, and dual-stream Res-U-Net demonstrate high reliability and accuracy, providing practical alternatives to polysomnography for home-based monitoring. However, challenges such as signal variability, susceptibility to noise, and the need for extensive clinical validation remain. Future research should focus on refining deep learning models with attention mechanisms, integrating ECG and PPG signals, and ensuring robust performance across diverse datasets. These advancements hold significant potential for continuous, unobtrusive monitoring, enabling early intervention and effective management of sleep apnoea.

RQ3: What key factors influencing sleep are identified and quantified through wrist-worn wearable device data?

For the RQ3 question, recent literature identifies four main factors influencing sleep quality, analysed through data from wrist-worn wearable devices. Regarding physical activity, four articles (Arias 2022; Hidayat, Budiarto, and Pardamean 2023; Park et al. 2022; Zamani et al. 2023) highlighted the importance of intensity, regularity, and timing of activity, with a significant impact on sleep efficiency and duration. Concerning light exposure, only one article (Park et al. 2022) explored how light at different times of the day affects sleep efficiency, with benefits observed for exposure in the late afternoon of the previous day. In the nutrition field, one article (Arias 2022) demonstrated the relationship between balanced dietary patterns and improvements in sleep duration and regularity. Finally, physiological variables, such as heart rate variability (HRV), were analysed in two articles (Di Credico et al. 2024; Park et al. 2024), which point to the critical role of these metrics as indicators of the interaction between external factors and sleep quality.

Table 2.6 summarises the key factors identified in the studies along with their corresponding references, providing a clear and concise overview to support the analysis of sleep quality.

Table 2.6: Factors and Associated Articles

Factor	Articles
Physical Activity	(Zamani et al. 2023) (Hidayat, Budiarto, and Pardamean 2023) (Park et al. 2022) (Arias 2022)
Light Exposure	(Park et al. 2022)
Nutrition	(Arias 2022)
Physiological Variables	(Park et al. 2024) (Di Credico et al. 2024)

Recent literature has extensively explored the relationship between behavioural and physiological factors and sleep quality, mainly using data collected through wearable devices. These devices enable continuous, real-time information collection on physical activity, nutrition, and physiological variables, allowing for monitoring sleep patterns and a deeper understanding of

how these factors influence sleep quality. Analysing these factors highlights the importance of data-driven interventions to improve sleep health.

Physical activity

Physical activity consistently emerges as one of the most significant and widely studied factors in modulating sleep quality. Recent studies emphasise that physical activities' intensity, timing, and regularity play critical roles (Soni, Gupta, and Uppal 2023).

The study employing the DLM-WESHMS model (Zamani et al. 2023) reveals that moderate daytime activities are positively associated with longer and more efficient sleep, while evening exercise can compromise rest. This study, with an accuracy of 97.54%, underscores the importance of aligning exercise routines with the circadian rhythm, demonstrating that personalised interventions can significantly benefit sleep health.

Additionally, a study utilising LSTM networks (Hidayat, Budiarto, and Pardamean 2023) demonstrated that temporal granularity in data collection is crucial for capturing subtle patterns between physical activity and sleep. Metrics such as heart rate (HR) and step count (NS) were identified as effective predictors of deep sleep duration, reinforcing that regular daytime exercise is associated with improved sleep efficiency. This model, which analysed data in 10-minute intervals, achieved robust results with an RMSE of 52.18 and an MAE of 40.89, highlighting the relevance of detailed temporal analyses.

Another significant study (Park et al. 2022) found that morning and afternoon physical activity on the previous day was associated with greater sleep efficiency (> 90%), while reduced nighttime activity improved overall sleep quality. A cumulative two-day analysis proved particularly effective for sleep quality prediction, suggesting continuous monitoring-based interventions can offer deep and personalised insights. Finally, an additional article (Arias 2022) emphasised that regular exercise positively impacts objective metrics such as sleep duration and efficiency. Data collected by wearables, such as calories burned, activity duration, and step count, highlight that both the intensity and timing of physical activities can influence different sleep stages, such as deep sleep and Rapid Eye Movement (REM).

Light Exposure

Light exposure is crucial in regulating the circadian rhythm, significantly impacting sleep quality. An article (Park et al. 2022) highlights that light exposure during specific periods of the day can directly influence sleep efficiency. Greater light exposure during the afternoon and evening of the previous day is associated with higher sleep efficiency (> 90%), while reduced light exposure during the afternoon and evening of the same day also contributes to better sleep quality. These findings suggest that light exposure later in the previous day positively prepares for subsequent sleep. In contrast, excessive light exposure on the same day, particularly before bedtime, may delay the circadian phase and impair sleep onset.

Nutrition

In nutrition, the literature points to a strong connection between balanced dietary patterns and sleep quality. An article (Arias 2022) highlights that diets rich in proteins, fibre, and micronutrients such as calcium, potassium, iron, sodium, and vitamin C are associated with longer sleep duration and more regular sleep patterns. Conversely, unbalanced diets are often linked to shorter and poorer-quality sleep, underscoring the importance of informed dietary choices for optimised sleep health. Applications such as MyFitnessPal and Noom, which allow self-reported nutrition data collection, were identified as valuable tools for integrating

dietary patterns with data collected through wearables. This integration enables a detailed analysis of how macronutrients and micronutrients influence rest quality.

Physiological variables

Physiological variables, such as heart rate variability (HRV), have emerged as critical indicators for understanding how non-sleep factors influence sleep quality. HRV, a measure of the variability between heartbeats, reflects the autonomic nervous system's response to external stressors and overall physiological balance, providing insights into how psychological stress, daily activities, and lifestyle choices impact rest. A study (Park et al. 2024) demonstrated that including HRV data in predictive sleep models significantly improves their accuracy (85%) and AUC (80%) using the XGBoost algorithm, suggesting that the physiological state during wakefulness plays a pivotal role in determining sleep efficiency. Additionally, another study (Di Credico et al. 2024) used photoplethysmography (PPG) sensors to measure HRV during periods of wakefulness, achieving a classification accuracy of 76.7%. These findings underline how HRV serves as a biomarker linking external influences such as psychological stress or physical activity with subsequent sleep quality. Together, these studies highlight HRV's potential for identifying how non-sleep factors interact with physiological systems to impact rest, offering valuable insights for developing targeted, real-world interventions to improve sleep.

The literature highlights the interaction between physical activity, nutrition, and physiological variables as the most influential factors in sleep quality. Physical activity is widely associated with longer and more efficient sleep, with intensity and timing being crucial. Nutrition plays an essential role, with balanced diets positively impacting sleep patterns. Lastly, physiological variables, such as HRV, provide deep insights into the relationship between sleep and psychological state. The integration of these factors, facilitated by wearable devices, not only allows precise monitoring but also provides a solid foundation for personalised interventions, paving the way for significant advancements in sleep health.

RQ4: How can explainable artificial intelligence (XAI) techniques be effectively applied to analyse and interpret wrist-worn wearable sleep data?

For the RQ4 question, Explainable Artificial Intelligence (XAI) techniques are essential for analysing and interpreting sleep data collected from wrist-worn wearable devices, enhancing the transparency, trust, and clinical applicability of automated models. These techniques were explored in five articles (Das Turja et al. 2024; Nam et al. 2024; Osathitporn et al. 2023; Sharma et al. 2024; Shen et al. 2021) that address different applications and methodologies.

Explainable Artificial Intelligence (XAI) techniques can be pivotal in analysing and interpreting sleep data collected from wrist-worn wearables, particularly by enhancing automated models' transparency, trustworthiness, and clinical applicability. Wearables increasingly capture physiological signals like heart rate variability (HRV), photoplethysmography (PPG), and accelerometer datanon-invasiveon-invasive, low-cost, and long-term monitoring capabilities. XAI methods make these models interpretable, allowing clinicians and users to understand how predictions are made, thereby increasing trust and enabling targeted interventions for sleep-related disorders.

For instance, a study (Nam et al. 2024) demonstrated the potential of combining PPG and accelerometer data from wearables with machine learning for monitoring sleep stages. By using models such as Gradient Boosting Machines (GBMs) and Explainable AI tools like SHAP (Shapley Additive Explanations), the study identified the importance of features like

HRV and activity patterns in classifying different sleep stages, such as REM and non-REM phases. XAI allowed clinicians to visualise the influence of these features on the classification process, ensuring that the model's predictions align with clinical expectations while improving the system's trustworthiness for personalised healthcare.

Additionally, another study (Sharma et al. 2024) explored the application of Deep Wavelet Scattering Networks to detect insomnia using Electrocardiogram (ECG) signals from wearables. This method, coupled with classifiers such as Weighted K-nearest neighbours and Trilayered Neural Networks (TNN), achieved accuracies above 99% across multiple datasets. By integrating SHAP for feature ranking, the study provided insights into how key signal characteristics contributed to model decisions, facilitating clinician trust and ensuring the applicability of these models in clinical settings.

Similarly, a study (Das Turja et al. 2024) highlighted the use of Heart Rate Variability (HRV) features derived from ECG signals in monitoring sleep apnoea. The study achieved high accuracy by employing an interpretable extreme gradient boosting classifier, showcasing how HRV data alone can be a reliable indicator for apnoea events. The interpretable nature of the model enabled the visualisation of HRV's contributions, enhancing both clinical utility and patient outcomes in wearable-based, long-term monitoring.

In a related development, the RRWaveNet study (Osathitporn et al. 2023) introduced an innovative end-to-end deep learning model for respiratory rate (RR) estimation, a crucial biomarker for various medical conditions, including sleep disorders. The model processes PPG signals directly without requiring feature engineering and demonstrated superior performance across multiple datasets, achieving mean absolute errors as low as 1.23 ± 0.61 breaths per minute in remote monitoring settings. Leveraging transfer learning further enhanced its performance on wrist-worn wearable data. Notably, the study integrated explainable AI(XAI) tools to reveal how signal features, such as peak intensities and their distributions, contribute to RR estimations. These insights are critical for fostering clinician trust and translating models into practical telemedicine applications for sleep monitoring.

In a more advanced application, a study (Shen et al. 2021) proposed a multiscale deep neural network (MSDA-1DCNN) for obstructive sleep apnoea (OSA) detection using HRV derived from single-lead ECG signals. The model achieved an accuracy of 89.4% and utilised dilated convolutions and attention mechanisms to extract high-quality features across multiple scales. The study also tackled data imbalance issues through weighted loss functions, emphasising that combining multiscale feature extraction with explainable AIcan highlight which HRV features most indicate apnoea events, improving model interpretability and performance.

These studies emphasise how XAI techniques such as SHAP and attention-based mechanisms empower clinicians and researchers to comprehend the inner workings of machine learning models. XAI bridges the gap between complex algorithms and their practical applications by providing visualisations and importance rankings of features like HRV, PPG, and activity data. This fosters trust among end-users and healthcare professionals and enhances the personalisation of sleep monitoring systems. Integrating XAI will be essential for advancing diagnostic accuracy and ensuring equitable, user-centred healthcare solutions as wearable technologies evolve.

2.2.9 Integrative Discussion

The integration of data from wearable devices with artificial intelligence (AI) models and explainable artificial intelligence (XAI) techniques offers a unique opportunity to enhance the understanding and management of sleep quality. The advancements presented in this analysis highlight the potential of deep learning algorithms and machine learning techniques to monitor sleep patterns, identify influencing factors, and diagnose sleep-related disorders such as obstructive sleep apnoea (OSA).

A deeper analysis of the differences between the AI models discussed reveals their distinct advantages and challenges. For instance, models like MSDA-1DCNN (Shen et al. 2021), which employ dilated convolutions and attention mechanisms, achieved 89.4% accuracy in OSA detection. These models excel at handling multiscale features in heart rate variability (HRV) signals, standing out for their ability to extract high-quality features. However, such gains in accuracy often come with increased computational costs, which may limit their adoption in wearable devices due to energy and processing constraints. Moreover, existing approaches frequently address sleep disorder detection as a binary classification task, typically distinguishing only between the presence and absence of conditions like obstructive sleep apnoea (OSA). This binary framing inherently limits the capability to capture nuanced severity levels or subtypes of disorders, reducing the overall clinical utility of such methods.

On the other hand, lighter approaches, such as the use of Random Forest and Gradient Boosting Machines (GBM) combined with explainability techniques like SHAP (Nam et al. 2024), offer greater interpretability and computational efficiency. These models are particularly valuable for clinical applications, as they allow healthcare professionals to understand how metrics like HRV and activity patterns contribute to sleep predictions. However, these methods may face limitations in complex, high-dimensional data scenarios where deep learning approaches, such as CNNs or LSTMs, tend to perform better (Das Turja et al. 2024; Sharma et al. 2024).

Another key point is the use of multimodal data in hybrid models, such as the integration of ECG and PPG signals (Sumitra et al. 2023; Ye et al. 2021). These approaches have shown promise in overcoming the limitations of noise and artifacts often associated with standalone PPG data (Topalidis et al. 2023). For example, the combination of BiGRU-based models and ADASYN demonstrated a robust balance between sensitivity and specificity in classifying apnoea episodes (Zhou, He, and K. Kang 2022). However, the increased technical complexity of these models can pose challenges for scalability and adoption in consumer devices. Additionally, advanced physiological sensors such as ECG are not yet common in most commercially available wrist-worn wearables due to their higher cost, complexity, and energy consumption.

In the context of explainability, methods like SHAP and attention mechanisms in neural networks have played a crucial role in making AI models more accessible and trustworthy for healthcare professionals and end-users. Studies such as RRWaveNet (Osathitporn et al. 2023) have demonstrated how explaining key features, such as peak intensities in PPG signals, can enhance the clinical acceptance of wearable devices. However, models like MSDA-1DCNN (Shen et al. 2021) and SomnNET (A. John, Nundy, et al. 2021), despite their high performance, face challenges in balancing explainability with model complexity.

Finally, the applicability of these models to wearables depends on their ability to operate in energy-constrained environments and handle heterogeneous data. The development of efficient, low-cost pipelines will be crucial to ensure that technologies like SomnNET or

ApSense (Choksatchawathi et al. 2024; A. John, Nundy, et al. 2021) can be widely implemented without compromising accuracy or user trust.

In conclusion, the integrated analysis of AI models applied to wearable devices underscores the need to balance performance, computational efficiency, and interpretability. Future advancements should prioritize the combination of machine learning techniques, model optimization, and multimodal data integration to create scalable, accessible, and effective solutions. Additionally, the use of XAI will be indispensable for ensuring transparency, fostering trust, and driving the adoption of wearables as fundamental tools for improving sleep health. It is also important to highlight the clear lack of traditional machine learning models with acceptable performance. These models are generally more explainable and computationally efficient, which makes them especially valuable in clinical contexts and wearable applications. Addressing this gap could contribute to the development of more interpretable and deployable systems, without significantly compromising diagnostic capabilities.

Chapter 3

Methodology and Technological Challenges

This chapter details the complete methodology pipeline developed in this thesis. Three well defined tasks are addressed and will be revisited in separate sections:

1. **Sleep stage classification** from wrist photoplethysmography (PPG).
2. **Sleep apnoea event stratification** from wrist PPG complemented by tri-axial accelerometry (ACC).
3. **Integrated sleep system** combining sleep stage classification and sleep apnoea detection into a unified framework using multimodal physiological signals for enhanced sleep monitoring and personalized health insights.

For sleep stage classification (3.1) and sleep apnoea event stratification (3.2), we follow the same structured, six-step pipeline:

1. *Reading*: raw signal import and sanity checks.
2. *Pre-processing*: signal quality assessment, filtering, and artefact suppression.
3. *Feature extraction*: physiological and motion descriptors engineered per fixed-length epoch.
4. *Data preparation*: class balancing and feature scaling.
5. *Model training and testing*: hyper-parameter optimisation and evaluation metrics.
6. *Explainable AI*: interpretation and visualisation of model decisions using XAI techniques.

Detailed numerical results are deferred to Chapter 4; the focus here is strictly on reproducible methodology.

3.1 Sleep–stage classification

Figure 3.1 illustrates the workflow adopted in this work for automatic five-class sleep stage annotation from wrist photoplethysmography (PPG). All signal processing and machine-learning steps were implemented in Python 3.11, using `numpy`, `scipy`, `scikit-learn`, `imbalanced-learn`, among other libraries.

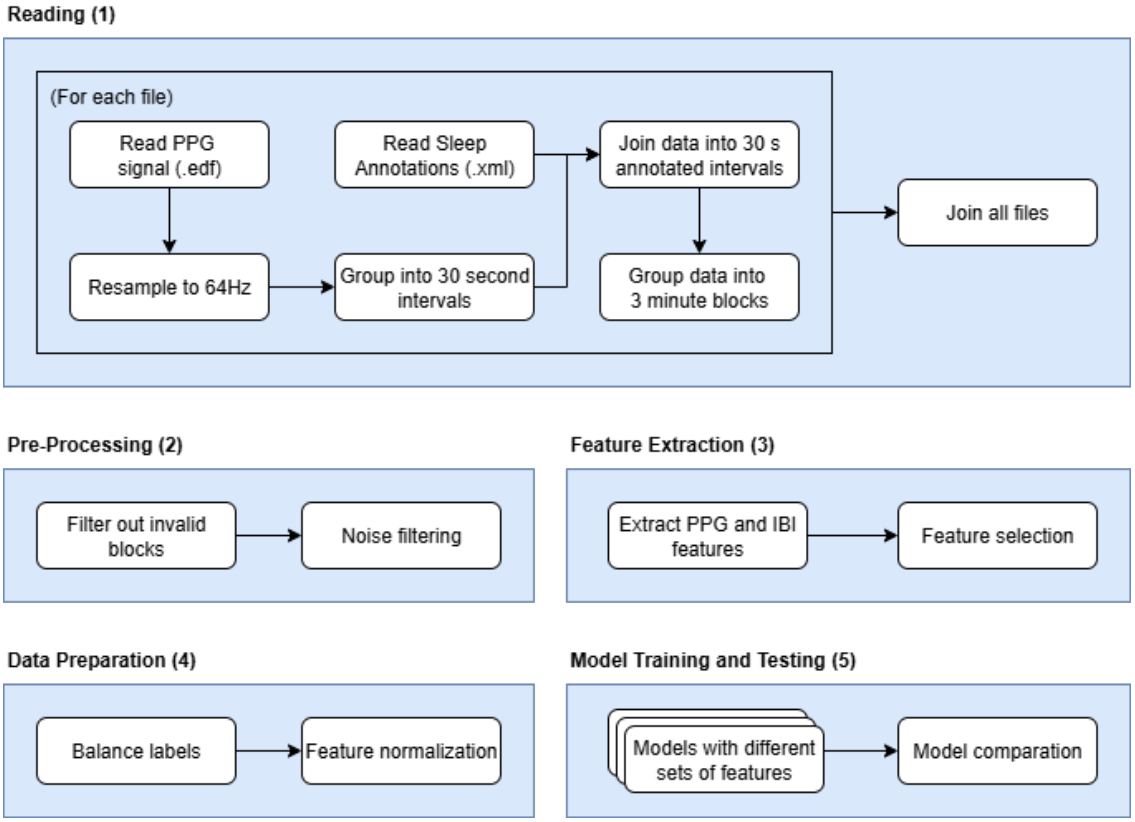

Figure 3.1: Workflow for automatic five-class sleep stage annotation from wrist PPG.

3.1.1 Dataset

The *MESA Sleep* cohort, a subset of the Multi-Ethnic Study of Atherosclerosis (MESA), was employed in this work. The MESA Sleep dataset comprises overnight laboratory polysomnography (PSG) recordings from a large, ethnically diverse community-based population in the United States, collected between 2010 and 2013 (Xiaoli Chen et al. 2015). In addition to the standard PSG channels, the database includes simultaneous wrist photoplethysmography (PPG) recorded at 256 Hz, as well as comprehensive demographic and clinical data for each participant. The full cohort consists of more than two thousand individuals aged between 45 and 84 years, representing four major ethnic groups: White, African-American, Hispanic, and Chinese-American.

To keep the computational burden tractable whilst maintaining ethnic, sex, and age diversity, 48 subjects were selected at random from the database (mean age $= 56.3 \pm 15.8$ y; 52% female). Sleep stage annotations for each subject follow the 2012 AASM rules, providing the canonical five stages: Wake, N1, N2, N3, and REM.

3.1.2 Signal reading and segmentation

The raw PPG signals were extracted from vendor-supplied EDF files and resampled to 64 Hz using polyphase FIR interpolation, ensuring compatibility with typical consumer-grade wearable devices (Virtanen et al. 2020). Each overnight recording was initially divided into

consecutive, non-overlapping epochs of 30 seconds and aligned temporally with sleep stage annotations obtained from polysomnography (PSG).

Various studies recommend the use of stationary data intervals of at least 120 seconds for reliable heart rate variability (HRV) analysis (Huthart et al. 2020). Given sufficient data availability, epochs were grouped into sliding windows of 3 minutes, consisting of six consecutive epochs. These windows moved forward every 60 seconds, following literature recommendations to balance longer signal duration with the diversity and size of the dataset. This overlapping approach increased the number of training examples and enhanced feature robustness.

Each 3-minute window was labeled according to the most frequent sleep stage among the six epochs it included. In cases where a tie occurred, the deeper sleep stage was chosen, such as selecting N2 instead of N1. This choice is supported by evidence that deeper sleep stages are physiologically linked to greater autonomic stability, reduced heart and respiratory rates, and more consistent HRV patterns. Such stable physiological patterns contribute significantly to improving the reliability and interpretability of classification results (Biczuk et al. 2024; Chouchou and Desseilles 2014).

Figure 3.2 illustrates the process of signal extraction, segmentation, and sleep stage labeling.

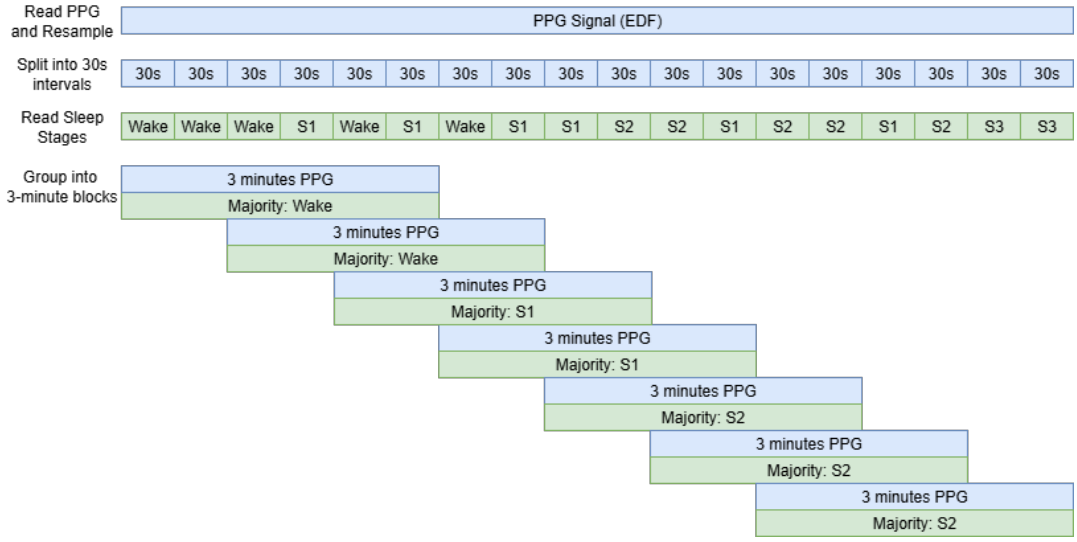

Figure 3.2: Illustration of the signal reading, segmentation, and sleep stage labelling process.

The final class distribution of the labeled 3-minute windows in the selected MESA Sleep dataset is presented in Table 3.1.

Table 3.1: Distribution of labelled 3-minute windows in the MESA subset used for this study.

Stage	Wake	N1	N2	N3	REM
Windows	3 057	611	5 965	1 236	1 361
Share (%)	25.0	5.0	48.8	10.1	11.1

3.1.3 Pre-processing

Two complementary procedures were employed to ensure high-quality feature extraction from the PPG signals:

Signal–quality screening. Each window was subjected to a quality control step and rejected if the peak-to-peak amplitude fell outside the range $0.05 \leq A \leq 2$ (arbitrary units), or if the number of detected systolic peaks was fewer than 90 or greater than 330, corresponding to heart rate values below 30 bpm or above 110 bpm. This peak range was chosen as it encompasses the typical resting heart rate observed under normal physiological conditions during wakefulness (60–100 bpm) (Avram et al. 2019), while also accounting for the lower heart rates commonly recorded during sleep.

Band-pass filtering. Windows that passed the screening were processed using a zero-phase fourth-order Butterworth band-pass filter with a passband of 0.5 Hz to 5 Hz (H. Liu et al. 2021). This filtering step served to suppress baseline drift, motion artefacts, and high-frequency sensor noise, thus enhancing the reliability of downstream analyses.

3.1.4 Feature extraction

Photoplethysmography (PPG) signals encode physiological information across two distinct time scales. The beat-to-beat morphology of the pulse waveform reflects peripheral vasodynamics, while slower fluctuations in the sequence of beat intervals capture aspects of autonomic regulation. To comprehensively represent both these components, each 3 min window was characterised by two complementary families of features, as listed in Tables 3.2 and 3.3. The direct PPG feature set quantifies properties of the raw optical waveform in the time, frequency, and shape domains. The IBI-HRV set is derived from the same window after extracting the inter-beat interval (IBI) series and calculating standard heart rate variability (HRV) descriptors. The joint use of these feature families enables the combination of noise-resilient HRV statistics with fine-grained morphological information, thus providing a richer characterisation of sleep physiology.

Table 3.2: PPG-based features (22 per window).

Group	Features	Brief description
Descriptive	`ppg_mean`, `ppg_median`, `ppg_std`, `ppg_mad`, `ppg_iqr`	Basic location and spread statistics
Morphology	`ppg_rise_time`, `ppg_decay_time`, `ppg_slope`	Pulse shape, rise/fall time, global trend
Frequency domain	`ppg_psd`, `ppg_peak_freq`, `ppg_cwt_energy`, `ppg_fft_energy`	Spectral power and frequency features
Heart rate	`ppg_bpm`, `ppg_peak_rate`	Estimated heart rate and detected peak rate
Signal shape	`ppg_skewness`, `ppg_kurtosis`	Higher-order moment descriptors
Complexity (Hjorth)	`ppg_hjorth_activity`, `ppg_hjorth_mobility`, `ppg_hjorth_complexity`	Hjorth mobility, activity, complexity

Table 3.3: IBI-based features (10 per window).

Group	Feature	Brief description
Time-domain HRV	`ibis_mean`, `ibis_std`, `ibis_rmssd`, `ibis_pnn50`, `ibis_nn50`, `ibis_cv`	Mean, spread, and short-term variability of intervals
Frequency domain	`ibis_vlf_power`, `ibis_lf_power`, `ibis_hf_power`, `ibis_lfhf_ratio`, `ibis_total_power`	Spectral power in VLF, LF, and HF bands and ratios
Non-linear	`ibis_sampen`, `ibis_sd1`, `ibis_sd2`, `ibis_dfa`	Entropy, Poincaré plot, DFA

3.1.5 Data preparation

With the features extracted in the previous stage and the corresponding sleep stage labels, the dataset was prepared for machine learning through a two-step process designed to mitigate the effects of class imbalance and to improve model convergence.

Class balancing. Sleep stages are naturally unequally distributed throughout the night, resulting in significant imbalance in the extracted 3-minute windows. To address this, class balancing was performed separately for the development (training) and test sets. For the development set, a two-step strategy was implemented: minority classes were oversampled using the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples via interpolation in feature space; simultaneously, majority classes were randomly undersampled. The target was set to 4,473 samples per class, corresponding to the most represented original class. This approach resulted in a balanced development set with equal representation of all five sleep stages. Figure 3.3 clearly demonstrates the impact of the class-balancing procedure, contrasting the markedly skewed original distribution of 3-minute windows with the uniform distribution achieved for each sleep stage after the combined over and undersampling steps. For the hold-out test set, only random undersampling was applied, reducing all classes to the size of the smallest class (153 windows). No synthetic samples were introduced in the test set, ensuring an unbiased and realistic evaluation.

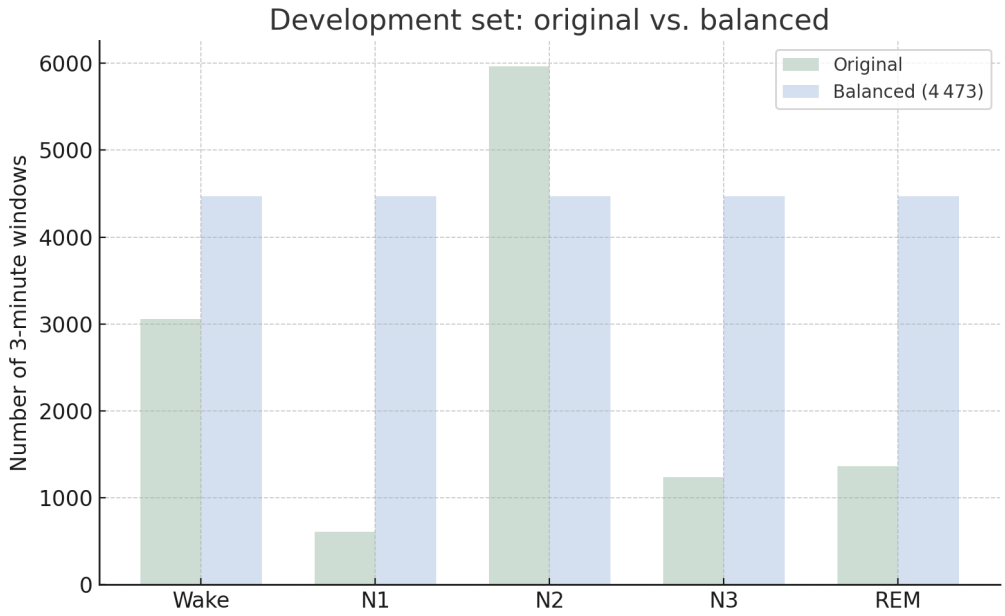

Figure 3.3: Distribution of 3 min windows in the development set before and after balancing.

Normalization. To promote efficient model training and avoid issues related to disparate feature scales, all features were individually normalized to the $[0, 1]$ interval using min–max scaling. The transformation is given by:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

where X is the original feature value, X_{min} and X_{max} are the minimum and maximum values for that feature, estimated exclusively from the development set. This normalization facilitates model convergence and is particularly beneficial for algorithms sensitive to feature scaling, such as support vector machines and k-nearest neighbours.

3.1.6 Model selection and optimisation

To evaluate the performance of different machine learning approaches for sleep stage classification, six standard algorithms were selected: a single classification tree, an ensemble random forest, Gaussian Naïve Bayes (GNB), k-nearest neighbours (KNN), a support vector machine (SVM) with RBF kernel, and the eXtreme Gradient Boosting library (XGBoost). These models were chosen to represent a diverse set of methodologies, spanning probabilistic, instance-based, tree-based, and ensemble learning paradigms.

Hyperparameter optimisation was performed using an exhaustive grid search within a stratified five-fold cross-validation framework on the development set. The macro-averaged F_1 score was adopted as the selection criterion, ensuring a balanced evaluation of precision and recall across all five sleep stages.

Table 3.4 summarises the hyperparameter search spaces explored for each model, along with brief model descriptions. For each algorithm, the primary hyperparameters considered and their tested values are listed.

Table 3.4: Trained models and respective hyperparameter search spaces.

Model	Hyperparameter	Values tested
Classification Tree	criterion	gini, entropy
	max_depth	None, 5, 10
	min_samples_split	2, 5, 10
Random Forest	n_estimators	100, 150, 250
	max_depth	None, 5, 10
	min_samples_split	2, 5, 10
Gaussian NB	var_smoothing	1×10^{-9}, 1×10^{-8}, 1×10^{-7}
KNN	n_neighbors	3, 5, 10
	weights	uniform, distance
	metric	euclidean, manhattan
SVM	C	0.1, 1, 10
	kernel	linear, rbf
	gamma	scale, auto
	degree	2, 3, 4
XGBoost	n_estimators	50, 100, 150
	learning_rate	0.01, 0.1, 0.2
	max_depth	3, 5, 7

After identifying the optimal hyperparameter configuration for each model, the final versions were retrained on the entire development set. Their performance was then evaluated on the independent hold-out set, as described in Section 3.1.5.

3.1.7 Evaluation

Final model evaluation was performed by retraining each classifier on the entire balanced development set, followed by assessment on the independent hold-out set. Multiple metrics were reported to provide a comprehensive view of model performance: overall accuracy, Cohen's κ, per-class precision, recall, and F_1-score, as well as the full confusion matrix. This set of metrics enables robust and transparent comparison with results reported in the contemporary literature. A detailed discussion of these results is provided in the following sections.

3.1.8 Explainable AI

To ensure the interpretability of the Machine Learning models developed in this work, a component of Explainable Artificial Intelligence (XAI) will be incorporated. This component aims to provide a deeper understanding of the internal workings of the model and of the factors that most influence its predictions, both at a general and individual level.

The approach will be divided into two main categories of explanations:

Global Explanations

Global explanations aim to provide an overall view of the model's behavior in relation to the entire dataset. The following techniques will be applied:

- **Permutation Feature Importance:** This technique will be used to evaluate the relative importance of each feature by measuring the change in the model's performance

when the values of a given variable are randomly permuted. The importance will be quantified using the mean and standard deviation of performance degradation over multiple repetitions.

- **SHAP (SHapley Additive exPlanations) – Global:** The mean SHAP values will be computed for each feature and for each class, allowing identification of the variables that contribute most significantly to the model's predictions. This approach is based on Game Theory and provides consistent and fair explanations.
- **Partial Dependence Plots (PDPs):** PDPs will be used to illustrate how isolated variation in a given feature affects the predicted probability for each class. Since automatic PDP generation is not supported for this type of multiclass data, a manual simulation will be performed: the mean values of the other features are fixed, and the feature of interest is varied across its observed range. The predicted probability is then recorded at each step. This approach provides a simple and interpretable view of the individual effect of continuous variables on model predictions Friedman 2001.

Local Explanations

Local explanations focus on interpreting the model's decisions for individual instances, helping to understand why a specific prediction was made. The following techniques will be applied:

- **SHAP – Local:** Specific predictions will be analyzed using waterfall plots, which intuitively show how each feature contributed to the model's decision. One representative instance will be selected per class (predicted as that class) for individual analysis.
- **LIME (Local Interpretable Model-agnostic Explanations):** This technique approximates the model's behavior in the local neighborhood of the analyzed instance using a simple interpretable model, usually linear. Explanations will be generated for representative instances of each class, highlighting the variables with the greatest local impact on the prediction.

This combination of techniques will enable a comprehensive, transparent, and trustworthy analysis of the model's decision logic, fostering user trust and helping to identify bias or unexpected patterns.

3.2 Sleep Apnoea Detection

This section describes the methodology developed for the automatic detection of sleep apnoea events using wrist photoplethysmography (PPG) signals, supplemented by tri-axial accelerometry data. The overall workflow, illustrated in Fig. 3.4, mirrors the structure established for sleep stage classification, including segmentation, feature extraction, data preparation, model selection and optimisation, and performance evaluation.

Unless otherwise specified, dataset partitions, preprocessing strategies, and feature extraction procedures were consistent with those described previously. The use of accelerometry features was specifically explored to assess their added value in apnoea event detection. The following subsections detail each stage of the pipeline.

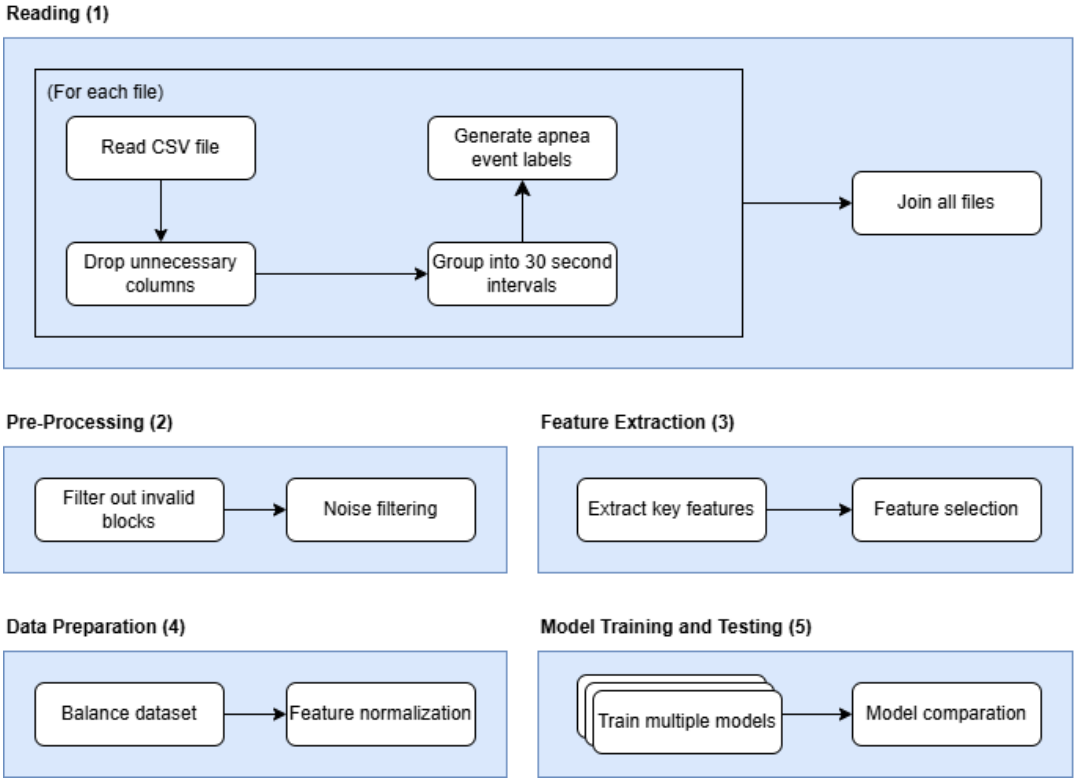

Figure 3.4: Workflow for sleep apnoea detection from wrist PPG and accelerometry.

3.2.1 Dataset

The publicly available DREAMT dataset (*Dataset for Real-time Sleep-stage Estimation using Multisensor Wearable Technology*) (W. K. Wang et al. 2024) comprises a single overnight polysomnography-validated (PSG) recording for each of 100 adults. The cohort spans both healthy volunteers and patients diagnosed with diverse forms of sleep-disordered breathing, ensuring a robust foundation for clinically oriented analysis.

Each participant's file contains multiple physiological signals recorded at various frequencies, such as 64 Hz for both the TIMESTAMP and BVP signals, 32 Hz for triaxial accelerometry (ACC_X, ACC_Y, ACC_Z), 4 Hz for both EDA and TEMP, and 1 Hz for HR, alongside the technician annotated sleep stage labels that are recorded every 30 seconds. Notably, DREAMT is unique in that it is the only dataset obtained using wearable technology specifically for real-time sleep stage estimation.

An assessment of missing values revealed that the dataset is remarkably complete, with only a few instances of missing sleep stage annotations typically due to re-setup events during the recording session. This high level of data completeness enhances the reliability of the derived insights. The average age is approximately 56.24 years, with participants represented from 21 years onward, thereby offering a comprehensive demographic spread. The gender distribution is balanced, with a slight predominance of females, which minimises potential gender bias in subsequent analyses. Investigation of body mass index (BMI) shows that overweight and obese individuals form the majority of the sample, a finding that may reflect the higher BMI prevalence typical of the U.S. population (Health n.d.).

The analysis extends to an examination of common medical conditions, with the most frequently occurring conditions clearly illustrated in Figure 3.5. Additionally, the prevalence of sleep disorders has been explored and notable conditions such as obstructive sleep apnoea (OSA) and insomnia are among those most frequently observed, underscoring the dataset's relevance for sleep disorder research. A moderate correlation between the Apnoea-Hypopnoea Index (AHI) and BMI suggests that increased body mass may be associated with more severe sleep-disordered breathing, while a strong inverse correlation between oxygen saturation (Mean_SaO2) and sleep apnoea emphasises the clinical significance of oxygen levels in sleep health assessments.

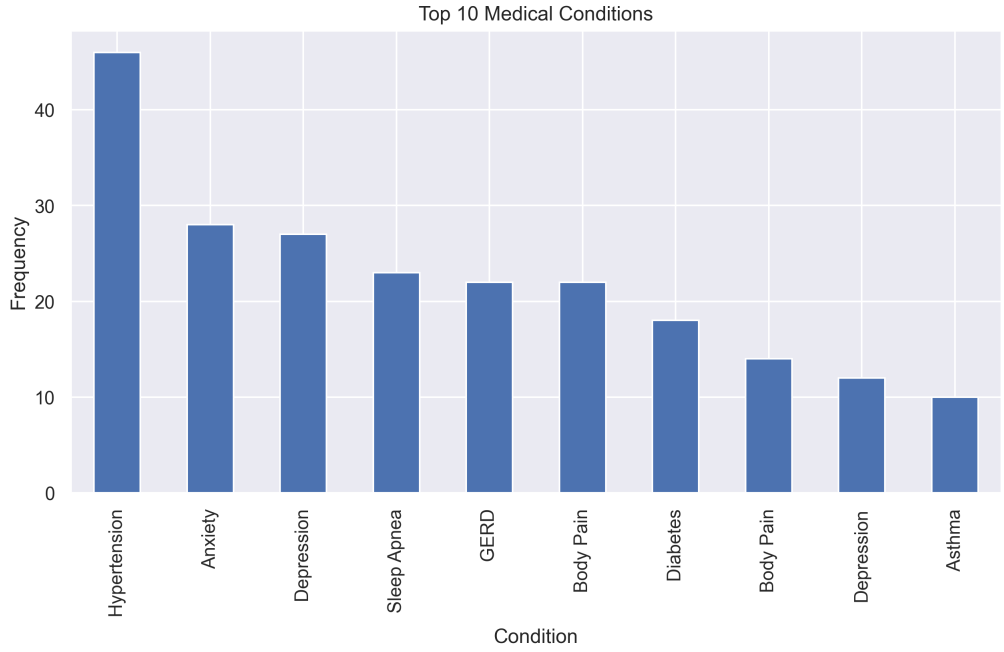

Figure 3.5: Top 10 medical conditions in the DREAMT dataset

In terms of age-related trends, the analysis of the apnoea index indicates that although no definitive trends were evident overall, there is a noticeable increase in the 30–40 age group, possibly influenced by outliers. These insights provide valuable information on the intricate relationships between physiological signals and sleep stages.

Overall, the exploration of the DREAMT dataset confirms its remarkable detail and specialised nature. The findings derived from this analysis lay a robust foundation for subsequent feature engineering, model development, and validation, and highlight the datasets potential to advance research in real-time sleep stage estimation using wearable technology.

3.2.2 Data Reading and Segmentation

This subsection details how the nightly CSV files supplied by the DREAMT repository were converted into a uniform collection of fixed-length, labelled segments, forming the foundation for subsequent analysis. Each file included wrist photoplethysmography (PPG) sampled at 64 Hz, tri-axial accelerometry (ACC), instantaneous heart rate, inter-beat intervals (IBI), and event counters reflecting clinical scoring. Channels unrelated to cardiorespiratory analysis, such as skin temperature and electrodermal activity, were discarded, retaining only the signals relevant for this study.

Timestamps were referenced as seconds from lights-off, and the streams were partitioned into consecutive, non-overlapping thirty-second epochs. The choice to maintain fixed 30-second windows in this context was driven by the characteristics of apnoea events. These episodes are typically short, often lasting between 10 and 30 seconds, and are annotated with precise timing (Berry et al. 2017). Aggregating such events into longer windows, as done for sleep staging, would either split individual episodes across segments or combine heterogeneous events, resulting in loss of specificity and clarity. Therefore, preserving the original segmentation ensured full event capture, better label consistency, and alignment with clinical annotation practices. Each segment retained the full vector of PPG and ACC samples, the corresponding mean heart rate, and the list of IBIs within the epoch.

Figure 3.6 illustrates the complete process of data reading, channel selection, and segmentation adopted in this work.

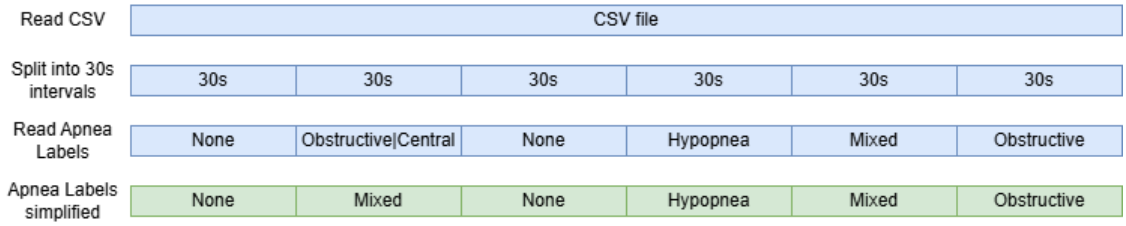

Figure 3.6: Workflow for reading and segmenting raw data from the DREAMT repository into uniform thirty-second epochs suitable for machine learning.

The four event counters provided, wich are obstructive apnoea, central apnoea, hypopnoea, and mixed events, were consolidated into a single categorical label for each segment. If only one event counter was non-zero, the corresponding event type was assigned; if none were active, the label `None` was applied; and if multiple counters were active, the label `Mixed` was used. The resulting labels were encoded as integer values suitable for machine learning algorithms. This procedure, applied to the entire DREAMT cohort, yielded 60,082 fully annotated thirty-second segments, which served as the basis for subsequent feature extraction and model development.

3.2.3 Preprocessing

To ensure high-quality signals for feature extraction, a comprehensive preprocessing pipeline was implemented.

Quality Checks. Segments labelled as "Missing" or "Wake" were excluded. Additionally, segments showing unrealistic or flat-line PPG signals were discarded.

Filtering. PPG signals underwent a second-order high-pass Butterworth filter (cut-off frequency 0.15 Hz) to remove baseline wander, followed by a fourth-order Butterworth band-pass filter (0.7–8 Hz) isolating cardiac-related frequencies and attenuating respiratory or high-frequency noise.

Artefact Removal. Motion artefacts were further mitigated by replacing samples exceeding a robust median-based z-score threshold (>5 standard deviations) with the median value of the segment and smoothing the result with a three-sample median filter.

An example illustrating the effect of the preprocessing pipeline on a raw wrist PPG segment is shown in Fig. 3.7.

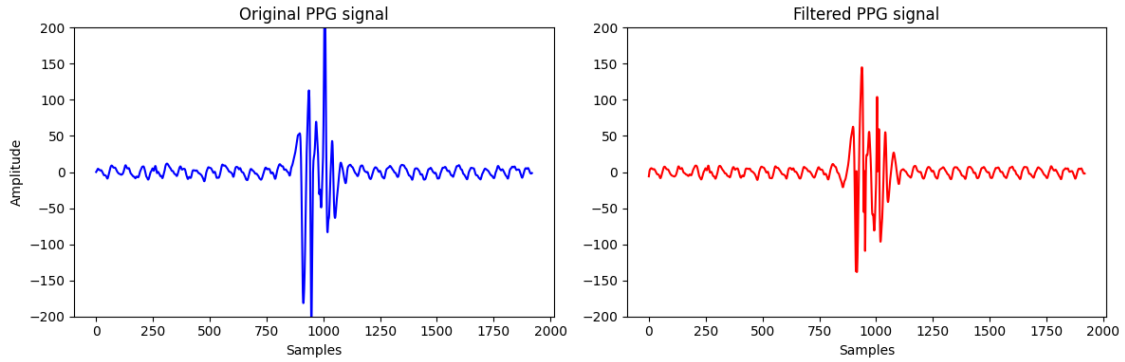

Figure 3.7: Example of raw and filtered wrist-PPG signal segment (30 s). Panel (a): Raw PPG signal with artefacts; Panel (b): Cleaned signal after high-pass and band-pass filtering.

3.2.4 Feature Extraction

Feature extraction for sleep apnoea detection leveraged information from three complementary sources: pulse-wave morphology (PPG), heart rate variability (IBI-derived), and wrist accelerometry. The complete feature set comprised 42 variables per window, as summarised in Tables 3.5–3.7.

Table 3.5 details the 22 PPG-based features, capturing classical statistics, pulse morphology, spectral content, beat timing, higher-order shape descriptors, and signal quality indices. Table 3.6 lists the 10 IBI-derived heart rate variability metrics, which reflect autonomic regulation and temporal complexity of cardiac cycles. Table 3.7 presents the 10 accelerometer features, describing the intensity and directionality of body motion.

Although this multi-source feature space is rich, including all variables can unnecessarily inflate model complexity and promote overfitting. To address this, a systematic feature selection process was conducted. Feature relevance was assessed using four complementary methods: mutual information and ANOVA F-tests for univariate ranking, principal component analysis (PCA) for global contribution, and pairwise Pearson correlation to detect redundancy. Features consistently ranked as low-importance or exhibiting strong redundancy were removed. Specifically, `ppg_mean` and `ppg_peak_freq` were discarded, yielding a final input space of 40 variables for model training.

Table 3.5: PPG-based features (22 per window).

Group	Feature	Brief description
Descriptive	`ppg_mean`, `ppg_median`, `ppg_std`, `ppg_mad`, `ppg_iqr`	Classical location and spread measures.
Morphology	`ppg_rise_time`, `ppg_decay_time`, `ppg_slope`	Global rise, fall, and trend of the pulse wave.
Cardiac power	`ppg_band_pow`, `ppg_peak_freq`	Welch power (0.7–4 Hz) and its peak frequency.
Time–frequency	`ppg_cwt_energy`, `ppg_fft_energy`	Energy from Gaussian-CWT map and half-sided FFT.
Beat timing	`ppg_hr_bpm`, `ppg_rr_mean`	Heart-rate and mean RR interval from detected peaks.
Shape/complexity	`ppg_skewness`, `ppg_kurtosis`, `hjorth_activity`, `hjorth_mobility`, `hjorth_complexity`	Higher-order moments and Hjorth parameters.
Quality/Respiratory	`ppg_snr_db`, `kurt_ratio`, `resp_amp_mean`, `resp_amp_std`	Fast SNR proxy, kurtosis ratio, and respiratory-band envelope statistics.

Table 3.6: IBI-based features (10 per window).

Group	Feature	Brief description
Time-domain HRV	`ibis_mean`, `ibis_std`, `sdnn`, `ibis_rmssd`, `rmssd`, `ibis_pnn50`	Mean, spread and short-term variability of successive intervals.
Poincaré	`ibis_sd1`, `ibis_sd2`	Short and long axis spread of the Poincaré plot.
Complexity	`ibis_sampen`	Sample entropy (irregularity) of the Inter-Beat Interval (IBI) series.
Spectral power	`ibis_total_power`	Integrated 0.003–0.4 Hz power of the HRV spectrum.

Table 3.7: Accelerometer features (10 per window).

Group	Feature	Brief description
Axis statistics	`acc_{x,y,z}_mean`, `acc_{x,y,z}_std`	Mean and standard deviation for each axis.
Axis energy	`acc_{x,y,z}_energy`	Sum of squared samples per axis (movement intensity).
Composite	`acc_mag_mean`	Mean vector magnitude (overall motion).

3.2.5 Data Preparation

Due to inherent class imbalance, data preparation involved careful balancing:

Stratified Splitting. The dataset (60,082 segments) was split into an 80% development set and a 20% independent hold-out set, preserving original event proportions.

Class Balancing. The training set underwent synthetic oversampling of minority classes using the SMOTE (Synthetic Minority Oversampling Technique) algorithm, followed by random undersampling of majority classes to create an evenly balanced training set. The final balanced training set consisted of 6,640 samples (1,328 per class), and the hold-out set contained 830 evenly distributed samples (166 per class) for unbiased model evaluation. Figure 3.8 clearly demonstrates the impact of this procedure.

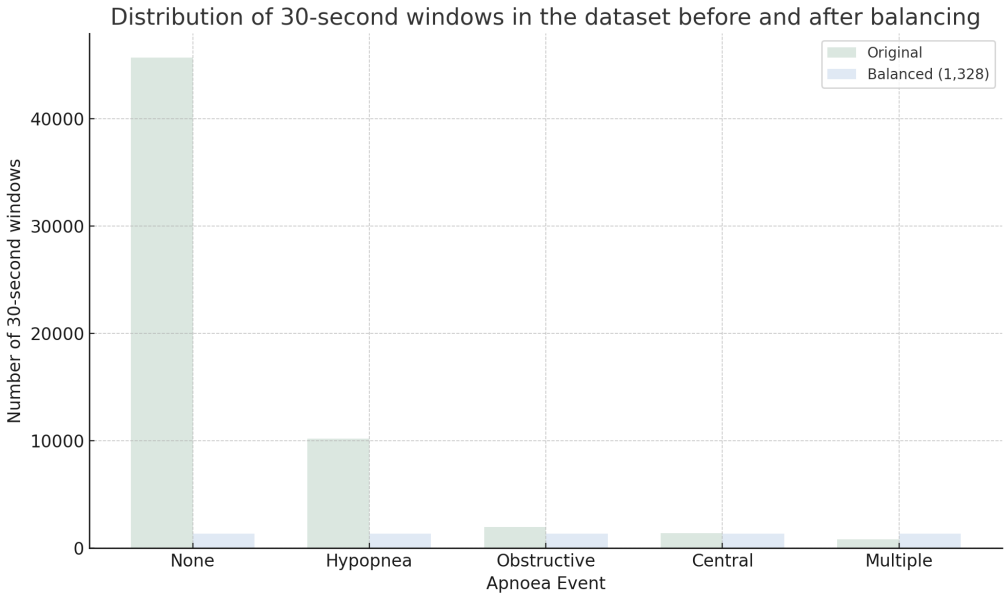

Figure 3.8: Distribution of 30-second windows in the dataset before and after balancing.

Feature Scaling. Features were normalised using min-max scaling, mapping values to the range [0,1].

3.2.6 Model Training and Hyperparameter Optimisation

Interpretable, lightweight machine learning models were selected and rigorously evaluated, including Decision Tree, Random Forest, AdaBoost, k-Nearest Neighbours (KNN), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). These algorithms were chosen to balance predictive power with interpretability and computational efficiency.

Hyperparameter optimisation was performed through either an exhaustive grid search or a 200-draw randomised search, depending on model complexity and computational cost. All search procedures were wrapped in stratified k-fold cross-validation to avoid optimistic bias and ensure robust generalisation. Hyperparameters explored included the number of estimators, maximum tree depth, minimum samples per split, number of neighbours, learning rates, and other model-specific complexity controls.

The full search spaces for each classifier are detailed in Table 3.8.

Table 3.8: Hyperparameter grids/ranges explored for each candidate model.

Model	**Parameter**	**Values tested**
Decision Tree	`criterion`	{gini, entropy}
	`max_depth`	{None, 5, 10}
	`min_samples_split`	{2, 5, 10}
Random Forest	`n_estimators`	200–1500 (14 values)
	`max_depth`	{5, 15, 25, . . . , 95}
	`min_samples_split`	{2, 5, 10, 20}
	`min_samples_leaf`	{1, 2, 4, 8}
	`criterion`	{gini, entropy, log_loss}
XGBoost	`n_estimators`	{100, 150, 200}
	`max_depth`	{None, 5, 10}
	`min_samples_split`	{2, 5, 10}
k-Nearest Neighbours	`n_neighbors`	{3, 5, 7}
	`weights`	{uniform, distance}
	`metric`	{euclidean, manhattan}
AdaBoost	`n_estimators`	{50, 100, 150}
	`learning_rate`	{0.01, 0.1, 1.0}
LightGBM	`n_estimators`	200–1500 (14 values)
	`max_depth`	{5, 15, 25, . . . , 95}
	`learning_rate`	0.001–0.30 (30 values)
	`num_leaves`	20–140 (step 10)
	`min_child_samples`	5–95 (step 10)

3.2.7 Evaluation Metrics

The trained models were evaluated on the independent hold-out set using a range of performance metrics commonly adopted in clinical sleep apnoea research. In addition to balanced accuracy, defined as the mean recall across all five event categories, Cohen's κ statistic was calculated to quantify inter-rater agreement beyond chance. Precision, recall, and F_1-score were reported both globally and on a per-class basis, providing a nuanced view of classifier

performance across the different apnoea event types. Confusion matrices were examined to further elucidate patterns of misclassification.

This comprehensive evaluation approach facilitates direct comparison with established results in the field. Detailed results and interpretation are presented in Chapter 4.

3.2.8 Explainable AI

The explainable AI methodology applied to sleep apnoea detection mirrors the previously described approach for sleep stage classification, maintaining consistency across tasks. Global feature importance (including permutation importance), SHAP-based analyses (both global and local), and representative instance explanations are reused to interpret the best-performing model for apnoea event stratification. This unified framework enables coherent analysis and facilitates direct comparison of interpretability results between the two classification problems.

3.3 Integrated sleep system

This section presents the integrated system developed to simultaneously analyse sleep stage occurrence and the incidence of apnoea events using signals collected from wrist-worn wearable devices. The goal is to provide a comprehensive view of an individual's nocturnal physiology by combining two complementary tasks: sleep stage classification and apnoea event detection. Both tasks are performed independently but synchronously, based on the same physiological signals, specifically photoplethysmography (PPG) and tri-axial accelerometry (ACC), allowing for unified and clinically relevant interpretation. The system pipeline is illustrated in Figure 3.9. Raw signals recorded throughout the night are first subjected to a preprocessing stage involving signal cleaning and normalisation. This step removes artefacts, non-physiological fluctuations, and invalid data segments, ensuring that only high-quality data are passed on to subsequent analysis. The signals are then directed into two parallel branches, each optimised for the requirements of its respective task.

For sleep stage classification, signals are segmented into three-minute overlapping windows, capturing the longer-term dynamics typical of sleep architecture. A set of physiological features is extracted from each window, describing waveform morphology, heart rate, and beat-to-beat variability. These features are used by a pre-trained machine learning model to assign one of the five canonical sleep stages: Wake, N1, N2, N3 or REM to each window.

In parallel, for apnoea detection, signals are divided into consecutive, non-overlapping thirty-second windows in line with clinical scoring practices. Feature extraction includes PPG-based markers, variability metrics from inter-beat intervals, and motion descriptors from accelerometry. These features are used to classify each segment into one of five categories: obstructive apnoea, central apnoea, hypopnoea, mixed event or no event, using a separately trained model.

The outputs from both models are then temporally aligned using timestamp information, producing a coherent time series in which both sleep stage and respiratory event predictions can be reviewed side by side. This fusion allows for integrated visual representations such as layered hypnograms and supports cross-analysis between sleep structure and disordered breathing. The modular design of the system also facilitates the independent upgrade or replacement of individual components without compromising overall functionality.

In addition to automated classification, the system incorporates a dedicated interpretability component that makes the models' decisions transparent. This is especially important in clinical contexts where trust in algorithmic output is essential. The explanations for each prediction are generated following the methodological protocol described earlier in this chapter, using explainable AI techniques such as feature importance, permutation analysis and Shapley values. These methods are applied consistently to both models, enabling the system to highlight which physiological or motion-based features contributed most to each decision, whether for sleep staging or apnoea event detection. The outputs include both global explanations, which summarise the overall importance of features, and local explanations for individual time windows, allowing for detailed inspection and validation.

The entire processing pipeline described above will be implemented in a backend environment using the Flask framework in Python. This includes signal preprocessing, feature extraction, classification for both sleep staging and apnoea detection, as well as the generation of global and local explanations using explainable AI techniques. The backend will expose the results through RESTful Application Programming Interface (API) endpoints, making it possible to retrieve structured outputs such as predicted labels, timestamps, feature importance scores, and explanatory visualisations. By centralising all computational steps in a dedicated backend, the system ensures scalability, maintainability, and clear separation of concerns between data processing and presentation.

To enhance accessibility and usability, a user-friendly frontend will be developed using web technologies such as JavaScript, HTML and CSS. This interface will allow clinicians, researchers or end-users to interact with the system in an intuitive environment, where both sleep stage and apnoea event predictions can be visualised in synchrony. Visual elements such as layered hypnograms, feature contribution plots, and annotated timelines will facilitate the interpretation of results, supporting informed decision-making. The frontend also ensures modular integration with the backend models, enabling real-time interaction and promoting practical deployment of the system in clinical or home-based settings.

In summary, the integrated sleep system proposed here enables automated and interpretable monitoring of sleep structure and breathing disturbances using wearable technology. This approach represents a significant step forward in accessible, non-invasive sleep analysis, combining accuracy, transparency and practical applicability.

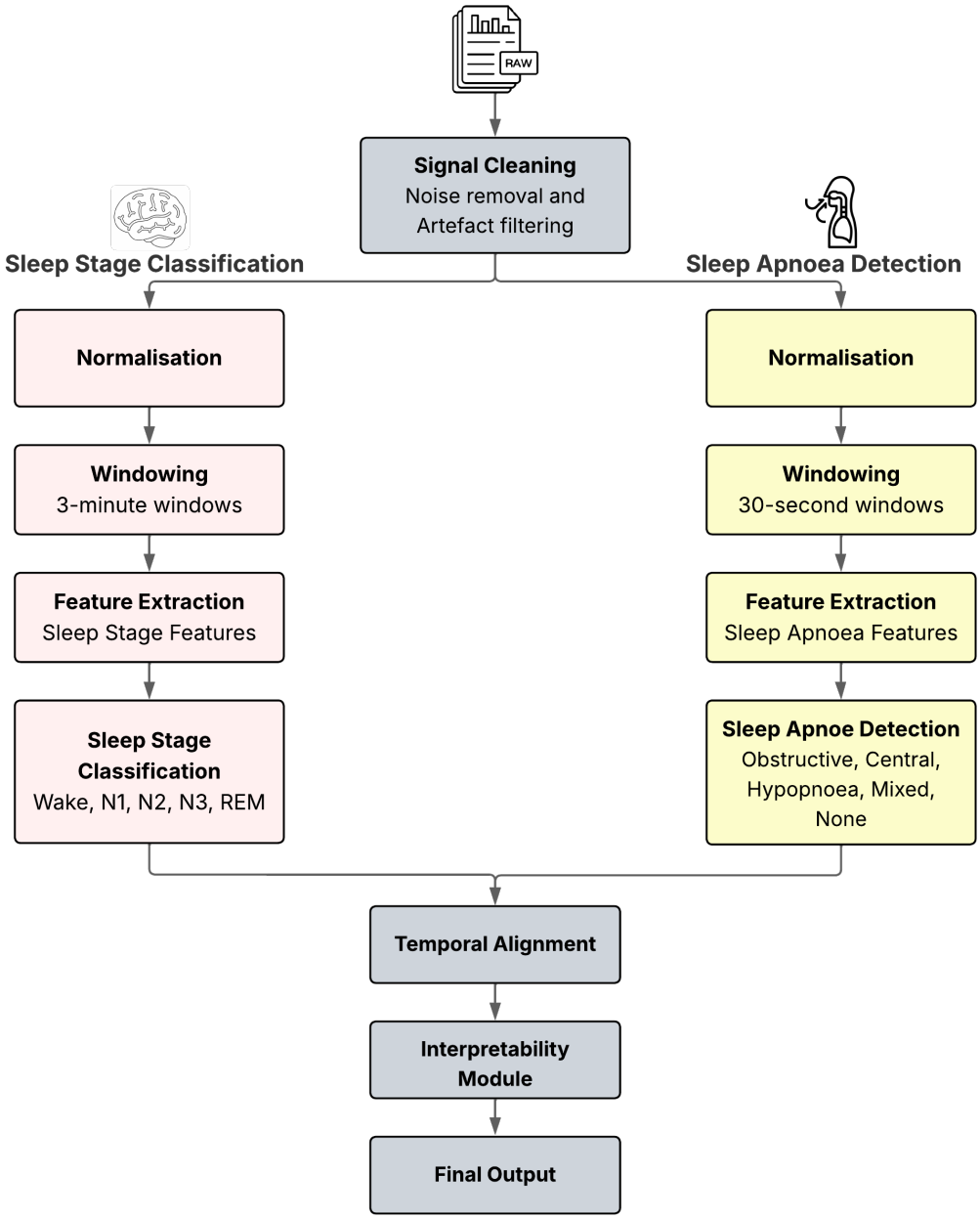

Figure 3.9: Architecture of the integrated sleep analysis system.

Chapter 4

Results and Ethical-Social Considerations

This chapter presents and discusses the empirical findings of the thesis. The material is organised around the two analytical tasks introduced in Chapter 3. Section 4.1 examines how accurately wrist photoplethysmography (PPG) can discriminate the five canonical sleep stages, while Section 4.2 evaluates multi-class detection of sleep-disordered breathing events using a combination of PPG and tri-axial accelerometry.

Beyond the core analytical results, Section 4.3 demonstrates the complete system integration, including the orchestration of all modules through RESTful APIs and the generation of synchronised outputs with local and global interpretability features. These examples illustrate not only the technical feasibility of the system but also its practical applicability in real-world monitoring scenarios.

The chapter also includes a commentary that relates the empirical findings to current literature and highlights the main limitations and open challenges that remain.

Finally, Section 4.5 examines the ethical, legal, and societal implications of deploying AI in healthcare. Particular attention is given to compliance with the General Data Protection Regulation (GDPR) and the European AI Act, as well as broader concerns of transparency, fairness, and data protection. This reflection situates the technical contributions within the regulatory and societal landscape that increasingly shapes AI-driven healthcare innovation.

4.1 Sleep Stage Classification

This section presents the results for automatic multi-class classification of sleep stages from wrist-worn photoplethysmography (PPG) and inter-beat interval (IBI) features. All models were evaluated on an independent, stratified test set comprising five mutually exclusive classes. The subsections that follow first report overall model performance metrics and then analyse per-class behaviour and error patterns.

4.1.1 Overall Model Performance

The overall classification performance, as detailed in Table 4.1, highlights the relevance of combining photoplethysmography (PPG) and inter-beat interval (IBI) features in sleep stage prediction. Among all evaluated models, Random Forest stands out with the highest accuracy (72.16%) and Cohen's Kappa (0.652), indicating strong agreement beyond chance. K-Nearest Neighbors (71.63%, $\kappa = 0.645$) and XGBoost (69.54%, $\kappa = 0.619$) also delivered

competitive results, reinforcing the robustness of ensemble and instance-based methods in handling physiological variability.

In contrast, models such as Support Vector Machines (56.73%, $\kappa = 0.459$) and Gaussian Naïve Bayes (32.03%, $\kappa = 0.150$) exhibited markedly lower performance. These results suggest that simpler or linear classifiers may struggle to capture the complex and non-linear dynamics inherent in biosignals, especially when class distributions are imbalanced or overlapping.

The variation in performance across models highlights the critical importance of model selection when working with physiological data. Ensemble-based approaches appear particularly well-suited to the task, likely due to their ability to model feature interactions and adapt to signal variability.

In the following sections, we delve deeper into class-wise performance and interpretability aspects, offering a more granular understanding of each model's strengths and limitations.

Table 4.1: Overall performance on the test set: accuracy and Cohen's κ for each model.

Model	**Accuracy (%)**	**Cohen's Kappa**
Random Forest	**72.16**	**0.652**
K-Nearest Neighbors	71.63	0.645
XGBoost	69.54	0.619
Classification Tree	58.43	0.480
Support Vector Machine	56.73	0.459
Gaussian NB	32.03	0.150

4.1.2 Per-Class Performance

Detailed per-class metrics for Random Forest, KNN, and XGBoost are presented in Table 4.2.Among the evaluated models, Random Forest consistently achieved the best overall performance, particularly excelling in the classification of Wake (W) and deep sleep (S3) stages, with a high precision of 0.86 and recall of 0.88, respectively. The REM stage was also accurately identified, with balanced precision and recall contributing to a high F_1 Score (F_1)-score of 0.82. In contrast, Random Forest struggled to classify stage S1, with a notably low precision of 0.69 and recall of 0.41, likely due to physiological overlap with neighbouring stages.

The KNN model demonstrated relatively balanced performance across most sleep stages, achieving solid results in REM and S3. This suggests its capability to capture well-defined physiological patterns. However, its performance was less consistent in transitional stages such as S2 and S1, where overlap and signal ambiguity are more pronounced.

XGBoost showed comparable effectiveness in detecting stages like Wake and REM, closely matching Random Forest in terms of precision and recall. Nonetheless, it faced significant difficulties in classifying stage S1, similarly to the other models. This consistent underperformance across models highlights the inherent complexity of accurately distinguishing transient and ambiguous sleep stages such as S1.

Table 4.2: Per-class metrics for Random Forest, KNN, and XGBoost.

Class	Random Forest			KNN			XGBoost		
	Prec	**Rec**	**F_1**	**Prec**	**Rec**	**F_1**	**Prec**	**Rec**	**F_1**
W	0.69	0.89	0.77	0.70	0.80	0.74	0.76	0.86	0.81
S1	0.69	0.41	0.51	0.64	0.53	0.58	0.47	0.29	0.36
S2	0.59	0.70	0.64	0.61	0.70	0.65	0.56	0.77	0.65
S3	0.86	0.78	0.82	0.84	0.77	0.80	0.84	0.75	0.79
REM	0.80	0.83	0.82	0.78	0.78	0.78	0.79	0.82	0.81

4.1.3 Confusion Matrices and Error Patterns

Confusion matrices for Random Forest, KNN, and XGBoost models (Tables 4.3, 4.4, and 4.5) offer detailed insights into classification performance and error patterns.

Table 4.3: Confusion Matrix - Random Forest (Accuracy: 72.16%, Cohen's Kappa: 0.652)

	W	S1	S2	S3	REM
W	136	5	7	3	2
S1	32	62	35	9	15
S2	14	14	107	7	11
S3	6	2	22	120	3
REM	10	7	9	0	127

Table 4.4: Confusion Matrix – k-Nearest Neighbors (Accuracy: 71.63%, Cohen's Kappa: 0.645)

	W	S1	S2	S3	REM
W	122	11	13	4	3
S1	20	81	26	7	19
S2	12	15	107	7	12
S3	4	4	25	117	3
REM	9	7	9	7	121

Table 4.5: Confusion Matrix - XGBoost (Accuracy: 69.54%, Cohen's Kappa: 0.619)

	W	S1	S2	S3	REM
W	132	4	13	3	1
S1	32	45	55	5	16
S2	11	8	118	7	9
S3	3	1	32	114	3
REM	11	4	15	0	123

The Random Forest model achieved an accuracy of 72.16% and a Cohen's kappa of 0.652. It correctly classified the majority of Wake instances (136/153) and REM epochs (127/153), showing strong performance in these classes. However, it confused 32 S1 epochs as Wake and 35 as S2, highlighting the difficulty of distinguishing transitional stages. Stage S2 also showed complexity, being correctly classified in 107 cases but misclassified as S1 in 14 instances and as REM in 11 instances.

The KNN model obtained an accuracy of 71.63% and a Cohen's kappa of 0.645. It classified Wake with reasonable precision (122 correct) but misclassified 11 epochs as S1 and 13 as S2. S1 achieved 81 correct classifications, yet 26 were labelled as S2 and 19 as REM, reinforcing the challenge of early sleep stage discrimination. Similarly, S2 was well recognised in 107 instances, but 15 were mistaken for S1 and 12 for REM, showing overlap across adjacent stages.

XGBoost reached an accuracy of 69.54% and a Cohen's kappa of 0.619. It performed strongly for REM (123 correct) and S3 (114 correct), but significantly struggled with S1. Out of all S1 epochs, only 45 were correctly classified, while 32 were misclassified as Wake and 55 as S2, indicating substantial confusion in this transitional stage. S2, despite being correctly identified 118 times, was also misclassified as S1 in 8 cases and as REM in 9 cases.

In summary, although all models showed solid performance in detecting well-defined stages such as Wake, S3, and REM, they consistently struggled with stage S1. The frequent misclassifications observed in the confusion matrices emphasise the intrinsic challenge of distinguishing transitional sleep stages based solely on PPG and IBI features.

4.1.4 Model Explainability and Interpretation

To understand the decision-making logic of the sleep-stage classifier, the same tiered approach adopted for apnoea was applied: global, class-specific and local perspectives obtained with permutation importance, SHAP, LIME and partial-dependence plots.

Global Feature Importance

Permutation importance computed on the min–max normalised space highlights the prominence of nonlinear PPG-HRV descriptors. *Sample Entropy* (*ibi_sampen*) and basic PPG morphology metrics (*ppg_median*, *ppg_kurtosis*, *ppg_skewness*) top the ranking, followed by heart-rate indices (*ppg_bpm*, *ppg_peak_rate*) and signal-complexity measures (*hjorth_complexity*, *ibi_higuchi_fd*). The ten most influential variables are summarised in Table 4.6.

Table 4.6: Top 10 features according to permutation importance (sleep-stage model).

Feature	Importance Mean	Importance Std
ibi_sampen	0.0468	0.0041
ppg_median	0.0441	0.0094
ppg_kurtosis	0.0297	0.0078
ppg_skewness	0.0268	0.0058
hjorth_complexity	0.0241	0.0044
ppg_bpm	0.0209	0.0068
ppg_peak_rate	0.0195	0.0050
ibi_higuchi_fd	0.0187	0.0094
ibi_mean	0.0180	0.0051
ppg_cwt_energy	0.0174	0.0054

These results suggest that nonlinear and statistical descriptors derived from PPG and IBI signals, particularly those reflecting signal irregularity and morphology, are crucial for distinguishing between sleep stages. The prominence of *ibi_sampen* and *ppg_median* highlights the importance of entropy and central tendency in heart rhythm analysis, while features such as *ppg_kurtosis* and *hjorth_complexity* underline the contribution of waveform shape and complexity. This reinforces the relevance of combining both time-domain and complexity-based metrics in sleep-stage classification tasks.

Class-Specific SHAP Analysis

The SHAP breakdown in Table 4.7 reveals stage-dependent patterns of relevance:

- **Wake (0)** – Characterised by heightened cardiovascular activity, this stage is primarily associated with rate-based PPG indicators (*ppg_peak_rate*, *ppg_bpm*) and waveform sharpness (*ppg_kurtosis*), along with higher signal irregularity (*ibi_sampen*), reflecting sympathetic dominance and movement artefacts.
- **N1 (1)** – As a transitional stage, N1 is influenced by low-frequency heart rate variability (*vlf_power*) and complexity metrics (*higuchi_fd*, *sd2*, *sdnn*), highlighting reduced autonomic stability and the beginning of parasympathetic engagement.
- **N2 (2)** – This light sleep stage shows dominant contributions from entropy-based (*ibi_sampen*) and spectral features (*vlf_power*), with shape-based descriptors (*ppg_kurtosis*, *ppg_skewness*) capturing increased waveform regularity as sleep stabilizes.
- **N3 (3)** – Deep sleep is strongly associated with high *vlf_power* and entropy, indicating profound parasympathetic activity. Morphological features like *ppg_skewness* and *ppg_kurtosis* reflect the symmetric, low-variability nature of the waveform, while geometric spread (*sd2*) signals autonomic consistency.
- **REM (4)** – This stage exhibits a mixed autonomic profile, where median pulse amplitude (*ppg_median*), entropy (*ibi_sampen*, *spectral_entropy*), and spectral power (*vlf_power*) dominate, capturing the instability and variability characteristic of REM physiology.

Table 4.7: Top 5 SHAP features for each sleep stage.

Rank	Wake (0)	N1 (1)	N2 (2)	N3 (3)	REM (4)
1	ppg_peak_rate 0.0237	vlf_power 0.0191	ibi_sampen 0.0198	vlf_power 0.0193	ppg_median 0.0263
2	ppg_bpm 0.0235	higuchi_fd 0.0154	vlf_power 0.0182	ibi_sampen 0.0193	ibi_sampen 0.0208
3	ppg_kurtosis 0.0188	sd2 0.0149	ppg_kurtosis 0.0178	ppg_skewness 0.0181	vlf_power 0.0199
4	ibi_sampen 0.0160	sdnn 0.0137	mean 0.0119	ppg_kurtosis 0.0179	ppg_skewness 0.0173
5	ppg_median 0.0147	ppg_kurtosis 0.0127	ppg_skewness 0.0108	sd2 0.0178	spectral_entropy 0.0164

Partial Dependence Plots (PDPs)

To assess the impact of the three highest-ranked predictors, which are *ibi_sampen* (in the figure as *sampen*), *ppg_median* and *ppg_kurtosis*, each feature was systematically perturbed across its normalised range while the model's class probabilities were tracked (Figure 4.1).

Despite their global importance, the partial-dependence curves reveal moderate sensitivity. The largest shifts (0.10–0.15) occur for *Wake* and *N1*: *Wake* probability declines and *N1* rises as *ibi_sampen* exceeds roughly 0.3, indicating a mild preference for more irregular inter-beat intervals during light sleep. A similar trade-off appears when *ppg_median* surpasses 0.6, with a subsequent uptick in *REM* above 0.75. Probabilities for *N2* and *N3* remain within ±0.05 across all three features, and *ppg_kurtosis* triggers only a brief adjustment near the origin before plateauing. These results confirm that even the model's most influential features do not single-handedly dominate predictions; their effects are modest and largely stage-specific.

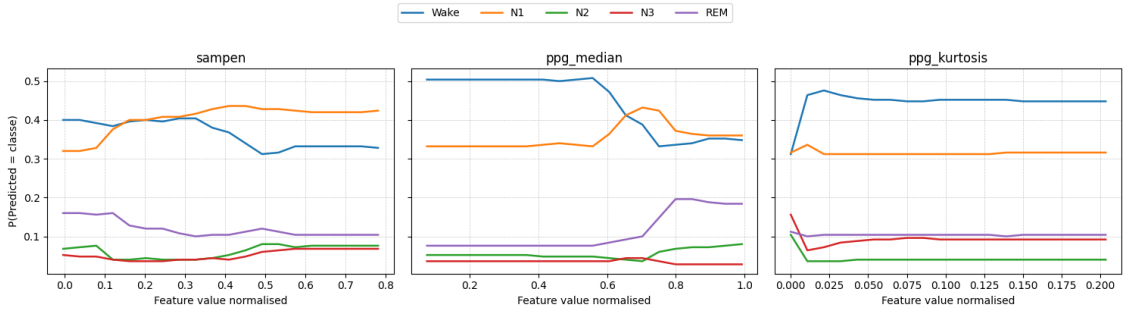

Figure 4.1: Partial-dependence curves for the three most important features in sleep stage classification

For an expanded version of these curves with enhanced readability, please refer to Appendix B (see Figure B.1).

Local Interpretability: SHAP and LIME (Sleep Staging)

To understand how the sleep stage classifier forms individual predictions, we generated both SHAP and LIME explanations for a representative test instance from each stage (*Wake*, *N1*,

N2, *N3*, *REM*). Explanations were computed in the min–max normalised feature space to enable cross-feature comparison. Table 4.8 lists the top five contributing features per stage for SHAP and LIME.

Agreement between the two methods is generally good (typically three or more overlapping variables), though effect direction and magnitude sometimes differ because SHAP integrates the model's non–linear interactions while LIME fits a local surrogate that is condition-dependent. For *Wake*, pulse rate and pulsatility metrics (e.g., `ppg_bpm`, `ppg_peak_rate`) dominate SHAP, whereas LIME emphasises low spectral power and signal-complexity thresholds. *N1* is strongly associated with low-frequency variability (`vlf_power`) and nonlinear complexity (`higuchi_fd`), with supporting evidence from HRV dispersion (`sd2`, `std_dev`). Deep sleep (*N3*) shows increased parasympathetic tone (`hf_power`) together with higher signal irregularity (`ibi_sampen`) and HRV spread (`sdnn`); decreases in pulsatile rate (`ppg_peak_rate`) also contribute. REM predictions are strongly modulated by PPG amplitude/distributional shape (`ppg_median`, `ppg_skewness`) and by central autonomic markers (`higuchi_fd`, `vlf_power`). Overall, the local attributions align with the global importance rankings and map plausibly onto known sleep physiology: lighter stages favour higher variability and rate reactivity, deep sleep reflects vagal dominance and waveform regularisation, and REM is marked by distinctive PPG amplitude shifts.

Table 4.8: Top 5 local features (absolute contribution) for a representative instance of each sleep stage using SHAP and LIME. Values are shown in the min–max normalised feature space.

Stage	SHAP		LIME	
	Feature	**Value**	**Feature**	**Value**
Wake	ppg_bpm	0.0412	higuchi_fd $\leq$ 0.41	0.0189
	ppg_peak_rate	0.0398	vlf_power $\leq$ 0.00	-0.0184
	hjorth_complexity	0.0171	ppg_bpm $>$ 0.23	-0.0172
	ppg_kurtosis	0.0150	ppg_kurtosis $\leq$ 0.00	-0.0158
	ppg_median	-0.0094	ppg_peak_rate $>$ 0.23	-0.0143
N1	vlf_power	0.0457	higuchi_fd $\leq$ 0.41	0.0190
	higuchi_fd	0.0313	vlf_power $>$ 0.01	0.0122
	sd2	0.0252	sdnn $>$ 0.17	0.0104
	ppg_bpm	0.0243	std_dev $>$ 0.17	0.0102
	std_dev	0.0239	0.61 $<$ ppg_median $\leq$ 0.68	0.0100
N2	pnn50	0.0425	ppg_median $>$ 0.74	-0.0223
	nn50	0.0305	higuchi_fd $\leq$ 0.41	0.0196
	ppg_median	0.0210	ppg_kurtosis $>$ 0.00	0.0141
	hjorth_mobility	0.0201	vlf_power $>$ 0.01	0.0136
	ppg_bpm	0.0199	sdnn $>$ 0.17	0.0128
N3	hf_power	0.0483	ppg_bpm $>$ 0.23	-0.0182
	ibi_sampen	0.0393	ppg_median $\leq$ 0.61	0.0165
	sdnn	0.0359	ppg_peak_rate $>$ 0.23	-0.0140
	ppg_peak_rate	-0.0246	hjorth_mobility $>$ 0.44	-0.0118
	std_dev	0.0232	ibi_sampen $>$ 0.41	-0.0098
REM	ppg_median	0.0808	ppg_median $>$ 0.74	-0.0246
	ppg_skewness	0.0647	higuchi_fd $\leq$ 0.41	0.0197
	higuchi_fd	0.0519	ppg_kurtosis $\leq$ 0.00	-0.0154
	vlf_power	0.0356	vlf_power $>$ 0.01	0.0102
	ppg_bpm	0.0287	hjorth_complexity $>$ 0.38	-0.0094

4.2 Sleep Apnoea Detection

This section presents the experimental results for automatic multi-class classification of sleep-disordered breathing events from wrist-worn signals. The models were evaluated on an independent, stratified test set comprising five mutually exclusive classes: *central apnoea*, *hypopnoea*, *mixed apnoea*, *obstructive apnoea*, and the negative class (*None*), as described in the previous chapter.

4.2.1 Overall Model Performance

Table 4.9 provides a detailed overview of the global performance metrics achieved by all candidate models on the test set, considering both classification accuracy and Cohen's κ as primary indicators of success.

The ensemble models, particularly Random Forest and LightGBM, achieved the highest performance across all evaluated methods. Random Forest attained an accuracy of 62.05% and a Cohen's κ of 0.526, with LightGBM presenting very similar results (61.81% accuracy and 0.523 κ). These outcomes clearly demonstrate the effectiveness of ensemble learning strategies for handling complex, multi-class physiological data derived from wrist-worn sensors. The ability of these models to capture subtle, overlapping patterns is reflected in both their robust accuracy and moderate-to-strong agreement beyond chance.

Intermediate models such as XGBoost and K-Nearest Neighbours (KNN) also performed well, with XGBoost achieving 59.76% accuracy and a Cohen's κ of 0.497, and KNN obtaining 56.51% accuracy and 0.456 κ. These results further confirm that more sophisticated algorithms, which can leverage complex decision boundaries or neighbourhood information, are highly beneficial for this task.

While the simpler models, Decision Tree and AdaBoost, showed comparatively lower results (45.06% and 42.53% accuracy, with κ values of 0.313 and 0.282, respectively), their inclusion in this evaluation is crucial. They provide a clear benchmark and underscore the value added by advanced ensemble and optimised models. Their performance, while modest, demonstrates the inherent challenge of the classification problem and validates the need for more robust approaches.

Table 4.9: Overall performance on the test set: accuracy and Cohen's κ for each model.

Model	**Accuracy (%)**	**Cohen's Kappa**
Decision Tree	45.06	0.313
AdaBoost	42.53	0.282
KNN	56.51	0.456
XGBoost	59.76	0.497
LightGBM	**61.81**	**0.523**
Random Forest	**62.05**	**0.526**

Having established the overall effectiveness of each model, it is important to examine how well the best-performing approaches distinguish between individual classes of sleep events. While aggregate metrics such as accuracy and Cohen's κ provide a useful summary, they may mask important differences in model performance across specific event categories. Therefore, the following section provides a more detailed per-class analysis.

4.2.2 Per-Class Analysis

In order to better understand the strengths and limitations of the proposed approach, we conducted a detailed per-class analysis of the two best-performing models: Random Forest and LightGBM.

Table 4.10 presents the per-class precision, recall, and F_1-score for the two top-performing models, Random Forest and LightGBM. This breakdown enables a more nuanced understanding of model performance beyond the aggregate metrics, highlighting the strengths and limitations in distinguishing between specific sleep event categories.

Both Random Forest and LightGBM deliver high sensitivity and precision for central apnoea (recall 0.73, precision 0.68–0.70) and for the "None" class (recall 0.68–0.69, precision 0.64).

These results indicate that the models are particularly effective in identifying events with clear and distinct physiological signatures, achieving consistent and reliable predictions for these categories.

Obstructive apnoea is also detected with good recall (0.70 for both models), but with lower precision (0.52–0.54). This imbalance suggests a tendency towards false positives, which may arise due to overlapping features between obstructive apnoea and other respiratory events, especially in cases where the physiological presentation is less clear-cut.

The most challenging categories for both models are hypopnoea and mixed apnoea. For hypopnoea, F_1-scores remain moderate (0.61), reflecting difficulties in consistently distinguishing these milder events, which often exhibit less pronounced or more variable sensor patterns. Mixed apnoea presents the greatest challenge, with F_1-scores of 0.51 (Random Forest) and 0.47 (LightGBM). The low recall (0.41 and 0.37, respectively) in this class further underlines the complexity of accurately identifying hybrid or ambiguous physiological events based solely on wrist-worn sensor data.

Table 4.10: Per-class metrics for Random Forest and LightGBM (support: 166 segments per class).

Class	**Random Forest**			**LightGBM**			**Support**
	Prec	**Rec**	**F_1**	**Prec**	**Rec**	**F_1**	(n)
Central	0.70	0.73	0.71	0.68	0.73	0.71	166
Hypopnoea	0.63	0.58	0.61	0.61	0.60	0.61	166
Mixed	0.67	0.41	0.51	0.66	0.37	0.47	166
None	0.64	0.68	0.66	0.64	0.69	0.66	166
Obstructive	0.52	0.70	0.59	0.54	0.70	0.61	166

These tendencies highlight the complexities of multi-class classification in sleep medicine, especially when using non-invasive wearable sensors. To gain further insight into the nature and distribution of these misclassifications, the following section examines the confusion matrices for both models, providing a detailed view of where and how errors are most likely to occur.

Given that Random Forest achieved the best aggregate and per–class performance while requiring considerably less hyper-parameter tuning time than LightGBM, the remainder of this section concentrates on a one–vs–rest ROC evaluation *solely* for the Random-Forest model. As will be shown in Fig. 4.2, the resulting class-wise AUCs corroborate the precision/recall patterns observed above: excellent separability for *Central* events (AUC 0.92) and robust discrimination for *None*, *Hypopnoea* and *Obstructive* (AUCs 0.87–0.88), while *Mixed* apnoea remains the most challenging category (AUC 0.85). These Receiver Operating Characteristic (ROC) findings reinforce the choice of Random Forest as the reference architecture for the subsequent interpretability analyses.

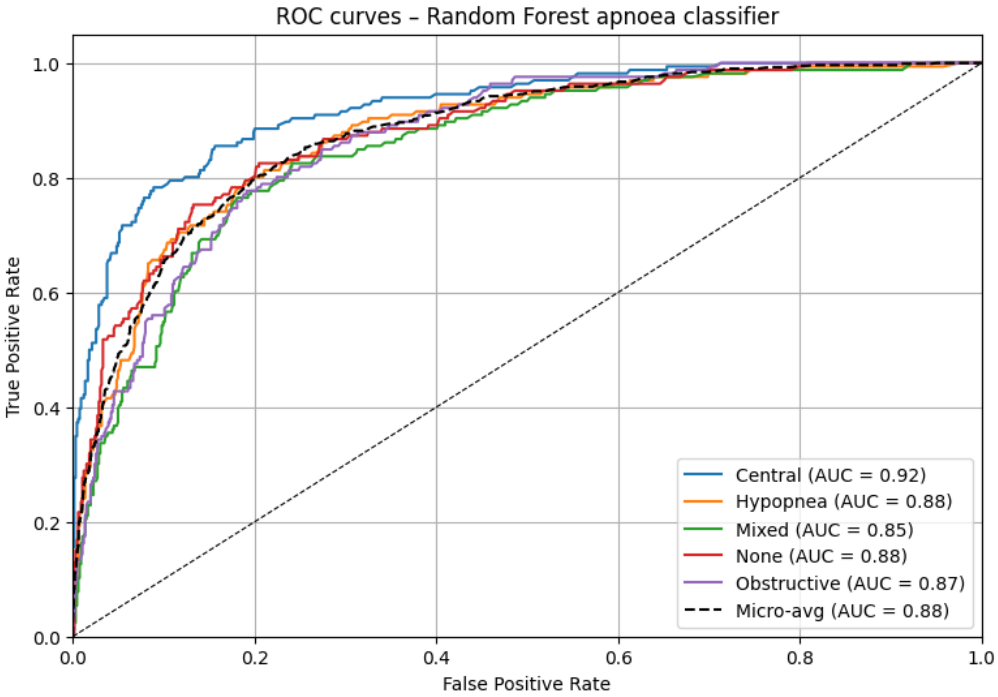

Figure 4.2: One–vs–rest - ROC curves for the Random-Forest apnoea classifier

4.2.3 Confusion Matrices and Error Patterns

The confusion matrices presented in Tables 4.11 and 4.12 provide a granular view of model behaviour, illustrating not only the frequency of correct classifications but also the specific nature of misclassifications between classes. This detailed breakdown is essential for understanding how and why errors occur, offering valuable insights for further model refinement and clinical interpretation.

Examining the Random Forest confusion matrix, we observe that the majority of central apnoea events (121 out of 161) are accurately identified. The remaining central apnoea cases are most commonly confused with obstructive apnoea (16 instances) or incorrectly labelled as None (7 instances). This pattern suggests that, while the model is generally adept at recognising central apnoea, a subset of cases shares physiological characteristics with other apnoea types, leading to occasional ambiguity in classification.

For hypopnoea, the model correctly classifies 97 segments, yet misclassifies a substantial proportion as None (33) or obstructive apnoea (21). This tendency likely reflects the subtle and less distinctive physiological signatures associated with hypopnoea, which can overlap both with normal sleep and with milder forms of obstructive events in the PPG and accelerometer data.

Mixed apnoea poses the greatest challenge for the model, with only 68 out of 166 events correctly identified. The majority of mixed apnoea events are misclassified as obstructive apnoea (58) or central apnoea (21), highlighting the inherent difficulty in distinguishing hybrid apnoeic events using peripheral sensor data. The physiological overlap and transition states present in mixed apnoea likely contribute to this confusion, as these cases may not exhibit clear or consistent patterns distinct from the other classes.

The None class is generally well identified (113 out of 166), although some segments are mislabelled as hypopnoea (22) or central apnoea (15), again reflecting the complexity of accurately differentiating normal sleep from mild or ambiguous apnoeic events.

Obstructive apnoea, meanwhile, demonstrates strong sensitivity (116 out of 167 correctly classified), but some events are incorrectly assigned to other classes, most notably mixed (13), hypopnoea (13), or central apnoea (13). This indicates that, while the model is reliable in detecting clear-cut obstructive events, there remains some degree of overlap with other respiratory disturbances.

Table 4.11: Confusion matrix – Random Forest (rows: true, columns: predicted).

	Central	Hypopnoea	Mixed	None	Obstructive
Central	121	10	7	7	16
Hypopnoea	4	97	11	33	21
Mixed	21	12	68	7	58
None	15	22	3	113	13
Obstructive	13	13	13	11	116

A similar pattern is seen in the LightGBM confusion matrix, which closely mirrors the results of Random Forest. Central apnoea, None, and obstructive apnoea are generally well captured, while mixed apnoea continues to represent the most significant source of confusion. The consistent misclassification patterns between the two top-performing models confirm the robustness of the analytical pipeline, while also pointing to underlying physiological complexities that present challenges for all current approaches.

Table 4.12: Confusion matrix – LightGBM (rows: true, columns: predicted).

	Central	Hypopnoea	Mixed	None	Obstructive
Central	121	6	7	16	16
Hypopnoea	5	100	9	30	22
Mixed	25	22	61	8	50
None	16	22	2	114	12
Obstructive	10	14	14	11	117

Taken together, the confusion matrices reveal several key insights:

- **Central apnoea and None** are both reliably identified, with the majority of errors involving classes with similar or overlapping physiological manifestations.
- **Hypopnoea** is frequently misclassified as None or obstructive apnoea, which can be attributed to its less pronounced and often ambiguous physiological signatures in the sensor data.
- **Mixed apnoea** stands out as the most challenging category, with the highest rate of misclassification. This reflects the inherent physiological ambiguity of mixed events, as well as the current limitations of using only peripheral sensor data for their detection.

- **Obstructive apnoea** exhibits strong sensitivity, as evidenced by a high number of true positives, but precision is somewhat lower, suggesting some tendency to over-predict this class, particularly in borderline cases with hypopnoea or mixed characteristics.

4.2.4 Model Explainability and Interpretation

To understand the decision-making logic of the developed models, a set of explainability techniques was applied. This section presents global, class-specific, and local interpretations using SHAP, LIME, and permutation feature importance.

Global Feature Importance

The analysis of permutation importance, computed on the *min–max normalised* feature space to ensure that all variables contribute on a comparable scale, reveals that accelerometry-based features dominate the model's behaviour, particularly *acc_z_mean* and *acc_y_mean*, which exhibit the highest mean importance scores (0.0255 and 0.0223, respectively). This suggests that body orientation and movement intensity along the vertical and sagittal axes are strongly associated with the classification task.

Physiological features also contribute meaningfully: *hr_mean* ranks fourth (0.0130), and PPG morphology and HRV indices, such as *ppg_skewness*, *ibis_mean*, and *ibis_pnn50*, appear consistently among the top 10. This distribution indicates that the model balances motor and cardiovascular cues, likely capturing both the presence of apnoeic movements and autonomic changes during disordered breathing. These observations are visually supported by the global importance ranking shown in Table 4.13.

Table 4.13: Top 10 features according to permutation importance.

Feature	Importance Mean	Importance Std
acc_z_mean	0.0255	0.0062
acc_y_mean	0.0223	0.0078
acc_z_energy	0.0139	0.0015
hr_mean	0.0130	0.0056
acc_y_energy	0.0100	0.0039
acc_x_mean	0.0092	0.0045
ppg_skewness	0.0086	0.0020
ibis_mean	0.0082	0.0054
ibis_std	0.0082	0.0036
ibis_pnn50	0.0075	0.0032

Class-Specific SHAP Analysis

The SHAP analysis per class, illustrated in Table 4.14, unveils patterns in feature relevance that differ substantially across apnoea types. For the **Central** class (Class 0), the top feature is *acc_x_energy*, followed closely by *ppg_kurtosis* and *acc_y_mean*. This suggests that the model recognises this subtype primarily through abrupt movement in the medio-lateral direction and waveform shape changes in the PPG signal.

In the **Hypopnoea** class (Class 1), PPG features dominate: *ppg_kurtosis*, *acc_z_mean*, and *hr_mean* appear among the most relevant. This points to subtle cardiorespiratory variations rather than movement-related features, consistent with the less intense physiological disruption associated with hypopnoea.

For the **Mixed** class (Class 2), no single modality stands out. The top features include *hjorth_mobility*, *ppg_hr_bpm* and *acc_x_std*, suggesting the model requires both motion and HRV fluctuations to resolve this hybrid category.

In the **None** class (Class 3), which represents absence of apnoea, the prominence of *acc_x_std*, *ppg_kurtosis* and *kurt_ratio* likely reflects the physiological baseline (i.e., regular movement and signal symmetry) used as a reference by the model.

Finally, in the **Obstructive** class (Class 4), *acc_y_mean* is the most influential feature by a wide margin (0.0282), with additional emphasis on *acc_x_energy* and *ibis_mean*. This confirms the association between sustained torso motion and obstructive respiratory effort, which aligns with clinical understanding.

Overall, these results confirm that the model adjusts its reliance on different physiological domains depending on the apnoea subtype, supporting both the flexibility and interpretability of the predictive process.

Table 4.14: Top 5 SHAP features for each class.

Rank	**Central** (Class 0)	**Hypopnoea** (Class 1)	**Mixed** (Class 2)	**None** (Class 3)	**Obstructive** (Class 4)
1	acc_x_energy 0.0237	ppg_kurtosis 0.0130	acc_x_std 0.0201	acc_x_std 0.0230	acc_y_mean 0.0282
2	ppg_kurtosis 0.0192	acc_z_mean 0.0121	hjorth_mobility 0.0161	ppg_kurtosis 0.0216	acc_x_energy 0.0169
3	acc_y_mean 0.0159	acc_y_mean 0.0110	acc_y_mean 0.0124	acc_z_std 0.0183	acc_y_energy 0.0157
4	acc_z_mean 0.0157	kurt_ratio 0.0108	ppg_hr_bpm 0.0111	kurt_ratio 0.0177	ibis_mean 0.0127
5	ibis_mean 0.0146	hr_mean 0.0100	acc_z_std 0.0108	hjorth_mobility 0.0168	acc_x_std 0.0096

Partial Dependence Plots (PDPs)

This experiment deliberately takes the opposite perspective of the sleep stage analysis (where the most influential predictors were highlighted). There, the goal was to show how strongly the classifier reacts to its top-ranked features, here, the aim is to demonstrate that it remains virtually *un*responsive when truly peripheral variables are perturbed. The three PPG summary statistics intentionally selected for their absence from every previous global analysis, which are *ppg_median*, *ppg_std*, and *ppg_mad*, display virtually flat partial-dependence curves (all class-wise fluctuations stay within ±0.02). As illustrated in Figure 4.3, their uniformly horizontal profiles confirm that these variables contribute negligible discriminative information to the classifier.

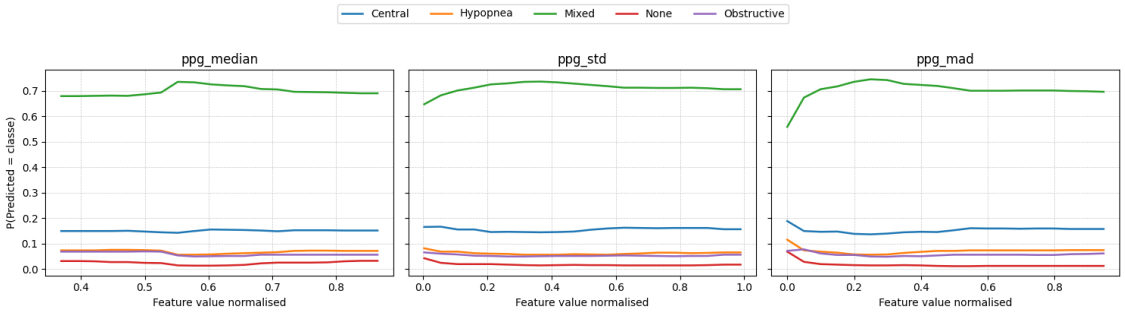

Figure 4.3: Partial-dependence curves for the three low-relevance features.

For an expanded version of these curves with enhanced readability, please refer to Appendix B (see Figure B.2).

Local Interpretability: SHAP and LIME

To explore how individual predictions are formed, both SHAP and LIME explanations were applied to representative samples from each class. These local explanations reveal which features were most influential in each case and how they contributed to the model's output.

Table 4.15 compares the top five local features for a selected instance per class, showing both SHAP and LIME contributions. While there is generally good agreement between the methods, slight differences highlight how each approach captures local influence using different approximation techniques.

The local explanations in Table 4.15 corroborate the global findings and add clinical nuance. For both SHAP and LIME, computed on the min-max normalised space, accelerometry dominates the apnoea positive classes: in *Central* and *Obstructive* events, high values of `acc_x_energy`, `acc_y_mean` and `acc_z_mean` provide the strongest positive evidence, consistent with brief arousal related motions and changes in arm orientation. Conversely, the *None* class is driven by the absence of movement and by clean PPG morphology; here `ppg_kurtosis` and `kurt_ratio` contribute negatively, signalling a stable, symmetrical pulse wave. Hypopnoea instances reveal a mixed pattern, moderate accelerometric activity coupled with subtle autonomic markers such as low `ppg_hr_bpm`, reflecting partial airway obstruction without full arousal. The high overlap between SHAP and LIME (three to four shared variables per class) reinforces confidence in the model's reasoning, while the occasional differences illustrate how the two methods weight non–linear versus locally linear interactions. Overall, the local analyses confirm that the integrated system fuses motor and cardiovascular cues in a physiologically plausible manner.

Table 4.15: Top 5 local features for a representative instance of each class using SHAP and LIME.

Class	SHAP		LIME	
	Feature	**Value**	**Feature**	**Value**
Central (0)	acc_x_energy	0.0870	hr_mean $\leq$ 0.16	-0.0167
	acc_x_mean	0.0539	ppg_kurtosis $>$ 0.01	-0.0137
	ppg_kurtosis	0.0386	kurt_ratio $>$ 0.00	-0.0120
	acc_z_mean	0.0306	acc_mag_mean $>$ 0.63	0.0119
	acc_z_energy	0.0296	acc_x_energy $>$ 0.50	-0.0091
Hypopnoea (1)	acc_y_mean	0.0224	0.01 $<$ ppg_kurtosis $\leq$ 0.01	-0.0114
	acc_x_mean	0.0200	0.55 $<$ acc_z_mean $\leq$ 0.76	-0.0106
	acc_y_energy	0.0155	0.00 $<$ kurt_ratio $\leq$ 0.00	-0.0103
	acc_x_std	0.0148	ppg_hr_bpm $>$ 0.59	-0.0099
	acc_x_energy	0.0139	ppg_mad $\leq$ 0.14	0.0072
Mixed (2)	ppg_median	0.0312	ppg_kurtosis $>$ 0.01	-0.0116
	acc_y_std	0.0292	kurt_ratio $>$ 0.00	-0.0104
	acc_x_std	0.0264	ppg_hr_bpm $>$ 0.59	-0.0101
	kurt_ratio	0.0259	acc_x_std $>$ 0.01	-0.0082
	ppg_kurtosis	0.0244	ppg_rr_mean $\leq$ 0.18	-0.0072
None (3)	acc_x_std	0.0477	ppg_kurtosis $\leq$ 0.00	0.0196
	kurt_ratio	0.0342	kurt_ratio $\leq$ 0.00	0.0137
	ppg_kurtosis	0.0235	ppg_skewness $\leq$ 0.60	0.0105
	ppg_skewness	0.0231	hjorth_complexity $\leq$ 0.10	0.0077
	acc_y_mean	-0.0224	0.21 $<$ hr_mean $\leq$ 0.27	0.0073
Obstructive (4)	acc_y_mean	0.0948	acc_z_mean $>$ 0.87	0.0181
	kurt_ratio	0.0309	hr_mean $\leq$ 0.16	-0.0167
	acc_x_energy	0.0304	ppg_kurtosis $\leq$ 0.01	-0.0126
	ppg_kurtosis	0.0229	kurt_ratio $\leq$ 0.00	-0.0109
	acc_z_mean	0.0210	acc_x_energy $\leq$ 0.03	0.0072

4.3 System Integration Results

To demonstrate the functional integration of the proposed sleep analysis system, a set of complete test cases was executed, combining full-night recordings with both sleep stage and apnoea event predictions. The pipeline, running fully on the backend, processes input signals, extracts features, performs classification using two independent models, and generates interpretability outputs that are rendered in the frontend.

Table 4.16 lists the main RESTful endpoints exposed by the backend. Each endpoint corresponds to a specific module of the system and is responsible for performing a core operation, including sleep stage prediction, apnoea classification, signal segmentation and explainability. This modular design ensures that components can be updated independently without disrupting the full pipeline, a critical aspect for future scalability and clinical adaptation.

Table 4.16: RESTful API endpoints provided by the backend.

Endpoint	HTTP Method	Description
`/sleep_stage`	POST	Predicts sleep stages from input signal
`/sleep_apnea`	POST	Classifies apnoea events in time windows
`/xai_local_sleep`	POST	Returns SHAP explanations for one sleep segment
`/xai_local_apnea`	POST	Returns SHAP explanations for one apnoea segment
`/signal_sleep`	POST	Extracts raw PPG signal for a selected sleep segment
`/signal_apnea`	POST	Extracts PPG and accelerometer signals for apnoea analysis

The system successfully returns time-aligned predictions for sleep stages and apnoea events, ensuring that both phenomena can be analysed within the same temporal frame. This temporal alignment is particularly important, since sleep disorders often manifest through interactions between disordered breathing and altered sleep architecture. For example, repeated apnoea events typically fragment sleep and reduce the amount of restorative deep and REM sleep. By presenting both outputs together, the system allows users to visually associate respiratory disturbances with transitions between stages, providing a richer context than isolated predictions. Figures 4.4 and 4.5 illustrate example outputs of the frontend: apnoea events displayed across the night (Figure 4.4) and a hypnogram-like representation of the predicted stages (Figure 4.5). Although these figures serve as representative examples of the system's functionality rather than validated clinical reports, they highlight the potential of integrated outputs for exploring relationships between apnoea burden and sleep structure.

Local interpretability is also incorporated into the system for both tasks. Figure 4.6 presents an example of a local explanation applied to a single sleep segment. For this prediction, the system identifies which features contributed most positively and negatively to the assigned class. This level of detail is valuable for users such as clinicians or researchers, who can immediately understand whether the model's decision relied on physiologically meaningful variables (e.g., heart rate variability or amplitude of PPG peaks) or on less relevant fluctuations. Such transparency is essential for building trust in automated predictions and facilitates error analysis when misclassifications occur. Once again, the figure illustrates a representative output rather than a definitive conclusion for a specific subject.

The frontend also supports combined visualisation of raw signals and interpretability outputs, offering a more holistic view of the data. As shown in Figure 4.7, the input signals (PPG and accelerometer) are displayed alongside the explanatory factors identified by the model. This alignment allows users to inspect whether sudden changes in signal morphology coincide with the features highlighted by the model. For instance, abrupt variations in PPG amplitude or accelerometer activity may correspond to model-identified markers of wakefulness or respiratory disturbance. This integration of raw and explanatory data makes it easier to verify consistency between the physiological evidence and the model's reasoning, adding an additional layer of validation.

At the global level, the system also aggregates interpretability results across the entire dataset. Figure 4.8 shows an example of global feature importance obtained from SHAP values, highlighting the features that most consistently influenced classification outcomes. These aggregated insights are useful for identifying which physiological variables the model

considers generally most discriminative. For example, dominance of inter-beat interval variability might suggest strong reliance on autonomic nervous system dynamics, while prominence of PPG-derived features could reflect sensitivity to vascular tone changes. Such global summaries can inform model refinement, guide feature engineering, and help clinicians understand the typical drivers behind automated classifications.

In summary, these examples demonstrate that the backend and frontend modules are not only functionally integrated but also capable of generating outputs that combine predictive accuracy with interpretability. The alignment of apnoea and sleep stage predictions enables joint analysis of respiratory and neurophysiological dimensions of sleep, while the inclusion of local and global interpretability provides transparency at both the single-prediction and dataset-wide levels. Together, these capabilities illustrate the practical functionality of the proposed architecture and its potential applicability to real-world scenarios of sleep monitoring and personalised healthcare.

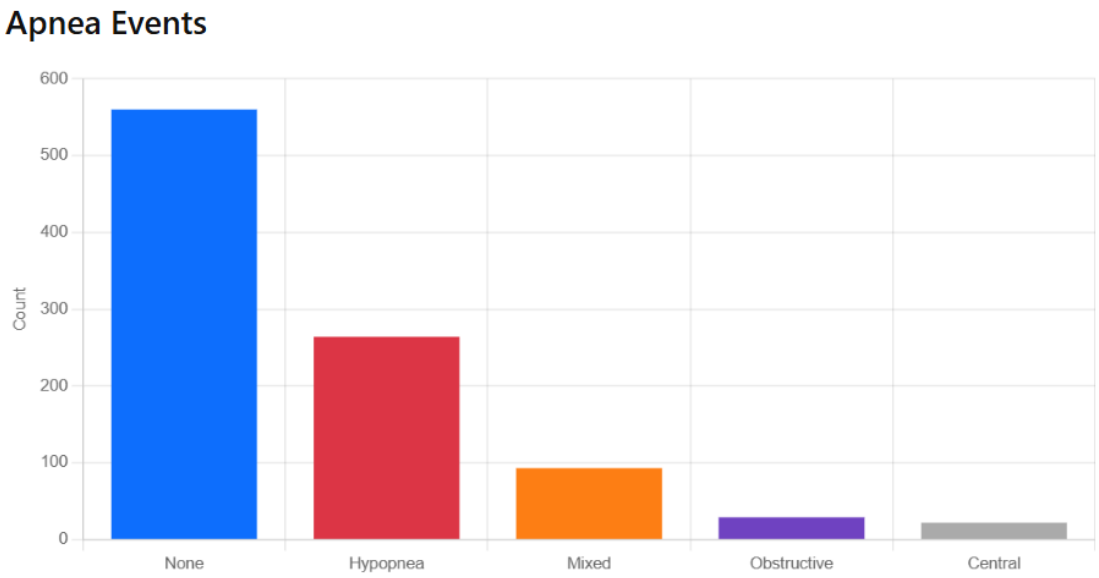

Figure 4.4: Predicted apnoea events over time.

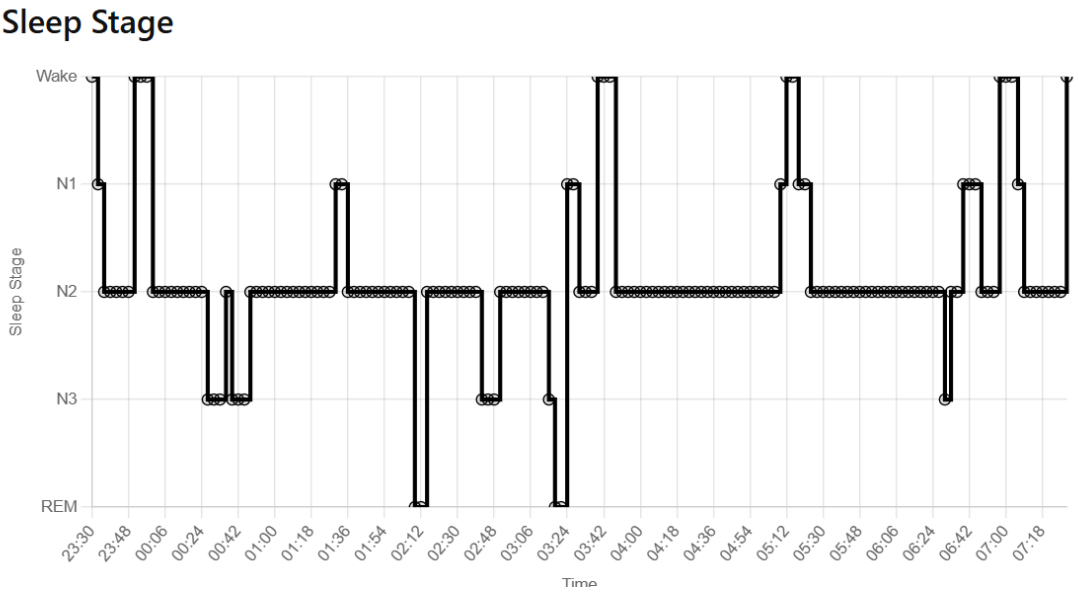

Figure 4.5: Predicted sleep stages over time.

Figure 4.6: Local SHAP and LIME explanations for a specific prediction.

Figure 4.7: Overlay of physiological signals and local XAI outputs.

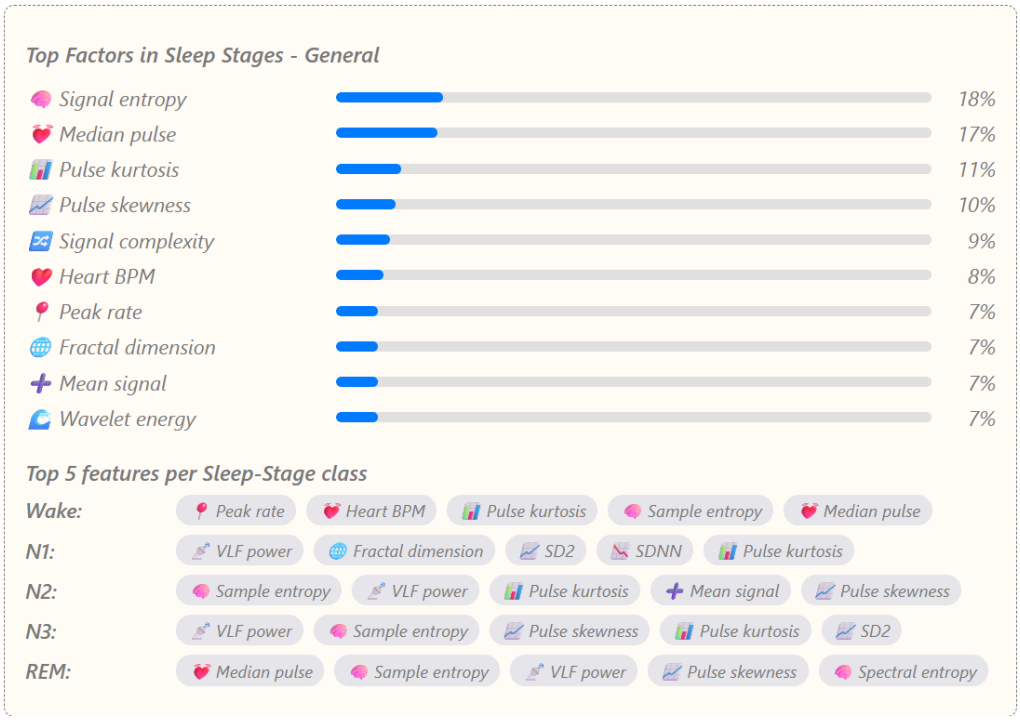

Figure 4.8: Global feature importance for sleep stage classification (SHAP values).

4.4 Synthesis of Results and Key Findings

Using wrist-derived photoplethysmography (PPG) alone, for sleep stage classification the proposed Random Forest (RF) pipeline achieved an overall accuracy of 72.2% and a Cohen's κ of 0.65 on the independent test set. Performance was highest for Wake, REM and deep-sleep (N3) stages (all $F_1 > 0.77$); the transitional N1 stage remained the main source of confusion. Three factors frame these results:

(i) **Task granularity** — The model distinguishes the five canonical stages (Wake, N1, N2, N3 and REM). Many wrist-worn studies collapse the problem to two or three classes (e.g. Wake/Sleep or Wake+REM/NREM), often reporting accuracies above 80% when fusing PPG and accelerometry (Kotzen et al. 2023; Olsen et al. 2024). When those same models are evaluated on this finer taxonomy, performance typically falls below 70% (Silva et al. 2024).

(ii) **Sensor modality** — Adding accelerometry, skin temperature or ECG generally yields 5–15 percentage-point improvements over PPG-only solutions (Peraza et al. 2020; Topalidis et al. 2023). The present study deliberately restricts itself to a single optical channel to keep hardware cost and energy usage compatible with commodity smart-bands.

(iii) **Model complexity** — Deep architectures such as SleepPPG-Net (temporal CNN) report 84% accuracy after extensive pre-training and transfer learning (Kotzen et al. 2023). The lightweight RF presented here attains a comparable macro κ while reducing computational cost by roughly two orders of magnitude relative to embedded CNN inference (Olsen et al. 2024).

These observations show that classical ensemble learners, paired with carefully engineered cardiovascular features, still provide an attractive accuracy–cost trade-off for fine-grained staging on resource-constrained wearables.

For the five-class classification of *none*, central apnoea, obstructive apnoea, mixed apnoea and hypopnoea, the best model (LightGBM) reached 62.4% accuracy and $\kappa = 0.52$. Central and obstructive events were detected with F_1 scores of 0.71 and 0.68, respectively, whereas mixed apnoea and hypopnoea remained challenging ($F_1 < 0.50$). A comparison with other wrist-based studies is summarised below:

(i) **Binary screening** — OSA screening with ECG or PPG routinely exceeds 90% accuracy, for example, 1D-SEResGNet on single-lead ECG (Yang et al. 2022) and ApSense on PPG (Choksatchawathi et al. 2024). These systems, however, do not differentiate phenotype, a clinically important gap for treatment titration.

(ii) **Multi-type systems** — Detectors that combine respiratory inductance or SpO_2 with motion signals achieve 65–75% accuracy (A. John, Cardiff, and D. John 2021; Ye et al. 2021), but require extra sensors and nightly calibration.

(iii) **PPG + ACC pipeline** — The present approach relies solely on wrist-PPG and tri-axial accelerometry yet attains comparable κ, highlighting the diagnostic value of posture and micro-movement features identified here and in prior actigraphy research (S. Chakraborty, Aich, and Kim 2021).

These findings reinforce the central role of cardiorespiratory coupling captured by PPG and the complementary contribution of body dynamics to apnoea phenotyping.

For **sleep staging**, $\kappa = 0.65$ lies in the Landis–Koch (Landis and Koch 1977) "substantial agreement" band and is close to the ~ 0.70 typically achieved by two expert scorers, so it is fit for home screening and longitudinal trend tracking. For **sleep apnoea classification**, $\kappa = 0.52$ falls into the "moderate agreement" band: good enough to flag high-risk nights and prioritise further testing, yet below the reliability usually expected for stand-alone diagnosis. Both models therefore serve well as low-cost triage tools.

Explainable-AI analyses (SHAP and LIME) further showed that model decisions align with established sleep physiology: heart-rate variability and waveform morphology dominated sleep-stage predictions, whereas movement intensity and posture changes drove apnoea detection. Prioritising interpretability alongside accuracy ensures transparency, essential for real-world adoption in health-care settings.

End-to-end validation on full-night recordings confirmed system robustness. The backend processed raw signals, executed the classifiers and served synchronised outputs via a RESTful API, while the frontend visualised predictions, signal windows and interpretability overlays in a single interface. Unlike many studies that report algorithmic performance in isolation, this framework demonstrates a modular, deployment-ready pipeline.

In summary, the proposed system balances performance, interpretability and practicality, positioning it as a competitive and transparent solution for wearable-based sleep analysis and as a solid foundation for future clinical and technological extensions.

4.5 Ethical and Social Challenges

In the context of the analysis and prediction of sleep patterns using Artificial Intelligence (AI), there are significant technological and social challenges to address. These challenges are largely driven by ethical considerations, regulatory frameworks such as the European Union's AI Act, and data protection laws such as the General Data Protection Regulation (GDPR). Ensuring the responsible use of AI in healthcare-related applications is crucial for maintaining trust, privacy, and fairness, especially when sensitive health data is involved (Gerke, Minssen, and Cohen 2020).

To visualise the interplay between these ethical pillars, Figure 4.9 presents a schematic overview of the key elements: data privacy (GDPR), regulatory compliance (European AI Act), transparency (XAI), and fairness (accountability). This diagram helps to underscore how these components together form the foundation for ethically compliant AI in sleep analysis. Even though this work relies exclusively on publicly available datasets (MESA and DREAMT), both databases were collected under institutional ethical approval, with explicit informed consent obtained from all participants.

Throughout this section, we will discuss in detail how each of these requirements is met, and the measures that have been or will be implemented to ensure their ongoing compliance.

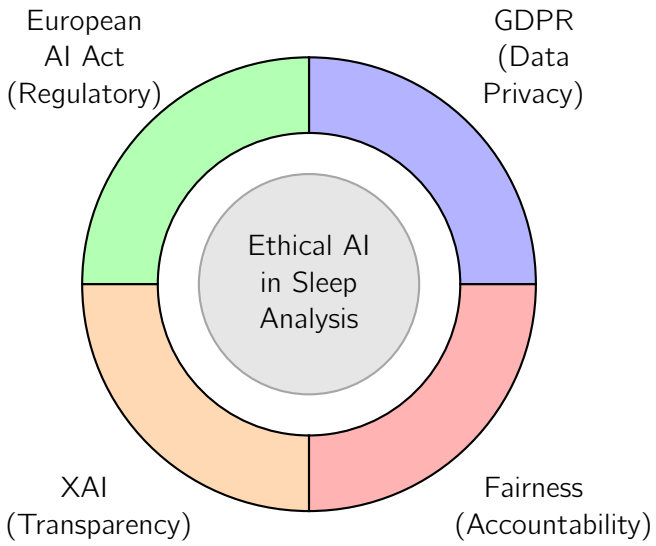

Figure 4.9: Four pillars (GDPR, European AI Act, XAI, Fairness) for ethically compliant AI in sleep analysis.

4.5.1 Ethical Considerations in Sleep Data Analysis

The use of artificial intelligence (AI) to analyse sleep data presents several ethical considerations, which are crucial to ensure that the technology is applied responsibly and for the benefit of society. One of the main ethical issues revolves around data privacy. Given that sleep data contains sensitive health information, it is essential that all practices related to the collection, storage, and processing of such data respect individuals' rights.

The integration of multimodal signals, such as PPG and accelerometry, offers improved accuracy in identifying sleep patterns and apnoea events. However, it also raises ethical concerns related to over-inference. When combined, these signals may unintentionally reveal additional information such as movement patterns, restlessness, or potential neurological issues that go beyond the intended scope of analysis. To address this, the principle of data minimisation must be respected, ensuring that only relevant features are extracted and that user profiling is explicitly avoided.

Moreover, ensuring fairness in AI applications is a significant ethical concern. Sleep data analysis can be susceptible to biases, especially if AI models are trained on limited or non-representative datasets. To ensure equitable outcomes, it is critical that the data used to train the model includes a wide variety of conditions and demographic profiles, thus preventing the system from producing biased or inaccurate results that could negatively impact certain groups (e.g., minority or vulnerable populations) (Hanna et al. 2024). This study mitigates potential algorithmic bias by selecting demographically diverse participants from both MESA and DREAMT datasets. The cohorts include variation across age, gender, ethnicity, and health status, which reduces the risk of under-representation. Nonetheless, to ensure sustained fairness, it is essential that models are periodically re-evaluated with real-world data, particularly from under-represented groups or wearable devices with differing sensor characteristics (e.g., skin tones affecting PPG quality).

Transparency is also a central ethical issue. While AI systems can generate valuable predictions, these outcomes can often be difficult to interpret due to the opaque nature of many models. To ensure that healthcare professionals can trust and effectively use AI results, it

is necessary to make the decision-making process of the model more transparent, explaining how each prediction is made and what factors influenced the decision (Ahmad, Eckert, and Teredesai 2018). This interpretability is essential in healthcare, where decisions based on AI can have a significant impact on patient care and outcomes. To address this need for transparency, explainable AI (XAI) techniques, such as SHAP and LIME, have emerged as valuable tools for making complex models more interpretable. These methods help uncover how specific input features, like heart rate variability and activity patterns, contribute to the model's predictions (Nam et al. 2024; Osathitporn et al. 2023). By revealing the rationale behind each decision, XAI not only fosters trust among clinicians but also supports personalised interventions and model refinement. However, this increased interpretability often comes at a cost. The so-called "glass box versus crystal ball" dilemma (Hulsen 2023) highlights the trade-off between transparency and accuracy: more interpretable models tend to be simpler and potentially less accurate, while highly accurate models are often opaque. As illustrated in Figure 4.10, finding the right balance between these extremes is especially critical in healthcare, where even minor errors can have serious consequences.

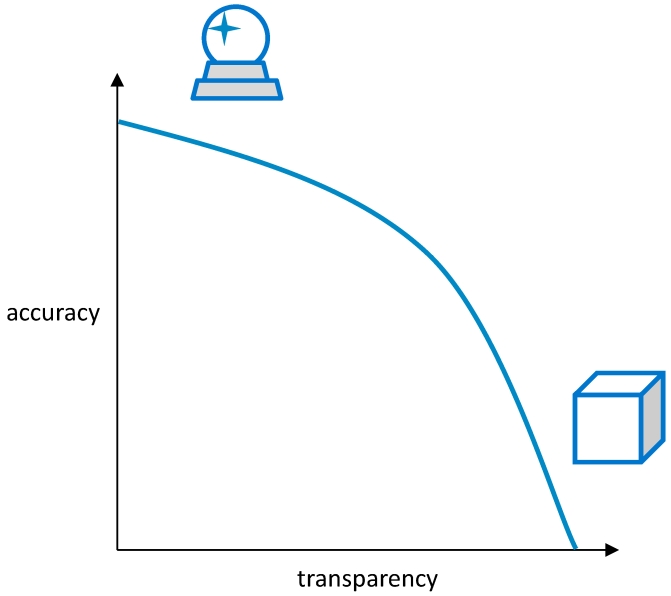

Figure 4.10: Relationship between transparency and accuracy in predictive models, moving from a "crystal ball" (higher accuracy) to a "glass box" (higher transparency). From (Hulsen 2023).

Finally, the autonomy of patients must be preserved, ensuring they can contest automated decisions that might affect their health. AI should be seen as a tool to assist healthcare professionals, not replace human judgment. The ultimate responsibility for medical decisions should remain with qualified professionals, ensuring that patients' ability to influence decisions about their care is not compromised (Gerke, Minssen, and Cohen 2020).

These ethical considerations are crucial to ensure that AI not only improves healthcare outcomes but does so in an ethical manner, respecting patients' rights and promoting fairness and transparency in the handling of sensitive data. However, there are also established ethical and legal frameworks that regulate these issues, and it is essential to ensure compliance with such regulations to maintain accountability and protect patient welfare.

Regulatory Compliance in AI-Driven Sleep Analysis As artificial intelligence continues to advance in healthcare applications, ensuring compliance with legal and ethical regulations is paramount. AI-driven sleep analysis systems must adhere to stringent frameworks to safeguard patient rights, data privacy, and the responsible use of AI technologies.

In the European Union, two major regulatory frameworks govern the deployment of AI in healthcare: the European AI Act and the General Data Protection Regulation (GDPR). The European AI Act categorizes healthcare AI as a high-risk application, imposing strict requirements on transparency, accountability, and bias prevention. Meanwhile, the GDPR establishes fundamental principles for handling sensitive personal data, including explicit consent, data minimization, and the right to contest AI-driven decisions.

This outlines the key legal requirements relevant to AI-based sleep analysis solutions, detailing the measures taken to align with both the AI Act and GDPR. By adhering to these regulations, AI models can be deployed in a manner that prioritizes safety, fairness, and ethical integrity, ultimately fostering trust in AI-driven healthcare technologies.

Compliance with the European AI Act

The European AI Act, introduced by the European Commission, aims to ensure the safe and ethical deployment of AI technologies across the EU. It categorizes AI applications into various risk levels, with healthcare-related AI models falling under the "high-risk" category due to the sensitive nature of health data . This mandates the implementation of stricter regulations and requirements to ensure AI models in healthcare are developed, deployed and monitored responsibly Parliament and European Union 2024. Figure 4.11 illustrates the risk-based framework established by the AI Act, highlighting the four main levels: unacceptable risk, high risk, limited risk, and minimal risk (AP4AI Initiative 2024).

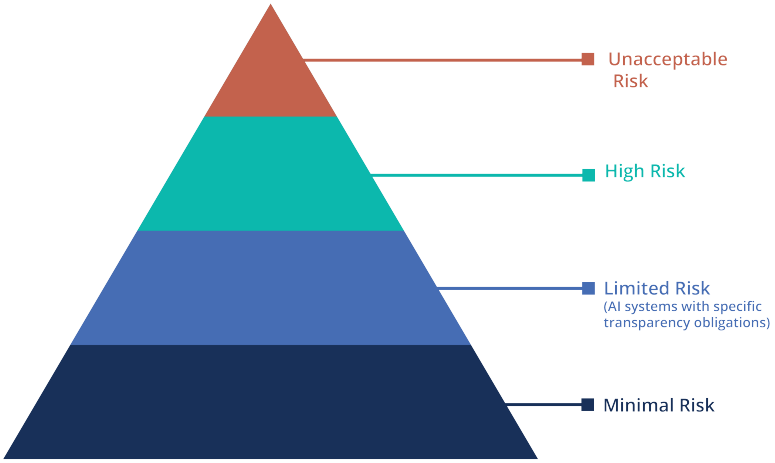

Figure 4.11: Risk-based classification of AI systems according to the European AI Act: from minimal risk at the base to unacceptable risk at the top.

The importance of the AI Act in the context of healthcare systems from the high potential impact, both positive and negative, of AI-driven decisions on individuals' health, safety, and fundamental rights. Unlike other sectors, healthcare involves deeply personal and sensitive data, and decisions made by AI systems can directly affect diagnoses, treatment plans, and

overall wellbeing. As such, the legislation seeks not only to prevent harm but also to promote trustworthy and human-centric AI that respects patients autonomy, ensures transparency in medical decision-making, and avoids biases that could disproportionately affect vulnerable populations. In this regard, the AI Act plays a pivotal role in reinforcing public trust and ensuring that innovation in healthcare remains aligned with ethical and legal standards.

In this study, the solution adheres to the European AI Act by implementing several key practices, ensuring transparency, fairness, and accountability throughout the entire AI model lifecycle.

- **Risk Management and Human Oversight**: In compliance with the requirements of the AI Act, high-risk systems must implement robust risk management measures. The model has been developed as a tool to support the work of healthcare professionals, rather than as a replacement. It provides recommendations and insights to assist in the analysis of conditions such as sleep apnoea, ensuring that the final decision remains the responsibility of the healthcare professional. Furthermore, the future integration of the model into wearable devices is planned, enabling continuous data monitoring for early detection. However, all results presented will include a clear warning emphasizing the need for validation and consultation with a qualified healthcare professional, ensuring that AI applies to principles of safety, ethics, and human oversight.
- **Data Quality and Bias Prevention**: The AI Act highlights the need to train AI models on high-quality, diverse, and representative data to reduce bias. In this study, the training datasets include a broad range of demographic factors such as age, gender, and ethnicity, ensuring that predictions remain fair and accurate for all groups. Furthermore, the physiological signals were preprocessed through filtering techniques and validated to ensure data integrity and reliability prior to model training.
- **Transparency and Documentation**: One of the fundamental principles of the AI Act is transparency, which is crucial for building trust in AI systems. To comply with this requirement, the solution provides comprehensive documentation on the AI model's development, training data, and performance evaluation. This includes clear explanations of how the data is processed, what features are being used, and how predictions are made. This transparency ensures that all clinicians and patients are fully informed about the AI system's capabilities and limitations.
- **Accountability and Monitoring**: The AI Act places significant emphasis on accountability, requiring high-risk AI systems to undergo regular monitoring to ensure compliance with safety and ethical standards. The models utilised will be subject to continuous oversight to validate their performance, with particular focus on their ability to generalise effectively to new, unseen data. Generalisation is a critical factor in ensuring that the models maintain accuracy and fairness when applied to diverse populations and varying conditions, rather than being limited to specific datasets or scenarios. This capability is especially important in the healthcare domain, where individual variability can greatly influence outcomes (Goetz et al. 2024). If issues such as reduced accuracy or fairness are detected, corrective measures will be implemented promptly to ensure compliance with the AI Act and uphold the reliability of the models.
- **Logging and Record-Keeping**: To ensure auditability and traceability, the integrated system developed in this study includes mechanisms for frequent logging into a dedicated log file. These logs capture key information such as input data, prediction

outcomes, and relevant explanations generated by the interpretability tools. By maintaining a continuous record of the model's operation, the system ensures that, if necessary, a detailed audit can be conducted in the future. This practice aligns with the AI Act's requirements for accountability and post-hoc analysis of high-risk AI systems.

- **Robustness, Accuracy, and Cybersecurity**: In accordance with the AI Act, high-risk AI systems must demonstrate technical robustness and resilience to ensure safe deployment in real-world settings. This study includes measures to evaluate the accuracy of the model across different subsets of the population and test its reliability under varying conditions. Although this research is conducted in a controlled environment, future developments will incorporate security measures to protect the model against adversarial attacks and unauthorised access, especially in scenarios where the AI system is deployed on wearable devices. These steps are essential to ensure that the model remains secure, stable, and trustworthy over time.
- **Transparency of AI Model Outputs**: To align with the transparency requirements of the AI Act, the solution integrates mechanisms that make AI-driven predictions more interpretable. The use of SHAP and LIME ensures that clinicians can understand the rationale behind each prediction, providing clarity on which features influenced the model's decision. This interpretability is particularly important in healthcare, where clinicians need to make informed decisions based on the outputs provided by the AI system.

By integrating these key principles of the European AI Act into the development and deployment of the proposed AI solution, this study ensures that the system not only achieves high technical performance but also operates within a framework of legal and ethical responsibility. This comprehensive approach enhances the reliability, transparency, and acceptability of AI in clinical settings, laying the groundwork for its safe adoption in real-world healthcare applications. Ultimately, aligning with the AI Act is not merely a legal requirement but a necessary step toward fostering trust and ensuring that innovation serves the best interests of patients and society.

General Data Protection Regulation (GDPR) Compliance

The General Data Protection Regulation (GDPR) (Parliament and European Union 2016) will be a key consideration when handling sensitive sleep data. Given that sleep data involves the collection of personal health information, it will be crucial that all processing of such data complies with the GDPR's principles, particularly regarding consent, data minimisation, and the right to be forgotten.

Under the GDPR, individuals will be required to provide explicit consent for their data to be collected and processed. In the context of AI-based healthcare systems, patients must be fully informed about how their data will be used, whether it will be stored, and for how long.

The right to access and correct data will be another important aspect of GDPR compliance. Patients will have the ability to request access to their data, verify its accuracy, and correct any inaccuracies. This will be particularly relevant in the context of sleep disorder detection, where incorrect data could lead to misdiagnosis or improper treatment recommendations.

The datasets used for training the AI models were carefully selected to ensure full compliance with GDPR regulations. This includes the DREAMT and MESA Sleep datasets, both of which adhere to strict ethical and data protection standards. The DREAMT dataset was collected under the supervision of the Duke Health Institutional Review Board (IRB

#Pro00108961), with written informed consent that explicitly allows the sharing of de-identified data. In the publicly available version, all direct identifiers have been removed and timestamps have been time-shifted to protect participant identities, ensuring full anonymisation in line with GDPR requirements.

The MESA Sleep dataset, although not publicly available, is accessible through formal access requests and is governed by robust data usage agreements. These agreements ensure that data are used exclusively for approved health-related research purposes, with explicit participant consent and proper de-identification measures. By selecting these datasets, the study ensures that no personally identifiable information is used and that all data processing is performed ethically and legally, respecting participants' privacy and consent.

Additionally, only the minimum necessary data from these datasets will be used to train the AI models, in line with the GDPR's principle of data minimisation. By using only the relevant data for sleep stage classification and apnoea detection, unnecessary or excessive data will be avoided. This approach will ensure that privacy is respected, as no sensitive data beyond what is required for the specific task will be used.

Furthermore, AI systems that involve automated decision-making, such as classifying sleep stages or detecting apnoea, must allow individuals to challenge decisions made solely by AI. Patients should be allowed to contest any automated conclusions that may impact their health and well-being. Throughout the process, it will be ensured that individuals rights under the GDPR, such as the right to access, correct, and contest decisions made by AI, are protected. The datasets that will be used will be transparent in their sources and the methods employed for obtaining consent from participants, thus ensuring compliance with the regulation's requirements for handling sensitive personal health data.

In summary, the careful selection and usage of pre-existing datasets for this AI-driven solution will ensure that the system remains fully compliant with GDPR, upholding privacy rights and ethical standards while facilitating accurate and effective healthcare outcomes.

4.5.2 Future Perspectives and Recommendations

Beyond the current ethical and regulatory challenges, the long-term success of AI-driven sleep analysis systems will depend on continued social engagement and the development of strategies that foster trust and adaptability. In this context, it is essential to establish clear communication channels between developers, healthcare professionals, and end users. Educating users about the capabilities and limitations of AI tools can help mitigate misconceptions and promote informed decision-making (Coeckelbergh 2019).

Moreover, continuous monitoring of model performance is vital to address any emerging biases or discrepancies as the system interacts with increasingly diverse populations. This includes the regular reassessment of training datasets to ensure they remain representative of real-world conditions (Verma et al. 2023), as well as the implementation of adaptive learning mechanisms that can update the models in response to new data and shifting societal needs (Watson and Fernandez 2021).

The development of transparent feedback mechanisms, where users and clinicians can report issues or suggest improvements, is also recommended. Such channels not only contribute to the refinement of the technology but also reinforce accountability, ensuring that the AI system evolves in alignment with ethical standards and public expectations (Khan et al. 2021).

Figure 4.12 illustrates a continuous improvement cycle, where feedback from clinicians and end users is collected to inform model retraining. Once refined, the updated models are verified, deployed, and subsequently monitored in real-world scenarios. This cyclical process ensures that the system remains adaptable to evolving conditions and user needs, thereby promoting trust, transparency, and long-term sustainability.

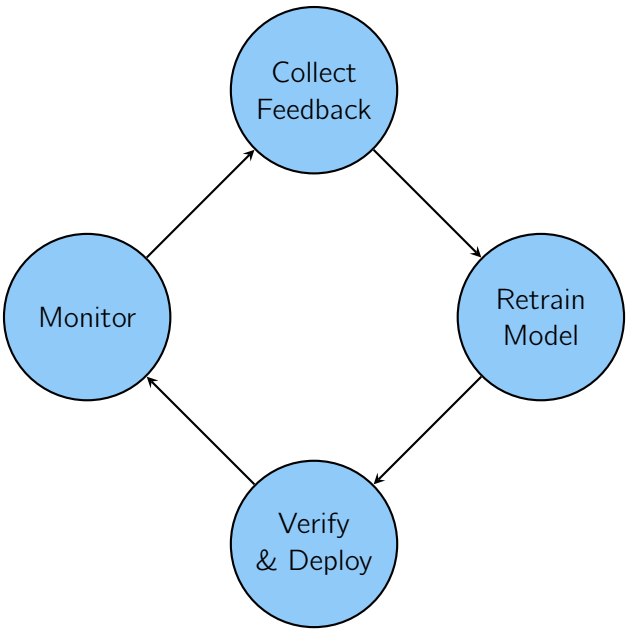

Figure 4.12: Iterative cycle for AI model refinement: from collecting feedback to monitoring performance.

Such a cyclical refinement process relies on the active input of multiple stakeholders, data scientists, clinicians, ethicists, and legal experts, to ensure that each iteration is both technically robust and ethically sound. It is essential that any modifications arising from user feedback or the identification of new biases are rigorously validated against established clinical standards and relevant legal and ethical guidelines. Moreover, fostering interdisciplinary collaboration is vital to ensure that technological innovations in AI-driven sleep analysis are not only effective from a technical standpoint, but also socially responsible and sustainable over time. By addressing these social and cultural dimensions, the integration of AI in sleep health can be realised in a manner that is both ethically sound and broadly endorsed by the community.

In conclusion, the successful integration of AI in sleep health requires more than just technological sophistication, it demands a commitment to ethical principles, legal compliance, and inclusive collaboration. By prioritising transparency, fairness, continuous learning, and active stakeholder involvement, it becomes possible to design AI systems that not only enhance clinical capabilities but also empower patients, protect rights, and foster societal trust. The long-term sustainability of AI in healthcare will ultimately depend on this balanced and human-centred approach.

Chapter 5

Conclusions

This chapter summarises the contributions of the thesis, reflects on how the research objectives have been addressed, and outlines potential directions for future work. The discussion is structured in three parts: a contextualisation of the study, an evaluation of the research goals and their fulfilment, and a perspective on next steps and ongoing developments.

5.1 Contextualisation of the Study

The work is positioned at the intersection of rapidly evolving wearable technologies and the growing integration of artificial intelligence in healthcare. Traditionally, sleep monitoring has relied on polysomnography, an accurate yet invasive and resource-intensive method. Wrist-worn devices, in contrast, offer a non-invasive, continuous, and cost-effective alternative for collecting physiological and behavioural data (M. P. Lee et al. 2024; Roberts et al. 2020).

In response to this paradigm shift, the present work developed an integrated framework that combines multimodal data (PPG and accelerometry) with machine learning and explainable AI (XAI) techniques. The system supports the automatic classification of sleep stages and the detection of sleep-disordered breathing events, while also providing interpretable outputs that facilitate clinical understanding and user trust (Moreno-Pino et al. 2022; Papini et al. 2020).

By addressing limitations observed in prior research, such as over-reliance on isolated metrics and limited generalisability to diverse populations, this thesis contributes to the development of scalable and adaptive approaches to sleep monitoring (S. Chen et al. 2015; Hamza et al. 2023). In particular, the application of XAI methods (e.g., SHAP and LIME) demonstrated that model predictions can be explained in a way that aligns with established physiological knowledge, supporting transparency and compliance with regulatory standards such as GDPR and the European AI Act (Gerke, Minssen, and Cohen 2020).

In summary, the work presented in this thesis establishes a solid methodological foundation for personalised, interpretable, and wearable-based sleep monitoring. It bridges the gap between algorithmic performance and clinical applicability, laying the groundwork for real-world implementations. Future research will focus on large-scale validation across heterogeneous datasets, on-device deployment for real-time monitoring, and longitudinal studies to evaluate health outcomes in practical settings.

5.2 Objectives and Their Fulfilment

The main objective of this thesis was to develop an integrated framework for personalised sleep monitoring using data from wrist-worn wearable devices and ensemble artificial intelligence techniques. The five research objectives outlined in Chapter 1 (OB1–OB5) guided the work and are summarised below, along with how each was addressed.

OB1: Investigate state-of-the-art methods for using wrist-worn wearable device data in sleep monitoring. A systematic review of current approaches was conducted and presented in Chapter 2, covering data sources, signal processing techniques, machine learning models, and evaluation strategies. This review identified key limitations in the field, including the underutilisation of explainable AI and the lack of generalisability across populations.

OB2: Identify and characterise specific sleep patterns, behaviours, and influencing factors derived from wearable data. Physiological and behavioural features were extracted from PPG and accelerometry signals, focusing on heart rate variability, waveform morphology, and motion intensity. These features enabled the characterisation of distinct sleep stages and respiratory events across multiple datasets.

OB3: Develop and apply ensemble AI models capable of integrating multiple data streams to generate high-quality, personalised insights into sleep quality. The pipeline incorporated ensemble models such as Random Forest, applied to both sleep stage classification and sleep apnoea detection. These models leveraged multimodal inputs and demonstrated performance metrics comparable to, or surpassing, those reported in recent literature.

OB4: Explore the use of explainable AI (XAI) techniques to interpret wearable sleep data, ensuring transparency and trustworthiness. SHAP and LIME were applied to both global and local predictions. The results demonstrated that model outputs could be explained through physiologically meaningful features, such as heart rate variability and movement patterns. These explanations were consistent across methods and support the interpretability of the system.

OB5: Validate and evaluate the developed framework using real-world wrist-worn wearable datasets. The framework was tested on multiple publicly available datasets, including MESA, and DREAMT. Models were evaluated using standard performance metrics (accuracy, F_1, Cohen's κ), and confusion matrices and ROC curves provided detailed insight into class-wise performance. The system demonstrated robustness across datasets and tasks.

In summary, all five research objectives were successfully addressed. The framework integrates multimodal feature extraction, ensemble learning, and explainability in a unified and modular architecture. The results support the feasibility of using wrist-worn devices for clinically relevant sleep monitoring, laying the groundwork for future deployment in real-world settings.

5.3 Contributions

The work developed in this thesis has resulted in a set of tangible and measurable contributions, spanning methodological innovations, system integration, and scientific dissemination. These contributions not only advance the state of the art in wearable-based sleep monitoring but also provide a foundation for future research and practical deployment. They can be summarised as follows:

- **Peer-reviewed Publication on Apnoea Detection:** The multi-class apnoea detection framework, based on lightweight tree ensembles and optimised feature sets, was published in the international, peer-reviewed conference paper *"Lightweight Tree Ensembles with Optimized Features for Five-Class Sleep Apnea Stratification"* (EPIA 2025). This contribution demonstrates that interpretable and computationally efficient models can achieve competitive accuracy, supporting the feasibility of real-world implementation on wearable platforms.
- **Forthcoming Journal Article on Sleep Stage Classification:** The author has co-authored the journal article *"Comprehensive Analysis of Machine Learning Models for Five-Class Sleep Stage Classification Using PPG Signals"*, submitted to *Scientific Reports*. This contribution consolidates the empirical findings of the thesis by systematically benchmarking a wide range of machine learning models, identifying their strengths and limitations, and positioning the proposed framework within the broader landscape of wearable-based sleep research. The submission to a high-impact, peer-reviewed journal underlines the scientific relevance and maturity of the work.
- **Manuscript in Preparation on XAI and System Architecture:** A third article is currently being prepared for submission to a major international conference. It focuses on the integration of explainable AI methods (SHAP and LIME) and the architecture developed in this thesis.

Taken together, these scientific outputs showcase the originality, robustness, and applicability of the research. They provide a regulation-ready foundation for large-scale, personalised sleep and health applications, bridging the gap between experimental machine learning models and clinically meaningful solutions.

5.4 Next Steps and Future Work

While the results presented in this thesis demonstrate the feasibility and effectiveness of using wrist-worn wearable data for automatic sleep stage classification and apnoea detection, several areas remain open for further investigation and development.

- **Cross-Dataset Generalisation and External Validation:** Although the framework was evaluated on multiple datasets, further validation on larger and more diverse populations is required. Future work should focus on cross-cohort experiments involving different age groups, ethnic backgrounds, and clinical profiles, to ensure generalisability and reduce dataset-specific bias.
- **Temporal Modelling and Longitudinal Analysis:** The current models rely on window-based predictions without explicitly modelling long-range temporal dependencies. Incorporating sequence-based architectures, such as temporal convolutional networks or transformers, could improve the modelling of sleep dynamics and transitions between stages.
- **Integration of Additional Sensors and Multimodal Signals:** Future work could explore the integration of additional biosignals, such as nocturnal oximetry, respiratory effort or skin temperature, to improve the detection of ambiguous events like mixed apnoea and refine the characterisation of sleep quality.
- **On-Device Deployment and Energy Efficiency:** To support real-time applications in wearable devices, further efforts are needed to optimise model size, inference speed and

energy consumption. Techniques such as model quantisation, pruning or edge-oriented inference strategies (e.g., TinyML) may enable practical deployment on resource-constrained platforms.

- **Clinical Validation and Human-Centred Evaluation:** Collaborating with clinical partners will be essential to assess the usability, interpretability, and clinical value of the system in practice. Prospective studies in real-world settings will help determine whether the insights provided by the system can support clinical workflows, early diagnosis, or personalised interventions.
- **Expansion of Explainability Features:** Although the current system integrates SHAP and LIME to provide transparency, future work may consider exploring user-adaptive explanations or interactive visualisation tools to enhance understanding by non-technical users, including patients and healthcare professionals.

In summary, while this thesis has delivered a working and interpretable system for wearable-based sleep analysis, future work should focus on enhancing robustness, broadening applicability, and bridging the gap between technical development and clinical adoption. Addressing these challenges will be crucial to enable the large-scale deployment of transparent and personalised sleep monitoring tools.

Bibliography

Ahmad, Muhammad Aurangzeb, Carly Eckert, and Ankur Teredesai (2018). "Interpretable Machine Learning in Healthcare". In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '18. Washington, DC, USA: Association for Computing Machinery, pp. 559–560. isbn: 9781450357944. doi: 10.1145/3233547.3233667. url: https://doi.org/10.1145/3233547.3233667.

Ahn, E. et al. (2023). "Elevated prevalence and treatment of sleep disorders from 2011 to 2020: a nationwide population-based retrospective cohort study in Korea". In: *BMJ Open* 14. doi: 10.1136/bmjopen-2023-075809.

Anggraini, Navira Anggraini et al. (2023). "Monitoring SpO2, Heart Rate, and Body Temperature on Smartband with Data Sending Use IoT Displayed on Android (SpO2)". In: *Jurnal Teknokes*. url: https://api.semanticscholar.org/CorpusID:268491280.

Anusha, A.S. et al. (2022). "Electrodermal activity based autonomic sleep staging using wrist wearable". In: *Biomedical Signal Processing and Control* 75, p. 103562. issn: 1746-8094. doi: https://doi.org/10.1016/j.bspc.2022.103562. url: https://www.sciencedirect.com/science/article/pii/S1746809422000842.

AP4AI Initiative (2024). *The EU AI Act – A Risk-Based Approach*. Accessed: 2025-07-08. url: https://ap4ai.eu/eu-ai-act.

Arias, Juan F. (2022). "Using Data from Wearables for Better Sleep". In: *2022 IEEE International Conference on Digital Health (IEEE ICDH 2022)*. Ed. by SI Ahamed et al. Chap. 0, pp. 37–39. doi: 10.1109/ICDH55609.2022.00014.

Arora, Anshika, Pinaki Chakraborty, and M. P. S. Bhatia (Dec. 2020). "Analysis of Data from Wearable Sensors for Sleep Quality Estimation and Prediction Using Deep Learning". In: *Arabian Journal for Science and Engineering* 45.12, pp. 10793–10812. issn: 2193-567X. doi: 10.1007/s13369-020-04877-w.

Avram, Robert et al. (2019). "Real-world heart rate norms in the Health eHeart study". In: *npj Digital Medicine* 2.1, p. 58. issn: 2398-6352. doi: 10.1038/s41746-019-0134-9. url: https://doi.org/10.1038/s41746-019-0134-9.

Bahrami, Mahsa and Mohamad Forouzanfar (Dec. 2022). "Deep Learning Forecasts the Occurrence of Sleep Apnea from Single-Lead ECG". In: *Cardiovascular Engineering and Technology* 13.6, pp. 809–815. issn: 1869-408X. doi: 10.1007/s13239-022-00615-5.

Bello, Yahuza and Emanuel Figetakis (Apr. 2023). "IoT-based Wearables: A comprehensive Survey". In: arXiv: 2304.09861.

Benedetti, Davide et al. (2022). "Obstructive Sleep Apnoea Syndrome Screening Through Wrist-Worn Smartbands: A Machine-Learning Approach." eng. In: *Nature and science of sleep* 14, pp. 941–956. issn: 1179-1608 (Print). doi: 10.2147/NSS.S352335. url: https://pubmed.ncbi.nlm.nih.gov/35611177/.

Berry, Richard B. et al. (2017). *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Version 2.4. Available from the American Academy of Sleep Medicine. American Academy of Sleep Medicine. Darien, IL.

Bhat, Ashwin and Arijit Raychowdhury (2023). "Explainable ECG Beat Classification On The Edge for Smart, Trustworthy and Low-Power Wearables". In: *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–5. url: https://api.semanticscholar.org/CorpusID:267045400.

Biczuk, Bartosz et al. (2024). "Sleep Stage Classification Through HRV, Complexity Measures, and Heart Rate Asymmetry Using Generalized Estimating Equations Models". In: *Entropy* 26.12. issn: 1099-4300. doi: 10.3390/e26121100. url: https://www.mdpi.com/1099-4300/26/12/1100.

Borazio, Marko et al. (2014). "Towards Benchmarked Sleep Detection with Wrist-Worn Sensing Units". In: *2014 IEEE International Conference on Healthcare Informatics*, pp. 125–134. doi: 10.1109/ICHI.2014.24.

Brutzman, Kathryn et al. (2024). "Integrating AI and ChatGPT in Wearable Devices for Enhanced Abnormal Activity Reporting: A Mixture of Experts Approach". In: *2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI)*, pp. 231–235. url: https://api.semanticscholar.org/CorpusID:272761170.

Calisti, Lorenzo and Emanuele Lattanzi (2024). "Real-Time Energy-Efficient Sensor Libraries for Wearable Devices". In: *IEEE Access* 12, pp. 126006–126018. doi: 10.1109/ACCESS.2024.3430049.

Cantanhede, Lorena Renata Costa et al. (2018). "COMPORTAMENTO DO CONSUMIDOR DE TECNOLOGIA VESTÍVEL: CARACTERÍSTICAS QUE INFLUENCIAM NA INTENÇÃO DE CONSUMO". In: *REAd. Revista Eletrônica de Administração (Porto Alegre)*. url: https://api.semanticscholar.org/CorpusID:172047644.

Chakraborty, Sabyasachi, Satyabrata Aich, and Hee-Cheol Kim (July 2021). "A Novel Sleep Scoring Algorithm-Based Framework and Sleep Pattern Analysis Using Machine Learning Techniques". In: *International Journal of System Dynamics Applications* 10.3, pp. 1–20. issn: 2160-9772. doi: 10.4018/IJSDA.2021070101.

Chen, Jeng-Wen et al. (Mar. 2023). "A Signal Segmentation-Free Model for Electrocardiogram-Based Obstructive Sleep Apnea Severity Classification". In: *Advanced Intelligent Systems* 5.3. doi: 10.1002/aisy.202200275.

Chen, Shuya et al. (2015). "Usability of a Low-Cost Wearable Health Device for Physical Activity and Sleep Duration in Healthy Adults". In: *Proceedings of the 2015 Workshop on Pervasive Wireless Healthcare*. MobileHealth '15. Hangzhou, China: Association for Computing Machinery, pp. 35–38. isbn: 9781450335256. doi: 10.1145/2757290.2757298. url: https://doi.org/10.1145/2757290.2757298.

Chen, Xianda et al. (June 2021). "ApneaDetector: Detecting Sleep Apnea with Smartwatches". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5.2. doi: 10.1145/3463514.

Chen, Xiaoli et al. (2015). "Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA)". In: *Sleep* 38.6. PMC4434554, pp. 877–888. issn: 1550-9109. doi: 10.5665/sleep.4732. url: https://pubmed.ncbi.nlm.nih.gov/25409106.

Chih, Hao-Yi et al. (Jan. 2024). "Multitask Learning for Automated Sleep Staging and Wearable Technology Integration". In: *Advanced Intelligent Systems* 6.1. doi: 10.1002/aisy.202300270.

Chirieleison, Carlos Augusto Marques et al. (dez. 2024). "Tecnologias vestíveis no monitoramento de cardiopatias: uma revisão de literatura". In: *Cuadernos de Educación y Desarrollo* 16.12 Edição Especial, e6483. doi: 10.55905/cuadv16n12-031. url: https://ojs.cuadernoseducacion.com/ojs/index.php/ced/article/view/6483.

Choksatchawathi, Tanut et al. (Oct. 2024). "ApSense: Data-Driven Algorithm in PPG-Based Sleep Apnea Sensing". In: *IEEE Internet Of Things Journal* 11.20, pp. 33915–33926. issn: 2327-4662. doi: 10.1109/JIOT.2024.3433500.

Chouchou, Florian and Martin Desseilles (2014). "Heart rate variability: a tool to explore the sleeping brain?" In: *Frontiers in Neuroscience* 8, p. 402. doi: 10.3389/fnins.2014.00402.

Coeckelbergh, Mark (May 2019). "Artificial Intelligence: Some ethical issues and regulatory challenges". In: *Technology and Regulation* 2019, pp. 31–34. doi: 10.71265/a9yxhg88. url: https://techreg.org/article/view/10999.

Das Turja, Partha Pratim et al. (2024). "Investigation of HR and QT Variability for Monitoring Sleep Apnea: An Interpretable Machine Learning Approach". In: *Applied Intelligence and Informatics*. Ed. by Mufti Mahmud et al. Cham: Springer Nature Switzerland, pp. 169–185. isbn: 978-3-031-68639-9.

De Assis, Gilda Aparecida et al. (Oct. 2024). "Jogos com tecnologia vestível como estímulo à saúde dos pés – avaliação de usabilidade". In: *Journal of Health Informatics* 16. issn: 2175-4411. doi: 10.59681/2175-4411.v16.iEspecial.2024.1252.

Dhar, Ruby, Arun Kumar, and Subhradip Karmakar (Dec. 2023). "Smart wearable devices for real-time health monitoring". In: *Asian Journal of Medical Sciences* 14.12, pp. 1–3. url: https://www.nepjol.info/index.php/AJMS/article/view/58664.

Di Credico, Andrea et al. (2024). "The Prediction of Sleep Quality Using Heart Rate Variability Modulations During Wakefulness". In: *9th European Medical and Biological Engineering Conference, Vol. 2 (EMBEC 2024)*. Ed. by T Jarm, R Smerc, and S Mahnic-Kalamiza. Vol. 113. IFMBE Proceedings. Chap. 0, pp. 316–325. doi: 10.1007/978-3-031-61628-0_35.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* 29.5, pp. 1189–1232.

Fukuda, Shuichi et al. (2020). "Predicting Depression and Anxiety Mood by Wrist-Worn Sleep Sensor". In: *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. International Conference on Pervasive Computing and Communications. Chap. 0. doi: 10.1109/percomworkshops48775.2020.9156176.

Fund, N. et al. (2019). "The epidemiology of sleep disorders in Israel: results from a population-wide study." In: *Sleep medicine* 67, pp. 120–127. doi: 10.1016/j.sleep.2019.10.010.

Gerke, Sara, Timo Minssen, and Glenn Cohen (2020). "Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare". In: *Artificial Intelligence in Healthcare*. Ed. by Adam Bohr and Kaveh Memarzadeh. Academic Press, pp. 295–336. isbn: 978-0-12-818438-7. doi: https://doi.org/10.1016/B978-0-12-818438-7.00012-5. url: https://www.sciencedirect.com/science/article/pii/B9780128184387000125.

Goetz, Lea et al. (2024). "Generalization—a key challenge for responsible AI in patient-facing clinical applications". In: *npj Digital Medicine* 7.1, p. 126. issn: 2398-6352. doi: 10.1038/s41746-024-01127-3. url: https://doi.org/10.1038/s41746-024-01127-3.

Gomes, Sophia Artiaga, Júlia França Montanini, and Hermínio Maurício da Rocha Sobrinho (Dec. 2024). "O uso da Inteligência Artificial na Medicina: os benefícios e desafios da parceria homem-tecnologia na saúde". In: *Revista Eletrónica Acervo Saúde* 24, e18374. issn: 2178-2091. doi: 10.25248/reas.e18374.2024.

Habib, Ahsan et al. (June 2023). "Performance of a Convolutional Neural Network Derived From PPG Signal in Classifying Sleep Stages". In: *IEEE Transactions on Biomedical Engineering* 70.6, pp. 1717–1728. issn: 0018-9294. doi: 10.1109/TBME.2022.3219863.

Hamza, Manar Ahmed et al. (Jan. 2023). "Wearables-Assisted Smart Health Monitoring for Sleep Quality Prediction Using Optimal Deep Learning". In: *SUSTAINABILITY* 15.2. doi: 10.3390/su15021084.

Hanna, Matthew et al. (2024). "Ethical and Bias Considerations in Artificial Intelligence (AI)/Machine Learning". In: *Modern Pathology*, p. 100686. issn: 0893-3952. doi: https://doi.org/10.1016/j.modpat.2024.100686. url: https://www.sciencedirect.com/science/article/pii/S0893395224002667.

Haval, Abhijeet Madhukar and Md Afzal (2024). "Role of Wearable Health Devices in Public Health: Developing Flexible Electronics for Seamless and Continuous Health Monitoring". In: *South Eastern European Journal of Public Health*. url: https://api.semanticscholar.org/CorpusID:272716996.

Health, National Institutes of (n.d.). *Overweight & Obesity Statistics - NIDDK*. url: https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity.

Hedrick, Traci L. et al. (Apr. 2020). "Wearable Technology in the Perioperative Period: Predicting Risk of Postoperative Complications in Patients Undergoing Elective Colorectal Surgery". In: *Diseases of the Colon and Rectum* 63, pp. 538–544. issn: 15300358. doi: 10.1097/DCR.0000000000001580.

Hidayat, Alam Ahmad, Arif Budiarto, and Bens Pardamean (2023). "Long Short-Term Memory-based Models for Sleep Quality Prediction from Wearable Device Time Series Data". In: *Procedia Computer Science* 227, pp. 1062–1069. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2023.10.616. url: https://www.sciencedirect.com/science/article/pii/S1877050923017830.

Hu, Yuzhu et al. (May 2024). "An Ensemble Classification Model for Depression Based on Wearable Device Sleep Data". In: *IEEE Journal of Biomedical and Health Informatics* 28.5, pp. 2602–2612. issn: 2168-2194. doi: 10.1109/JBHI.2023.3258601.

Hua, Jing, Jiajun Lyu, and Wenchong Du (2022). "Prevalence of Sleep Disorder in Chinese Preschoolers: A National Population-Based Study". In: *Nature and Science of Sleep* 14, pp. 2091–2095. doi: 10.2147/NSS.S383209.

Hulsen, Tim (2023). "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare". In: *AI* 4.3, pp. 652–666. issn: 2673-2688. doi: 10.3390/ai4030034. url: https://www.mdpi.com/2673-2688/4/3/34.

Huthart, Sam et al. (Dec. 2020). "Advancing PPG Signal Quality and Know-How Through Knowledge Translation—From Experts to Student and Researcher". In: *Frontiers in Digital Health* 2. doi: 10.3389/fdgth.2020.619692.

Iwasaki, Ayako et al. (Dec. 2021). "Screening of sleep apnea based on heart rate variability and long short-term memory." eng. In: *Sleep & breathing = Schlaf & Atmung* 25.4, pp. 1821–1829. issn: 1522-1709 (Electronic). doi: 10.1007/s11325-020-02249-0. url: https://pubmed.ncbi.nlm.nih.gov/33423183/.

Iyer, Sridhar, Sujay Gejji, and Rahul Jashvantbhai Pandya (2022). "Survey on Applications of Wearable Technology for Healthcare". In: *Handbook of Research on Lifestyle Sustainability and Management Solutions Using AI, Big Data Analytics, and Visualization*. url: https://api.semanticscholar.org/CorpusID:246831045.

Jahrami, Haitham A. et al. (2019). "Prevalence of sleep problems among medical students: a systematic review and meta-analysis". In: *Journal of Public Health* 28, pp. 605–622. doi: 10.1007/s10389-019-01064-6.

Jenefa, A et al. (2023). "Smart Sleep Monitoring: IoMT and GRU for Effective Sleep Disorders Management". In: *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1767–1773. doi: 10.1109/ICECA58529.2023.10394887.

Jeon, Sanghoon, Yang-Soo Lee, and Sang Hyuk Son (2023). "Cascade Windows-Based Multi-Stream Convolutional Neural Networks Framework for Early Detecting In-Sleep Stroke Using Wristbands". In: *IEEE Access* 11, pp. 84944–84956. doi: 10.1109/ACCESS.2023.3301872.

Jeon, YeongJun, KukHo Heo, and Soon Ju Kang (2020). "Real-Time Sleep Apnea Diagnosis Method Using Wearable Device without External Sensors". In: *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1–5. doi: 10.1109/PerComWorkshops48775.2020.9156119.

John, Arlene, Barry Cardiff, and Deepu John (2021). "A 1D-CNN Based Deep Learning Technique for Sleep Apnea Detection in IoT Sensors". In: *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE International Symposium on Circuits and Systems. Chap. 0. doi: 10.1109/ISCAS51556.2021.9401300.

John, Arlene, Koushik Kumar Nundy, et al. (2021). "SomnNET: An SpO2 Based Deep Learning Network for Sleep Apnea Detection in Smartwatches". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE Engineering in Medicine and Biology Society Conference Proceedings. Chap. 0, pp. 1961–1964. doi: 10.1109/EMBC46164.2021.9631037.

Kargarandehkordi, Ali, Christopher Slade, and Peter Washington (Mar. 2024). "Personalized AI-Driven Real-Time Models to Predict Stress-Induced Blood Pressure Spikes Using Wearable Devices: Proposal for a Prospective Cohort Study." In: *JMIR research protocols* 13, e55615. issn: 1929-0748. doi: 10.2196/55615.

Khajehpiri, Boshra et al. (June 2024). "SleepBP-Net: A Time-Distributed Convolutional Network for Nocturnal Blood Pressure Estimation From Photoplethysmogram". In: *IEEE Sensors Journal* 24.12, pp. 19590–19600. issn: 1530-437X. doi: 10.1109/JSEN.2024.3396052.

Khan, A. et al. (2021). "Ethics of AI: A Systematic Literature Review of Principles and Challenges". In: *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*. doi: 10.1145/3530019.3531329.

Kilic, Ozan, Berrenur Saylam, and Ozlem Durmaz Incel (2023). "Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning". In: *Proceedings of the 2023 8th International Conference on Machine Learning Technologies (ICMLT 2023)*. Chap. 0, pp. 116–120. doi: 10.1145/3589883.3589900.

Kotzen, Kevin et al. (Feb. 2023). "SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography". In: *IEEE Journal of Biomedical and Health Informatics* 27.2, pp. 924–932. issn: 2168-2194. doi: 10.1109/JBHI.2022.3225363.

Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, pp. 159–174.

Leclercq, Christophe et al. (Oct. 2022). "Wearables, telemedicine, and artificial intelligence in arrhythmias and heart failure: Proceedings of the European Society of Cardiology Cardiovascular Round Table". en. In: *Europace* 24.9, pp. 1372–1383.

Lee, Minki P. et al. (Jan. 2024). "Imputing missing sleep data from wearables with neural networks in real-world settings". In: *Sleep* 47. issn: 15509109. doi: 10.1093/sleep/zsad266.

Lee, Young Jeong et al. (Nov. 2024). "Performance of consumer wrist-worn type sleep tracking devices compared to polysomnography: a meta-analysis." In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*. issn: 1550-9397. doi: 10.5664/jcsm.11460.

Li, Qiao et al. (May 2021). "Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables." eng. In: *Physiological measurement* 42.4. issn: 1361-6579 (Electronic). doi: 10.1088/1361-6579/abf1b0. url: https://pubmed.ncbi.nlm.nih.gov/33761477/.

Li, Ryan T. et al. (Jan. 2016). "Wearable Performance Devices in Sports Medicine". In: *Sports Health* 8, pp. 74–78. issn: 19410921. doi: 10.1177/1941738115616917.

Li, Tianjing, Ian J. Saldanha, and Karen A. Robinson (2022). "Introduction to Systematic Reviews". In: *Principles and Practice of Clinical Trials*. Ed. by Steven Piantadosi and Curtis L. Meinert. Cham: Springer International Publishing, pp. 2159–2177. isbn: 978-3-319-52636-2. doi: 10.1007/978-3-319-52636-2_194. url: https://doi.org/10.1007/978-3-319-52636-2_194.

Liang, Zilu and Mario Alberto Chapa-Martell (2021). "A Multi-Level Classification Approach for Sleep Stage Prediction With Processed Data Derived From Consumer Wearable Activity Trackers." eng. In: *Frontiers in digital health* 3, p. 665946. issn: 2673-253X (Electronic). doi: 10.3389/fdgth.2021.665946. url: https://pubmed.ncbi.nlm.nih.gov/34713139/.

Lima, Marcio Roberto de (2023). "Affected body: a self-tracking experience with a wearable technology". In: *Revista Brasileira de Ciencias do Esporte* 45. issn: 21793255. doi: 10.1590/rbce.45.e20230055.

Liu, Chengzhi et al. (2024). "Enhanced Human Activity Recognition Framework for Wearable Devices Based on Explainable AI". In: *2024 IEEE International Symposium on Consumer Technology (ISCT)*, pp. 385–391. url: https://api.semanticscholar.org/CorpusID:274826116.

Liu, Haipeng et al. (June 2021). "Filtering-induced time shifts in photoplethysmography pulse features measured at different body sites: The importance of filter definition and standardization". In: *Physiological Measurement* 42. doi: 10.1088/1361-6579/ac0a34.

Ma, Yaopeng J.X. et al. (2023). "Automatic sleep-stage classification of heart rate and actigraphy data using deep and transfer learning approaches". In: *Computers in Biology and Medicine* 163, p. 107193. issn: 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2023.107193. url: https://www.sciencedirect.com/science/article/pii/S0010482523006583.

Mahadevan, Nikhil et al. (Mar. 2021). "Development of digital measures for nighttime scratch and sleep using wrist-worn wearable devices." eng. In: *NPJ digital medicine* 4.1, p. 42. issn: 2398-6352 (Electronic). doi: 10.1038/s41746-021-00402-x. url: https://pubmed.ncbi.nlm.nih.gov/33658610/.

Mantua, Janna, Nickolas Gravel, and Rebecca M.C. Spencer (May 2016). "Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography". In: *Sensors (Switzerland)* 16. issn: 14248220. doi: 10.3390/s16050646.

McArdle, N. et al. (2022). "Prevalence of common sleep disorders in a middle-aged community sample." In: *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*. doi: 10.5664/jcsm.9886.

Moraes, Joel et al. (Aug. 2023). "IMPACTO DA TECNOLOGIA DE INTELIGÊNCIA ARTIFICIAL NA MEDICINA DIAGNÓSTICA". In: *Revista Ibero-Americana de Humanidades, Ciências e Educação* 9, pp. 1303–1214. doi: 10.51891/rease.v9i7.10699.

Moreno-Pino, Fernando et al. (Nov. 2022). "Heterogeneous Hidden Markov Models for Sleep Activity Recognition from Multi-Source Passively Sensed Data". In: arXiv: 2211.10371.

Moses, Iyekekpolor Osamudiame (2024). "The Rise of Biochemical Sensors: Technology for Real-Time Health Monitoring (Applications and Future Scope)". In: *Journal of Multidisciplinary Science: MIKAILALSYS*. url: https://api.semanticscholar.org/CorpusID:273417960.

Motin, Mohammod Abdul et al. (July 2020). "Photoplethysmographic-based automated sleep-wake classification using a support vector machine". In: *Physiological Measurement* 41.7. issn: 0967-3334. doi: 10.1088/1361-6579/ab9482.

Mousavi, Ali et al. (Nov. 2023). "Recent advances in smart wearable sensors as electronic skin". In: *Journal of Materials Chemistry B* 11, pp. 10332–10354. issn: 20507518. doi: 10.1039/d3tb01373a.

N, Abinaya et al. (2024). "A Restful Night's Sleep: Predicting Stress with Machine Learning - 2024 2nd International Conference on Disruptive Technologies (ICDT)". In: *2024 2nd International Conference on Disruptive Technologies (ICDT)*. Chap. 0, pp. 1497–1500. doi: 10.1109/ICDT61202.2024.10489403.

Nam, Borum et al. (Feb. 2024). "InsightSleepNet: the interpretable and uncertainty-aware deep learning network for sleep staging using continuous Photoplethysmography." eng. In: *BMC medical informatics and decision making* 24.1, p. 50. issn: 1472-6947 (Electronic). doi: 10.1186/s12911-024-02437-y. url: https://pubmed.ncbi.nlm.nih.gov/38355559/.

Nur, Siti (2024). "The Role of Digital Health Technologies and Sensors in Revolutionizing Wearable Health Monitoring Systems". In: *International Journal of Innovative Research in Computer Science and Technology*. url: https://api.semanticscholar.org/CorpusID:274360021.

Ojalvo, Dávid, André Pekkola Pacheco, and Christian Benedict (Oct. 2023). "A useful tool or a new challenge? Hand-wrist-worn sleep trackers in patients with insomnia". In: *Journal of Sleep Research* 32. issn: 13652869. doi: 10.1111/jsr.13883.

Olawade, David B et al. (Dec. 2024). "Integrating AI-driven wearable devices and biometric data into stroke risk assessment: A review of opportunities and challenges." In: *Clinical neurology and neurosurgery* 249, p. 108689. issn: 1872-6968. doi: 10.1016/j.clineuro.2024.108689.

Olsen, Mads et al. (Aug. 2024). "A Deep Transfer Learning Approach for Sleep Stage Classification and Sleep Apnea Detection Using Wrist-Worn Consumer Sleep Technologies". In: *IEEE Transactions on Biomedical Engineering* 71.8, pp. 2506–2517. issn: 0018-9294. doi: 10.1109/TBME.2024.3378480.

Osathitporn, Pongpanut et al. (2023). "RRWaveNet: A Compact End-to-End Multiscale Residual CNN for Robust PPG Respiratory Rate Estimation". In: *IEEE Internet of Things Journal* 10.18, pp. 15943–15952. doi: 10.1109/JIOT.2023.3265980.

Page, Matthew J et al. (2021). "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". In: *BMJ* 372. doi: 10.1136/bmj.n71. eprint: https://www.bmj.com/content/372/bmj.n71.full.pdf. url: https://www.bmj.com/content/372/bmj.n71.

Papini, Gabriele B et al. (Aug. 2020). "Wearable monitoring of sleep-disordered breathing: estimation of the apnea-hypopnea index using wrist-worn reflective photoplethysmography." eng. In: *Scientific reports* 10.1, p. 13512. issn: 2045-2322 (Electronic). doi: 10.1038/s41598-020-69935-7. url: https://pubmed.ncbi.nlm.nih.gov/32782313/.

Park, Kyung Mee et al. (May 2022). "Prediction of good sleep with physical activity and light exposure: a preliminary study." eng. In: *Journal of clinical sleep medicine : JCSM*

: official publication of the American Academy of Sleep Medicine 18.5, pp. 1375–1383. issn: 1550-9397 (Electronic). doi: 10.5664/jcsm.9872. url: https://pubmed.ncbi.nlm.nih.gov/34989333/.

Park, Kyung Mee et al. (Dec. 2024). "Predicting sleep based on physical activity, light exposure, and Heart rate variability data using wearable devices." eng. In: *Annals of medicine* 56.1, p. 2405077. issn: 1365-2060 (Electronic). doi: 10.1080/07853890.2024.2405077. url: https://pubmed.ncbi.nlm.nih.gov/39297306/.

Parliament, European and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data*. Accessed: 2025-01-18. url: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

– (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 14 February 2024 on Artificial Intelligence*. Accessed: 2025-01-18. url: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Paul, Tanmoy et al. (Nov. 2024). "Lightweight and Low-Parametric Network for Hardware Inference of Obstructive Sleep Apnea." eng. In: *Diagnostics (Basel, Switzerland)* 14.22. issn: 2075-4418 (Print). doi: 10.3390/diagnostics14222505. url: https://pubmed.ncbi.nlm.nih.gov/39594171/.

Peraza, Luis R. et al. (2020). "Device agnostic sleep-wake segment classification from wrist-worn accelerometry". In: *2020 8th IEEE International Conference on Healthcare Informatics (ICHI 2020)*. IEEE International Conference on Healthcare Informatics. Chap. 0, pp. 463–465. doi: 10.1109/ICHI48887.2020.9374318.

PRISMA Group (2024). *PRISMA Statement: Transparent Reporting of Systematic Reviews and Meta-Analyses*. https://www.prisma-statement.org/. Accessed: 2024-12-22.

Qin, Hengji and Guanzheng Liu (2022). "A dual-model deep learning method for sleep apnea detection based on representation learning and temporal dependence". In: *Neurocomputing* 473, pp. 24–36. issn: 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.12.001. url: https://www.sciencedirect.com/science/article/pii/S092523122101835X.

Ratheesh, Karthika et al. (2023). "Open Tool-kit for AI-based Sleep Apnea Scoring - 2023 4th International Conference for Emerging Technology (INCET)". In: *2023 4th International Conference for Emerging Technology (INCET)*. Chap. 0, pp. 1–4. doi: 10.1109/INCET57972.2023.10170040.

Robbins, Rebecca et al. (Oct. 2024). "Accuracy of Three Commercial Wearable Devices for Sleep Tracking in Healthy Adults". en. In: *Sensors (Basel)* 24.20.

Roberts, Daniel M et al. (Mar. 2020). "Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography". In: *Sleep* 43.7, zsaa045. issn: 0161-8105. doi: 10.1093/sleep/zsaa045. eprint: https://academic.oup.com/sleep/article-pdf/43/7/zsaa045/33460801/zsaa045.pdf. url: https://doi.org/10.1093/sleep/zsaa045.

Schyvens, An-Marie et al. (2023). "Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP Versus Polysomnography: Systematic Review". In: *JMIR mHealth and uHealth* 12. doi: 10.2196/52192.

Sharma, Nishant et al. (Feb. 2024). "Automated accurate insomnia detection system using wavelet scattering method using ECG signals". In: *Applied Intelligence* 54.4, pp. 3464–3481. issn: 0924-669X. doi: 10.1007/s10489-024-05284-6.

Shen, Qi et al. (2021). "Multiscale Deep Neural Network for Obstructive Sleep Apnea Detection Using RR Interval From Single-Lead ECG Signal". In: *IEEE Transactions on Instrumentation and Measurement* 70. issn: 0018-9456. doi: 10.1109/TIM.2021.3062414.

Silva, Fernanda B. et al. (2024). "Sleep staging algorithm based on smartwatch sensors for healthy and sleep apnea populations". In: *Sleep Medicine* 119, pp. 535–548. issn: 1389-9457. doi: https://doi.org/10.1016/j.sleep.2024.05.033. url: https://www.sciencedirect.com/science/article/pii/S138994572400248X.

Skaer, Tracy L. and David A. Sclar (2010). "Economic implications of sleep disorders". In: *PharmacoEconomics* 28, pp. 1015–1023. issn: 11707690. doi: 10.2165/11537390-000000000-00000.

Song, Tzu-An, Samadrita Roy Chowdhury, et al. (May 2023). "AI-Driven sleep staging from actigraphy and heart rate". In: *PLOS ONE* 18.5, pp. 1–29. doi: 10.1371/journal.pone.0285703. url: https://doi.org/10.1371/journal.pone.0285703.

Song, Tzu-An, Chowdhury SR, et al. (2023). "AI-Driven sleep staging from actigraphy and heart rate." eng. In: *PloS one* 18.5, e0285703. issn: 1932-6203 (Electronic). doi: 10.1371/journal.pone.0285703. url: https://pubmed.ncbi.nlm.nih.gov/37195925/.

Soni, Tanishq, Deepali Gupta, and Mudita Uppal (2023). "Enhancing Accuracy of Sleep Disorder with Logistic Regression Model - 2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDeA)". In: *2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDeA)*. Chap. 0, pp. 292–295. doi: 10.1109/ICIDeA59866.2023.10295230.

Sumitra, Madiri Divya et al. (2023). "Deep Learning Model for ECG-based Sleep Apnea Detection - 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)". In: *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. Chap. 0, pp. 280–284. doi: 10.1109/ICAAIC56838.2023.10140417.

Sundararajan, Kalaivani et al. (Jan. 2021). "Sleep classification from wrist-worn accelerometer data using random forests." eng. In: *Scientific reports* 11.1, p. 24. issn: 2045-2322 (Electronic). doi: 10.1038/s41598-020-79217-x. url: https://pubmed.ncbi.nlm.nih.gov/33420133/.

Tat, Trinny et al. (Sept. 2022). "Smart Textiles for Healthcare and Sustainability". In: *ACS Nano* 16, pp. 13301–13313. issn: 1936086X. doi: 10.1021/acsnano.2c06287.

Topalidis, Pavlos I et al. (Nov. 2023). "From Pulses to Sleep Stages: Towards Optimized Sleep Classification Using Heart-Rate Variability." eng. In: *Sensors (Basel, Switzerland)* 23.22. issn: 1424-8220 (Electronic). doi: 10.3390/s23229077. url: https://pubmed.ncbi.nlm.nih.gov/38005466/.

Umutoni, Ritha Marie et al. (2023). "Integration of TinyML-based proximity and couch sensing in wearable devices for monitoring infectious disease's social distance compliance". In: *Proceedings of the 2023 12th International Conference on Software and Computer Applications*. url: https://api.semanticscholar.org/CorpusID:259205741.

Verma, R. et al. (2023). "Artificial intelligence in sleep medicine: Present and future". In: *World Journal of Clinical Cases* 11, pp. 8106–8110. doi: 10.12998/wjcc.v11.i34.8106.

Virtanen, Pauli et al. (Feb. 2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3, pp. 261–272. issn: 1548-7105. doi: 10.1038/s41592-019-0686-2. url: http://dx.doi.org/10.1038/s41592-019-0686-2.

Wang, Shaokui et al. (Apr. 2023). "Machine Learning Assisted Wearable Wireless Device for Sleep Apnea Syndrome Diagnosis." eng. In: *Biosensors* 13.4. issn: 2079-6374 (Electronic). doi: 10.3390/bios13040483. url: https://pubmed.ncbi.nlm.nih.gov/37185558/.

Wang, Will Ke et al. (June 2024). "Addressing Wearable Sleep Tracking Inequity: A New Dataset and Novel Methods for a Population with Sleep Disorders". In: *Proceedings of the fifth Conference on Health, Inference, and Learning*. Ed. by Tom Pollard et al. Vol. 248. Proceedings of Machine Learning Research. PMLR, pp. 380–396. url: https://proceedings.mlr.press/v248/wang24a.html.

Wang, Xiaowei et al. (June 2020). "Obstructive sleep apnea detection using ecg-sensor with convolutional neural networks". In: *Multimedia Tools and Applications* 79.23, pp. 15813–15827. issn: 1380-7501. doi: 10.1007/s11042-018-6161-8.

Wang, Zhiya et al. (2022). "Single-lead ECG based multiscale neural network for obstructive sleep apnea detection". In: *Internet of Things* 20, p. 100613. issn: 2542-6605. doi: https://doi.org/10.1016/j.iot.2022.100613. url: https://www.sciencedirect.com/science/article/pii/S2542660522000956.

Watson, N. and Chris Fernandez (2021). "Artificial intelligence and sleep: Advancing sleep medicine." In: *Sleep medicine reviews* 59, p. 101512. doi: 10.1016/j.smrv.2021.101512.

Wongtaweesup, Chanphot et al. (2023). "Using Consumer-Graded Wearable Devices for Sleep Apnea Pre-Diagnosis: A Survey and Recommendations". In: *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 517–522. doi: 10.1109/JCSSE58229.2023.10202122.

Yang, Quanan et al. (2022). "Obstructive sleep apnea detection from single-lead electrocardiogram signals using one-dimensional squeeze-and-excitation residual group network". In: *Computers in Biology and Medicine* 140, p. 105124. issn: 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.105124. url: https://www.sciencedirect.com/science/article/pii/S0010482521009185.

Ye, Guanhua et al. (Aug. 2021). "FENet: A Frequency Extraction Network for Obstructive Sleep Apnea Detection". In: *IEEE Journal of Biomedical and Health Informatics* 25.8, pp. 2848–2856. issn: 2168-2194. doi: 10.1109/JBHI.2021.3050113.

Zamani, Abu Sarwar et al. (2023). "The prediction of sleep quality using wearable-assisted smart health monitoring systems based on statistical data". In: *Journal of King Saud University - Science* 35.9, p. 102927. issn: 1018-3647. doi: https://doi.org/10.1016/j.jksus.2023.102927. url: https://www.sciencedirect.com/science/article/pii/S1018364723003890.

Zhou, Yu, Yinxian He, and Kyungtae Kang (2022). "OSA-CCNN: Obstructive Sleep Apnea Detection Based on a Composite Deep Convolution Neural Network Model using Single-Lead ECG signal - 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)". In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Chap. 0, pp. 1840–1845. doi: 10.1109/BIBM55620.2022.9995675.

Appendix A

Summary of Selected Articles from the Systematic Review

List of Articles

Table A.1: Articles selected in the systematic review

Year	Title	Citation
2020	Wearable monitoring of sleep-disordered breathing: estimation of the apnea-hypopnea index using wrist-worn reflective photoplethysmography.	(Papini et al. 2020)
2020	Analysis of Data from Wearable Sensors for Sleep Quality Estimation and Prediction Using Deep Learning.	(Arora, P. Chakraborty, and Bhatia 2020)
2020	Device agnostic sleep-wake segment classification from wrist-worn accelerometry	(Peraza et al. 2020)
2020	Predicting Depression and Anxiety Mood by Wrist-Worn Sleep Sensor.	(Fukuda et al. 2020)
2020	Obstructive sleep apnea detection using ecg-sensor with convolutional neural networks.	(X. Wang et al. 2020)
2020	Photoplethysmographic-based automated sleep-wake classification using a support vector machine.	(Motin et al. 2020)
2021	Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables.	(Q. Li et al. 2021)
2021	Development of digital measures for nighttime scratch and sleep using wrist-worn wearable devices.	(Mahadevan et al. 2021)
2021	Sleep classification from wrist-worn accelerometer data using random forests.	(Sundararajan et al. 2021)
2021	Screening of sleep apnea based on heart rate variability and long short-term memory.	(Iwasaki et al. 2021)

2021	A Multi-Level Classification Approach for Sleep Stage Prediction With Processed Data Derived From Consumer Wearable Activity Trackers.	(Liang and Chapa-Martell 2021)
2021	A Novel Sleep Scoring Algorithm-Based Framework and Sleep Pattern Analysis Using Machine Learning Techniques.	(S. Chakraborty, Aich, and Kim 2021)
2021	ApneaDetector: Detecting Sleep Apnea with Smartwatches.	(Xianda Chen et al. 2021)
2021	FENet: A Frequency Extraction Network for Obstructive Sleep Apnea Detection.	(Ye et al. 2021)
2021	A 1D-CNN Based Deep Learning Technique for Sleep Apnea Detection in IoT Sensors.	(A. John, Cardiff, and D. John 2021)
2021	SomnNET: An SpO2 Based Deep Learning Network for Sleep Apnea Detection in Smartwatches.	(A. John, Nundy, et al. 2021)
2021	Multiscale Deep Neural Network for Obstructive Sleep Apnea Detection Using RR Interval From Single-Lead ECG Signal.	(Shen et al. 2021)
2022	Single-lead ECG based multiscale neural network for obstructive sleep apnea detection	(Z. Wang et al. 2022)
2022	Obstructive sleep apnea detection from single-lead electrocardiogram signals using one-dimensional squeeze-and-excitation residual group network	(Yang et al. 2022)
2022	Electrodermal activity based autonomic sleep staging using wrist wearable	(Anusha et al. 2022)
2022	A dual-model deep learning method for sleep apnea detection based on representation learning and temporal dependence	(Qin and G. Liu 2022)
2022	OSA-CCNN: Obstructive Sleep Apnea Detection Based on a Composite Deep Convolution Neural Network Model using Single-Lead ECG signal	(Zhou, He, and K. Kang 2022)
2022	Obstructive Sleep Apnoea Syndrome Screening Through Wrist-Worn Smartbands: A Machine-Learning Approach.	(Benedetti et al. 2022)
2022	Prediction of good sleep with physical activity and light exposure: a preliminary study.	(Park et al. 2022)
2022	Using Data from Wearables for Better Sleep	(Arias 2022)
2022	Deep Learning Forecasts the Occurrence of Sleep Apnea from Single-Lead ECG.	(Bahrami and Forouzanfar 2022)

2023	Automatic sleep-stage classification of heart rate and actigraphy data using deep and transfer learning approaches	(Ma et al. 2023)
2023	The prediction of sleep quality using wearable-assisted smart health monitoring systems based on statistical data	(Zamani et al. 2023)
2023	Long Short-Term Memory-based Models for Sleep Quality Prediction from Wearable Device Time Series Data	(Hidayat, Budiarto, and Pardamean 2023)
2023	Open Tool-kit for AI-based Sleep Apnea Scoring – 2023 4th International Conference for Emerging Technology (INCET)	(Ratheesh et al. 2023)
2023	Deep Learning Model for ECG-based Sleep Apnea Detection – 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)	(Sumitra et al. 2023)
2023	Enhancing Accuracy of Sleep Disorder with Logistic Regression Model – 2023 IEEE 2nd International Conference on Industrial Electronics: Developments & Applications (ICIDeA)	(Soni, Gupta, and Uppal 2023)
2023	Cascade Windows-Based Multi-Stream Convolutional Neural Networks Framework for Early Detecting In-Sleep Stroke Using Wristbands	(S. Jeon, Y.-S. Lee, and Son 2023)
2023	Machine Learning Assisted Wearable Wireless Device for Sleep Apnea Syndrome Diagnosis.	(S. Wang et al. 2023)
2023	From Pulses to Sleep Stages: Towards Optimized Sleep Classification Using Heart-Rate Variability.	(Topalidis et al. 2023)
2023	AI-Driven sleep staging from actigraphy and heart rate.	(Song, SR, et al. 2023)
2023	Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning	(Kilic, Saylam, and Incel 2023)
2023	Wearables-Assisted Smart Health Monitoring for Sleep Quality Prediction Using Optimal Deep Learning.	(Hamza et al. 2023)
2023	SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography.	(Kotzen et al. 2023)
2023	Performance of a Convolutional Neural Network Derived From PPG Signal in Classifying Sleep Stages.	(Habib et al. 2023)
2023	RRWaveNet: A Compact End-to-End Multiscale Residual CNN for Robust PPG Respiratory Rate Estimation.	(Osathitporn et al. 2023)
2023	A Signal Segmentation-Free Model for Electrocardiogram-Based Obstructive Sleep Apnea Severity Classification.	(J.-W. Chen et al. 2023)
2024	Sleep staging algorithm based on smartwatch sensors for healthy and sleep apnea populations	(Silva et al. 2024)

2024	A Restful Night's Sleep: Predicting Stress with Machine Learning – 2024 2nd International Conference on Disruptive Technologies (ICDT)	(N et al. 2024)
2024	Lightweight and Low-Parametric Network for Hardware Inference of Obstructive Sleep Apnea.	(Paul et al. 2024)
2024	Predicting sleep based on physical activity, light exposure, and Heart rate variability data using wearable devices.	(Park et al. 2024)
2024	InsightSleepNet: the interpretable and uncertainty-aware deep learning network for sleep staging using continuous Photoplethysmography.	(Nam et al. 2024)
2024	Multitask Learning for Automated Sleep Staging and Wearable Technology Integration.	(Chih et al. 2024)
2024	Automated accurate insomnia detection system using wavelet scattering method using ECG signals.	(Sharma et al. 2024)
2024	SleepBP-Net: A Time-Distributed Convolutional Network for Nocturnal Blood Pressure Estimation From Photoplethysmogram.	(Khajehpiri et al. 2024)
2024	ApSense: Data-Driven Algorithm in PPG-Based Sleep Apnea Sensing.	(Choksatchawathi et al. 2024)
2024	Investigation of HR and QT Variability for Monitoring Sleep Apnea: An Interpretable Machine Learning Approach.	(Das Turja et al. 2024)
2024	A Deep Transfer Learning Approach for Sleep Stage Classification and Sleep Apnea Detection Using Wrist-Worn Consumer Sleep Technologies.	(Olsen et al. 2024)
2024	An Ensemble Classification Model for Depression Based on Wearable Device Sleep Data.	(Hu et al. 2024)
2024	The Prediction of Sleep Quality Using Heart Rate Variability Modulations During Wakefulness.	(Di Credico et al. 2024)

Appendix B

Additional Figures and Visualisations

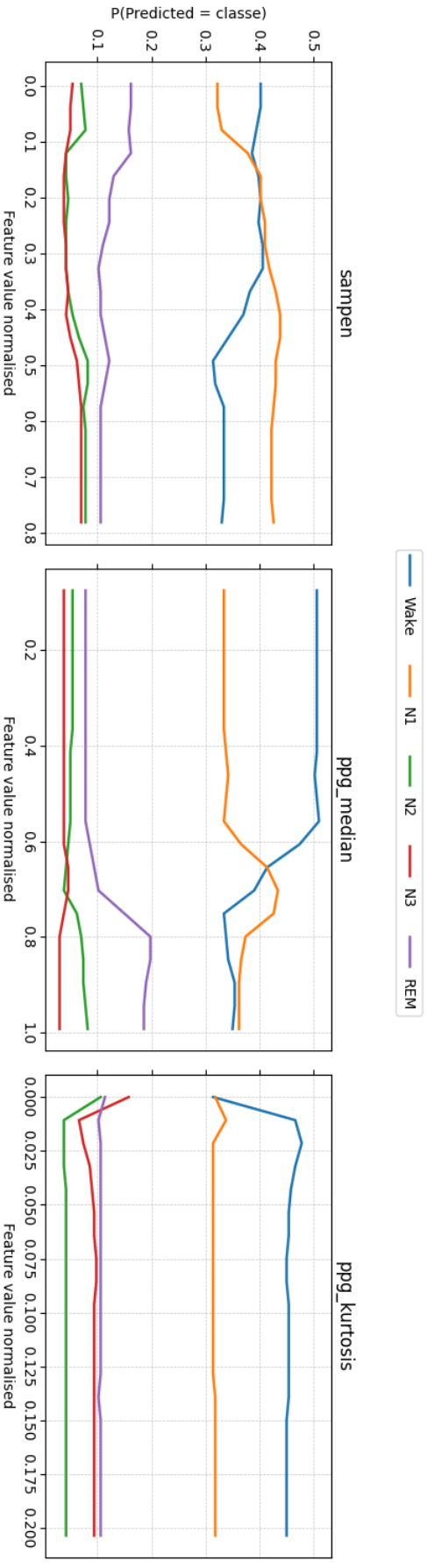

Figure B.1: Expanded partial-dependence curves for the three most important features in sleep stage classification. Larger version provided for readability.

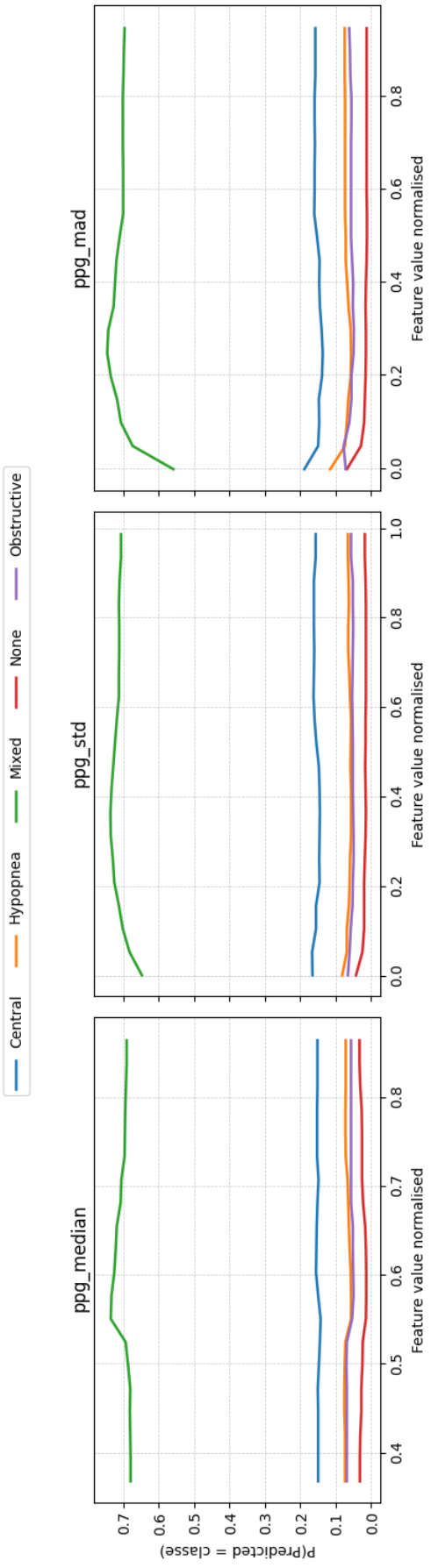

Figure B.2: Expanded partial-dependence curves for the three most important features in sleep apnea detection. Larger version provided for readability.