PhD-FSTM-2022-137

The Faculty of Science, Technology and Medicine

DISSERTATION

Presented on 25/10/2022 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN SCIENCES DE L'INGÉNIEUR

by

Marino Eliezer Gavidia Borrero

Born on 27/10/1992 in Mérida (Venezuela)

AI METHODS FOR DETECTION AND PREDICTION OF DISEASES FROM PHYSIOLOGICAL DATA

Dissertation Defence Committee:

Dr. Jorge Gonçalves, Dissertation Supervisor
Professor, Université du Luxembourg

Dr. Alexander Skupin, Chairman
Professor, Université du Luxembourg

Dr. Roberto Sassi
Professor, Università degli Studi di Milano

Dr. Laurent Mombaerts, Vice Chair
Applied Scientist, Amazon Web Services

Dr. Jan Soelter
Data Scientist, Hochschule Stralsund University, N4

Declaration

The author hereby certifies that, except for specific references to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification at this or any other university. Except as mentioned in the text and acknowledgments, this dissertation is entirely my work. This dissertation comprises less than 65,000 words, including a bibliography, footnotes, tables, equations, and less than 150 figures.

Marino E. Gavidia B.

October 2022

Acknowledgements

First and foremost, I would like to express my deep gratitude to Prof. Dr. Jorge Goncalves for allowing me to conduct my study in collaboration with the Systems Control Group at the University of Luxembourg and for his endless support over the last four years. Your encouragement enabled me to pursue my potential and I appreciate that you trusted me and gave me the flexibility to develop and build my ideas. This thesis would not have been possible without it.

Furthermore, I would like to thank all my colleagues who helped me during my Ph.D. journey, especially those who contributed directly and indirectly, by providing direction and feedback to the development of this thesis; Stefano, Daniele, Arthur, Jesus, Atte, Mehri, François, Andreas. Likewise, I would like to thank all my colleagues from the Systems Control Group and the Interventional Neuroscience Group, who has been like a second family to me during the previous four years. Thank you for all the good memories and for making the office a delightful place.

I want to express my gratitude to the thesis defense committee members for taking the time to examine my work; Prof. Dr. Jorge Goncalves, Prof. Dr. Alexander Skupin, Prof. Dr. Roberto Sassi, Dr. Laurent Mombaerts, and Dr. Jan Soelter. Also, thanks to my CET committee for all the discussions and feedback provided during the last years.

I would also like to thank my Ph.D. training unit, CriTiCS, for allowing me to participate in exciting activities and exchange expertise. Furthermore, I would like to thank the Fonds National de Recherche in Luxembourg for sponsoring my research under PRIDE15/10907093/CriTiCS.

Finally, I would like to express my deepest gratitude to all my friends for being there in the ups and downs moments, especially to my parents and my brother, for their unwavering support throughout my life.

Abstract

In recent years, health care spending has shown a trend toward global price increases. Therefore, health care systems must become more effective in treating patients to reverse this trend. Incentives are in place to encourage the development of personalized and data-based health care systems. With the increasing amount of health-related data generation, effective computational approaches are required to integrate these data in a meaningful manner and reveal unknown insights.

Given the large amount of available health care data, a significant question at this time is whether it is possible to develop tools that can be used to monitor and predict the future physiological condition of a patient. This can help transform the paradigm of reactionary patient care into one that helps patients receive more preventive care.

Consequently, this thesis sought to answer this question by providing two examples in which machine learning and neural networks can be applied to physiological time series data to continuously monitor, predict and/or detect life-threatening events in patients. These two examples are developed in the context of Atrial Fibrillation and Obstructive Sleep Apnea.

Atrial fibrillation (AF) is the most common abnormal heart rhythm, linked to a higher chance of being hospitalized, heart failure, or stroke. In this example, we address the question of whether it is possible to find early warning signals for AF, in the order of minutes. We show that the answer to this question is affirmative. We developed a deep convolutional neural network model, trained and cross-validated with a 24-hour R-to-R interval (RRI) signal from Holter electrocardiogram (ECG) recordings of 280 individuals and tested on 70. The suggested model can anticipate the onset of AF on average 31 minutes before the onset, with an accuracy of 83% and an F1 score of 85%. The proposed method is simple and could be embedded in standard wearable devices that record RRI data to continuously monitor the heart and provide early warnings to patients. It can potentially warn patients to take preventive steps (such as taking oral antiarrhythmic medications) and avoid AF onset.

Obstructive sleep apnea (OSA) is a common respiratory disorder characterized by respiratory tract obstruction and breathing disturbances during sleep, often diagnosed with overnight polysomnography (PSG) monitoring. However, continuous monitoring of PSG is not feasible as it is costly, time-

consuming, and uncomfortable for patients. To circumvent these issues, in this example, we propose an automatic detection method of OSA events using only data available from easy-to-use wearables: electrocardiogram, respiratory, and oximetry data. We used data from three sleep studies from the National Sleep Research Resource (NSRR), the largest public repository, consisting of 10,878 recordings. The developed method is based on combining deep convolutional neural networks and a light-gradient-boosting machine (LightGBM) for classification. In the test data, our model achieved the highest classification performance in the literature, with accuracy and F1-score of 91%. Since the trained model is simple and computationally efficient, we expect that our method can be implemented for the automatic detection of OSA in unsupervised home monitoring systems.

The performance of the proposed methods is superior to that of existing methods in the literature, showing that this thesis can serve as a proof of concept that personalized and data-based health care tools can be built to continuously monitor and predict life-threatening events in patients. Such tools can and will have a substantial impact on the healthcare system since they can be implemented on inexpensive and non-invasive wearable devices. Thus, helping medical experts by providing them with a better understanding of the condition of patients, assisting them to intervene accordingly, reducing costs to healthcare systems, and improving patient care overall.

Table of contents

List of figures

List of tables

Glossary

AF	Atrial Fibrillation.
AI	Artificial Intelligence.
AUPRC	Area under precision-recall curve.
AUROC	Area under receiver operating characteristic curve.
CEC	Constant error carousel.
CNN	Convolutional Neural Networks.
DL	Deep Learning.
ECG	Electrocardiograms.
LSTM	Long Short-Term Memory.
ML	Machine Learning.
NN	Neural network.
OSA	Obstructive Sleep Apnea.
PSG	Polysomnography.
RNN	Recurrent neural networks.
ROC	Receiver operating characteristic.
RRI	R-to-R interval between heartbeats.
SR	Sinus rhythm.

Chapter 1

Introduction

1.1 Introduction

The reduction in mortality rates across all age groups in the world's population during the past century can be attributed to early detection of diseases, advances in medical diagnostics, and improved access to healthcare facilities [Thevenot et al., 2017]. Predictions and diagnoses come from the experience of the physician in identifying diseases early. However, the knowledge of a physician can sometimes be wrong and result in erroneous judgments. Furthermore, the healthcare sector generates large amounts of patient data, which contain previously unknown patterns or associations that can be used to advance early disease diagnosis [Naz et al., 2020].

The amount of health care data generation is growing exponentially, proportional to the rate of progress in information and communication technology. For example, the US healthcare system alone generated more than 150 exabytes (150×10^{18}) of data in the year 2011 [Raghupathi et al., 2014], and the total number of health records in the world was estimated to exceed 40 yottabytes (40×10^{24}) by the year 2020 [Nepal et al., 2015]. Big data impacts public health by anticipating outbreaks, improving patient care, and achieving better health outcomes [Mayer-Schönberger et al., 2013]. However, these multiple applications are impossible without statistical and automated models for data analysis, pattern identification, prediction, and decision making. Artificial intelligence (AI) and Machine Learning (ML) provide these capabilities. They have been used to find patterns and correlations with little or no human help [Cioffi et al., 2020]. And its popularity in the health domain has grown steadily over the past decade (see Fig. 1.1).

Recently, with advances in graphic processing units (GPUs), computing, and optimization, deep learning has emerged as a powerful tool that allows us to solve challenging tasks in both Computer

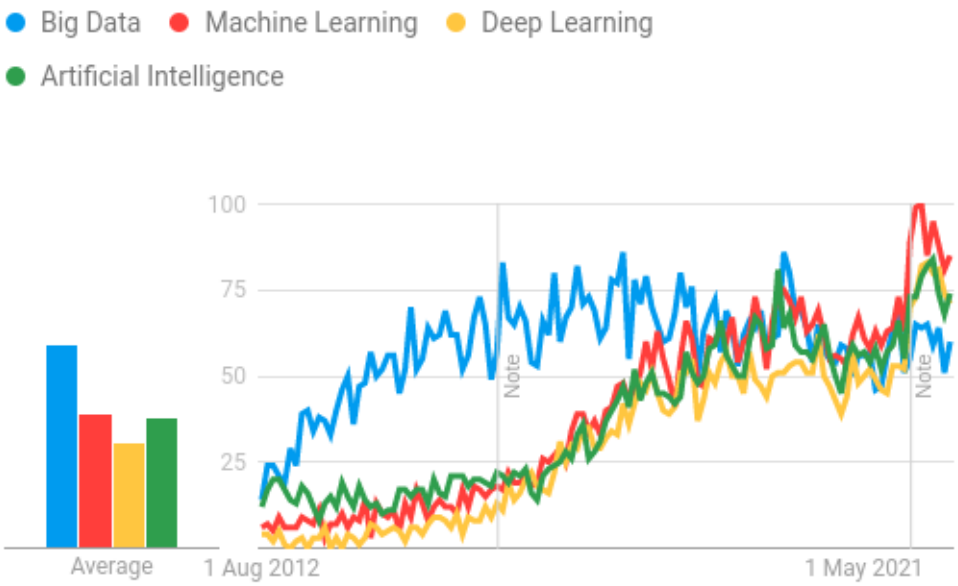

Fig. 1.1 Google trends on Big Data, Machine Learning, Deep Learning, and Artificial Intelligence.

Vision [Krizhevsky et al., 2012] and Medical Imaging [Esteva et al., 2017]. One of the most significant applications of AI and ML in the medical field is computer-aided diagnosis, which aims to help clinicians and physicians based on different sources of information. However, AI and ML cannot generally function without much data and skill, which refers to domain-specific knowledge [Ng, 2016].

1.1.1 Knowledge discovery from data

In recent years, there has been a significant expansion in the amount of data available in the medical field. Data mining techniques are increasingly gaining ground in various industries due to the large amounts of data saved in electronic form. The clinical sector has also shown significant interest in data mining due to its potential as a resource to help diagnose or improve medical processes [Jothi et al., 2015].

The step-by-step procedure for extracting meaningful information from the data is known as knowledge discovery in the database process, abbreviated as (KDD). Although many different ways describe the KDD process, most agree on the fundamental components, it is a process that consists of nine steps [Fayyad et al., 1996; Kotsiantis et al., 2006; Maimon et al., 2009].

1. Determine the purpose of the procedure and compile all the necessary background information in the application domain.

2. Determine the most appropriate data collection to use to derive knowledge.

3. Prepare the data in advance. As part of this process, unwanted or noisy data records are removed, and specific parameters are chosen, such as how to treat missing attribute values in data collection. The choice of variables used with the data set is also part of this stage.

4. Convert the collected data into a format that can be easily represented by eliminating any variables or factors that are not beneficial to achieving the assignment's goal.

5. Determine which data mining strategy will best serve the predetermined goal of the KDD process and implement it.

6. Select the data mining algorithm used after deciding on an overall data mining approach. It is important to reiterate once again that the choice of an alternative is often based on the preferences of the end users.

7. The algorithm is then applied to the pre-processed data and will then examine the data to determine whether it includes useful information.

8. Analyze the patterns the algorithm finds and, if required, go back to a previous step to modify the KDD process' setup.

9. The last stage of the KDD in databases involves using the interpreted results for additional tasks, such as using them for more study or applying a system to a real-world problem.

The KDD process may involve numerous loops and iterations, as mentioned in Step 8. For example, after analyzing the results of the algorithm, one may decide that the selected method was incorrect and go to step 5. Similarly, it might be assumed that pre-processing was done after Step 4 reduced the data to a representative structure.

1.1.2 Data Mining

In the KDD process, the step referred to as data mining is when the best approach and algorithm are chosen and applied to the data set [Fayyad et al., 1996], to find patterns or structures present in the data and use them to classify it into various classes [Mikut, 2008; Frické, 2009], and, in the long run, to build knowledge and wisdom (see Fig. 1.2). Depending on how much knowledge an algorithm has about the many classes currently existing in data collection, data mining methods can be divided into distinct categories [Battula et al., 2013]:

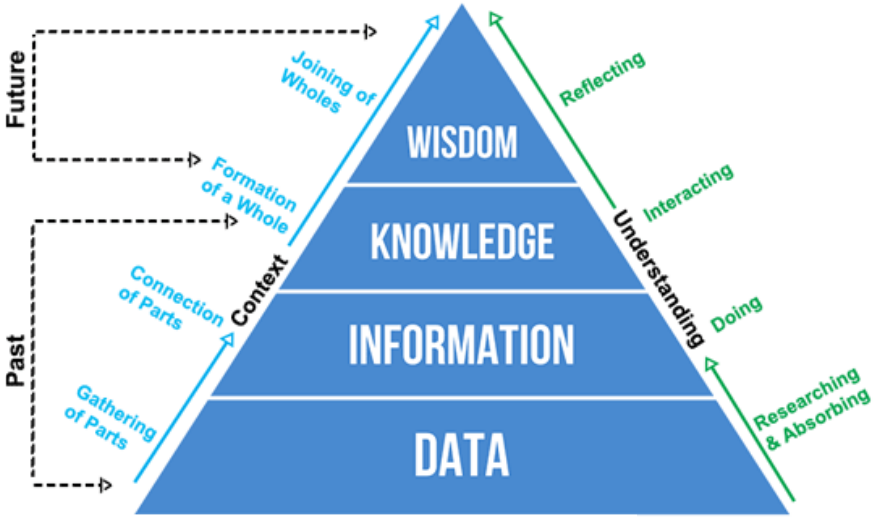

Fig. 1.2 The data mining pyramid. Information comes from data, knowledge comes from information, and wisdom comes from knowledge. From [Maimon et al., 2009]

- Supervised learning: involves any method in which the algorithm receives input and output data to work with. The external information the algorithm is allowed to use, such as feature values and meta-data, is known as the input values. In contrast, the output values are the exact labels of the classes. Therefore, the data format is decided beforehand, and the algorithm's objective is to classify the new input data into relevant classes.

- Unsupervised learning: encompasses any method that does not have access to output values and, as a result, attempts to discover structures within the data by independently creating classes. This is in contrast to supervised learning, which has access to output values.

The two main objectives of data mining are verification and discovery. While discovery looks through the data for patterns that have never been noticed before, verification aims to show that the user is right in his or her premise. The two sub-steps of the discovery process are description and prediction. First, the model searches for patterns in the data to display them in a way that is understandable during the description sub-step. Next, the model attempts to predict future data outcomes using patterns during the prediction substep. It is possible to divide the subgroup into regression and classification tasks. Regression tasks generate continuous values as output instead of classification tasks, which generate fixed labels for each class attribute associated with each data record [Tascini et al., 1996].

One of the most significant distinctions that can be made between people and computers has been that people have a natural tendency to become better at solving problems as time goes on. People can reflect on past mistakes and use that knowledge to improve their problem solving methods or develop

entirely new ones. Because traditional computer programs do not consider the results of their activities, they cannot modify their actions in response to new information. This specific issue is addressed head on within the discipline of ML, which involves the development of computer programs that can acquire knowledge and, as a result, enhance their capabilities by accumulating additional data and experience.

1.1.3 Medical machine learning

Today, machine learning models are used for various tasks in both the public and private sectors. Healthcare, self-driving, and decision making are just a few examples of these applications. Recent scientific developments have focused on deep learning, a subclass of ML algorithms. Given that they are modeled after the human brain, they are utilized to learn challenging tasks. In reality, neural network based algorithms have achieved state-of-the-art performance in several benchmark data sets, including handwritten character identification [Ruiz-Garcia et al., 2017], sentiment analysis [Deng et al., 2009], object classification [Voulodimos et al., 2018], heart disease detection [Aljanabi et al., 2018], sleep apnea detection [Mostafa et al., 2019], proteins folding [Ruff et al., 2021], etc.

There is potential for improvement; neural networks are now as capable as humans in some tasks because of their increased skill. For example, the reigning European champion of the traditional game of GO was defeated by the deep learning algorithm known as AlphaGo [Silver et al., 2016]. Furthermore, a deep learning system made public in 2016 can perform as well as humans on an intelligence test (IQ), if not better. The use of the Google neural network to detect cancer metastases in the medical field, which outperformed a clinical expert [Deng et al., 2009], is a great example.

Understanding the functioning state of the human body and accurately diagnosing and prognosizing any disease requires extracting important data-related insights. Sensors typically capture physiological data and are then used in disease diagnosis and patient physical therapy [Mporas et al., 2015]. Due to recent advances in medical technology, we can now obtain vital sign data from patients treated in a hospital setting or with wearable devices for continuous monitoring. These data include vital signs, events, photographs, waveforms, test results, and a variety of others [Vijayan et al., 2021]. Unfortunately, most of the collected data has not been used well and the healthcare industry often cannot fully utilize what the data offer. The data collected can be used to perform classification and regression analyzes.

Furthermore, it is critical to study deterioration events and see if there is a signature that can be used to identify them. One of the issues discussed here is the development of forecasting algorithms

that recognize and account for this signature. A prediction model of this type could be brought to the patient's bedside and reduce the burden on the healthcare system [Ghassemi et al., 2020].

The physiological state of a patient can worsen unexpectedly and without prior warning. When something like this occurs, the standard of care in the healthcare industry currently dictates that a medical expert must act quickly to assist the patient. The term "reactive patient care paradigm" describes this kind of practice. The early warning systems in place now follow a set of predetermined guidelines. Unfortunately, it is difficult for a human clinician to analyze continuous and quantitatively complicated medical data in a practical clinical context. Furthermore, there is a significant gap between academic research and actual clinical use [Bennett et al., 2013].

The current healthcare system used in Europe and the rest of the world has drawn criticism from the European Association for Predictive Preventive and Personalized Medicine [Golubnitschaja et al., 2014]. In addition, it has motivated its delegation to look at cutting-edge healthcare research to discover solutions to the reactive nature of the current healthcare system. The investigation of the application of ML methods to predict and, more specifically, prevent potentially dangerous clinical events has been motivated by all these concerns.

1.2 Thesis objectives

This doctoral research aims to investigate and construct methods of efficient and generalizable machine learning (ML) frameworks to improve the accuracy of disease prognosis and detection using physiological time-series data that could be recorded from affordable wearable devices with little or no expert intervention. Furthermore, the construction of such methods could be used as an aid in the early intervention of unfavorable clinical conditions such as atrial fibrillation (AF) and obstructive sleep apnea (OSA). The following research objectives must be met to achieve this goal.

- Analyze the different ways in which various forecasting approaches can be implemented using various cutting-edge ML algorithms to anticipate and detect potentially life-threatening clinical events based on physiological time series data.
- Study the applicability of ML algorithms to analyze time-series data from various vital sign sensors to predict and detect clinical events in the context of AF and OSA.

1.3 Summary of contributions

Our contribution is broken down into five different chapters. Starting with Chaper 1 where we explain the motivation and objectives that led us to conduct this research. The remainder of this thesis has been broken into separate chapters and is ordered as follows.

In Chapter 2, we present an overview and methodology of data-driven ML methods for biomedical data. Specifically, this chapter introduces the technical fundamentals of the ML approaches implemented in Chapters 3 and 4, as well as the data pre-processing and performance evaluation.

In Chapter 3, we focus on detecting early warning signals of atrial fibrillation (AF). This chapter begins with a biological review of the structure of the heart, the function of the heart, the pathophysiology of AF, the risk and treatment of AF, and the use of electrocardiograms (ECG) or the R-to-R interval (RRI) to assess heart activity. A detailed review of the literature on the field's current state is presented at the end of this section. Next, we present our contribution: a method to detect early warning of AF, called WARN (Warning of Atrial fibRilatioN), which is based on a deep convolutional neural network and recurrence plots of the R-to-R interval data. Finally, this chapter shows the results obtained and the discussions.

In chapter 4, we focus on the detection of events of obstructive sleep apnea (OSA). This chapter begins with a review of sleep disorders and the use of polysomnography (PSG) to collect data from multiple sensors that monitor different physiological vital signs and evaluate sleep disorders. Then it expands on the pathophysiology of OSA, and the risk and treatment of OSA. A detailed review of the literature on the current state of the field is presented at the end of this section. Next, we present our contribution: a data-driven method for detecting OSA events from physiological time series data. Our approach eliminates the requirement of separate feature extraction and selection processes by developing a hybrid learning model that combines deep neural networks and a light gradient boosting machine. Finally, this chapter shows the results obtained and the discussions.

In Chapter 5, we present our conclusion, highlighting our findings and the important contributions we made. In addition, this chapter details the research findings that were disseminated and offers a view into future research.

Chapter 2

Data driven methods for biomedical data

This chapter provides the technical foundation for the subsequent chapters of the thesis. First, it gives an overview of machine learning by explaining its fundamental concepts and the two most common approaches: feature engineering and feature learning. Furthermore, it expands on the methods that are used in this thesis in the context of early warning of Atrial Fibrillation and the detection of events of obstructive sleep apnea. Finally, it describes the data preprocessing techniques and the evaluation methods and metrics discussed throughout the thesis.

2.1 Machine Learning

The idea of machine learning (ML) was first proposed in 1952 [Samuel, 1967]. ML was initially only a small group of engineering methods and statistical studies [Mitchell, 2006]. Nevertheless, it has grown over time into a sizable subject rich in applications and theory. The terms "ML" and "data mining" can be used to refer to a variety of distinct concepts or ideas [Alpaydin, 2020]. A common interpretation of ML is to use statistical theory to create mathematical models that allow computers to draw conclusions from data without explicitly programming [Alpaydin, 2020]. Another typical interpretation of ML is that it is related to the use of ML in applications. This further emphasizes the modeling component of data mining, defined as the entire process of data collection, preparation, preprocessing, modeling, and interpretation. Modeling has the same significance as other processes because none would be viable or relevant without the others, particularly in manufacturing, data mining, and machine learning [Alpaydin, 2020].

Machine learning methods are typically classified according to the learning techniques that lie beneath them. These learning techniques are frequently determined by the level of inference that can be drawn from the data by the computer program [Jukes, 2018]:

- Rote learning: This approach is used in all classic computer programs. Since the application cannot draw any inferences or make any transformations based on the information provided, it is the responsibility of the programmer to implement all their knowledge directly.
- Learning from instruction: It refers to the conversion of information from the language used to input it into an internal language of computers. Although the programmer still provides information on how to successfully carry out this transformation, very little inference is required from the computer program for this to work. In contrast to learning by rote, this shows another level of the learning system.
- Learning by analogy: It entails numerous transformations of previously acquired knowledge to produce new skills that are almost exact replicas of previously held skills and, therefore, are easy to learn. It introduces novel aspects for which the original computer program was not intended.
- Learning from examples: It is one of the most popular learning methods in the modern era. This is because it offers the highest degree of flexibility and allows computer programs to learn entirely new skills or discover previously undiscovered structures and patterns in data [Carbonell et al., 1983]. Learning from examples is a common technique in data mining and classification tasks. This approach aims to predict the class label of incoming data entries based on a group of continuously changing instances previously seen.

A successful ML method requires a meaningful transformation into a mathematical model and a precise characterization of the problem in the relevant domain [Mikut, 2008]. The questions could be asked before or after the data collection procedure.

2.1.1 Data collection

Data collection is the process of collecting raw data. This procedure could involve taking manual measurements of some physical parameters or employing sensors, cameras, or similar tools. Raw data is collected and saved in various formats, such as text files, CSV files, and SQL databases. Data preparation is the process of combining data from various sources into a single standard format and occasionally changing the names of those sources to make the data easier to visualize or analyze.

The statistical characteristics of the data have not yet undergone any changes. The stages of data preparation and modeling are interconnected. The goal of "data preprocessing" is to change how the data is represented so that it can be used in later stages of the machine learning (ML) process.

Processes that include feature extraction, selection, normalization, etc., are frequently used in data preprocessing. The statistical properties of the data will be modified during this phase. Modeling is the practice of using machine learning (ML) algorithms to extract statistical regularities from input data to provide predictions. These algorithms include conventional ML techniques and more sophisticated neural networks (NNs). Evaluation and visualization of the significance of the data to aid in decision making is the process known as "interpretation of results." This is achieved by combining the conclusions drawn from the data analysis with an understanding of the particular domain of the problem and its mathematical description. The selection of ML algorithms is greatly influenced by feature extraction, which can be divided into feature engineering and feature learning.

Data preparation requires hard work in ML since the generated data features or representations are essential to efficient ML modeling [Bengio et al., 2013]. The ML modeling process can be divided into two groups [LaCasse et al., 2019]:

- Feature engineering combined with traditional machine learning.
- Feature learning from neural networks.

No distinct line can be formed between the two groups from a mathematical or even methodological perspective [Zhou, 2021]. This is an important fact. These two groups subscribe to opposing philosophical viewpoints. Unlike the latter, the former emphasizes the understandability of the extracted features, ML methods, outcomes, and the integration of domain expertise into feature engineering. However, the latter approach requires less work to use domain expertise to extract features. In addition, it creates general ML algorithms that may be applied to various applications.

2.2 Feature engineering combined with traditional machine learning

Feature engineering (FE) is the practice of modifying features to make machine learning (ML) algorithms work using domain expert knowledge or mathematics [LaCasse et al., 2019]. The representation that describes the data significantly affects how well the ML algorithms perform when applied. Feature engineering combines human ingenuity and prior knowledge to account for the fact that some ML models are ineffective and delicate enough to overlook subtle discriminatory features hiding in the data [Bengio et al., 2013]. Feature engineering techniques include the following.

- Use mathematical operations to create new single features from single or multiple raw features [Waring et al., 2020].
- Single feature extraction from data with temporal or spatial patterns, such as segmenting [Junno et al., 2004], subsampling [Park et al., 2004], computing statistic attributes such as maximum, variance [Zhang et al., 2015], filtering [Jain et al., 1995], frequency domain [Hamilton, 2020] or time-frequency domain [Hamilton, 2020], etc.
- Using the principle component analysis (PCA) [Jolliffe et al., 2016] or linear discriminant analysis (LDA) [Martinez et al., 2001], for example, to reduce the raw features to a lower-dimensional space.

Compared to feature learning, the pipeline consisting of feature engineering and traditional ML has the advantage of producing models that are more understandable and open to scrutiny. The results of the model are easy to interpret in terms of determining which features are more significant and how they affect the quality of the model.

In this thesis, the following ML methods were used as benchmarks to determine whether additional effort, such as developing and implementing more complex and computationally demanding ML models, would be worthwhile in terms of resources and performance.

2.2.1 Logistic regression

The classification method called logistic regression (LR) is often used to estimate the likelihood of a specific outcome given a set of input variables [Wright, 1995]. A binary outcome, which might have two values such as "healthy" and "not healthy," "yes" and "no," "true" and "false," etc., is the most common type of logistic regression model. It should be emphasized that, despite its name, it is not a regression algorithm. Instead, logistic regression uses estimated probabilities to determine the link between categorical independent variables (input) and dependent variables (output). LR is defined as follows.

$$L(x) = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}} \tag{2.1}$$

When dealing with binary data, it is normal to consider the existence of nonlinear relationships between x and $L(x)$. The sigmoid function $L(x)$ transforms the entire real domain into [0, 1]. This is required to make the outcome understandable as a probability. When $L(x)$ is closer to 0 or 1, rather than 0.5, a change in x often has less effect than when it is closer to either of the extremes. An LR

model is a generalized linear model with a binomial random component and a logit link function. As a result, logistic regression models are often referred to as logit models. Although x must, as stated above, be in the range [0,1], the logit can take any actual value.

When x increases, the sign of β determines whether $L(x)$ increases or decreases. The rate at which the function grows or shrinks is determined by the magnitude of $|\beta|$. When $\beta = 0$, the response variable $L(x)$ is independent of x. Given the symmetry of the logistic density, the function $L(x)$ approaches 1 at the same rate as it approaches 0.

It is possible to manipulate and calculate the logarithm of both sides of equation 2.1 to obtain the log-odds, which are linear in x. It should be noted that the LR model differs from the Linear Regression model in that the logarithmic odds change by β for every unit increase in x.

$$log(\frac{L(x)}{1-L(x)}) = \alpha + \beta x \tag{2.2}$$

The maximum likelihood method is preferred for calculating the coefficients in the logistic function. The idea is that we want to obtain estimates for α and β so that the probability obtained $L(x)$ in Equation 2.2 corresponds to the probability observed in the data set. Equation 2.3 represents the likelihood function of Equation 2.1.

$$l(\alpha,\beta) = \prod_{i:y_i=1} L(x_i) \prod_{i':y_{i'}=1} (1-L(x_i)) \tag{2.3}$$

Logistic regression can be extended to handle multiple variables. Let $x = x_1, \ldots, x_n$ be a vector of n variables. Logistic regression can be defined as follows.

$$L(x) = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots\beta_n x_n)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots\beta_n x_n)}} \tag{2.4}$$

The parameter β_i represents the effect of x_i on the logarithmic probabilities that $Y = 1$ controls the other parameters x_j. Consequently, at fixed values of the other variables x_j,$e^{(\beta_i)}$ is the multiplicative effect on the probabilities of a unit increase in x_i when all other variables x_j ($j \neq i$) remain constant.

2.2.2 Support Vector Machine

A support vector machine (SVM) is used for classification and regression analysis, in addition to data analysis and pattern recognition [Burges, 1998]. Given a collection of training examples with two distinct classes, the SVM training approach creates a model that divides subsequent examples into either one of the two categories or neither of the two, making it a nonprobabilistic binary linear

classifier. As points in space, samples are represented by the Support Vector Machine (SVM) model. The samples of the various categories are divided by a hyperplane that is as wide as possible according to how the samples are mapped. When implemented, input samples are mapped, and classification is performed depending on which side of the hyperplane instance the inputs lie on (see Fig. 2.1). The hyperplane is defined as follows:

$$f(x) = w^T x + w_0 = 0 \tag{2.5}$$

where w is the direction of the hyperplane, w_0 is its exact position, and $x(i = 1, \ldots, N)$ is the vector of features of the data set X, which corresponds to classes w_1 or w_2. The distance between two points and the hyperplane is given by:

$$z = \frac{|f(x)|}{||w||} \tag{2.6}$$

When w and w_0 are scaled, the value of $g(x)$ is equal to one at the nearest points for w_1 and negative at the nearest points for w_2. The goal of linearly separable data is to maximize the margin between the two classes.

$$\frac{1}{||w||} + \frac{1}{||w||} = \frac{2}{||w||} \tag{2.7}$$

Given the constraints:

$$\begin{aligned} w^T x + w_0 \geq 1, \forall x \in w_1 \\ w^T x + w_0 \leq -1, \forall x \in w_2 \end{aligned} \tag{2.8}$$

When $y = 1$ for $x_i \in w$ and $y = -1$ for $x_i \in w$, the objective function can be expressed as follows:

$$\begin{aligned} min \quad J(w, w_0) = \frac{1}{2}||w^2|| \\ \text{where:} \quad y_i(w^T x_i + w_0) \geq 1, \quad i = 1, \ldots, N \end{aligned} \tag{2.9}$$

Let $L(w, w_0, \lambda)$ be the Lagrangian function, and let λ be the vector of Lagrange multipliers:

$$L(w, w_0, \lambda) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i [y_i (w^T w_i + w_0) - 1] \tag{2.10}$$

The following constraints must be met to satisfy the Karush-Kuhn-Tucker (KKT) requirements for optimality [Gordon et al., 2012]:

$$\begin{aligned} \frac{\partial}{\partial w} L(w, w_0, \lambda) &= 0 \\ \frac{\partial}{\partial w_0} L(w, w_0, \lambda) &= 0 \\ \lambda_i [y_i (w^T x_i + w_0) - 1] &= 0 \\ \lambda_i \geq 0, \quad i &= 1, \ldots, N \end{aligned} \tag{2.11}$$

When Equation 2.10 is substituted in Equations 2.11, the result is as follows:

$$\begin{aligned} w &= \sum_{i=1}^{N} \lambda_i y_i x_i \\ \sum_{i=1}^{N} \lambda_i y_i y_i &= 0 \end{aligned} \tag{2.12}$$

The problem can then be written in its corresponding Lagrangian duality form:

$$\begin{aligned} & max_\lambda \left(\sum_{i=1}^{N} \lambda_i \frac{1}{2} - \sum_{i,j} \lambda_i \lambda_j y_i y_j {x_i}^T x_j \right) \\ \text{where:} \quad & \sum_{i=1}^{N} \lambda_i y_i = 0 \\ & \lambda \geq 0 \end{aligned} \tag{2.13}$$

Once λ is calculated, w and w_0 can be found using the equation 2.12.

When the data cannot be separated linearly, a variable ξ is included to penalize misclassified outliers. The objective function becomes as follows.

$$\begin{aligned} min \quad & J(w, w_0, \xi) = \frac{1}{2} ||w^2|| + C \sum_{i=1}^{N} \xi_i \\ \text{where:} \quad & y_i (w^T x_i + w_0) \geq 1 - \xi_i, \quad i = 1, \ldots, N \\ & \xi_i \geq 0, \quad i = 1, \ldots, N \end{aligned} \tag{2.14}$$

Here, C is a positive constant that balances the width of the margin and the number of misclassified points. The equivalent dual representation is written as follows:

$$\begin{aligned} & max_\lambda \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j {x_i}^T x_j \right) \\ \text{where:} \quad & \sum_{i=1}^{N} \lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq C, \quad i = 1, \ldots, N \end{aligned} \tag{2.15}$$

When the number of data in each class is unbalanced, the penalty parameter C for each class is multiplied by a weight proportional to the size of that class. The objective is to aggressively penalize outliers from the class with fewer data, reducing the classifier's bias towards the class with more data. A kernel function is frequently used in the nonlinear scenario to project the input feature space into a higher dimension, where a hyperplane more easily separates the classes. This is achieved by defining the inner product of the vectors in the higher-dimensional space as a function of the inner product of the vectors in the original feature space [Suthaharan, 2016].

Using the kernel technique, SVM can be generalized to compute nonlinear decision boundaries when dealing with binary classification issues that do not have a simple hyperplane as an effective separation criterion [Burges, 1998]. Furthermore, when kernel functions are used, it is possible to compute a separate hyperplane without explicitly performing the map in the feature space.

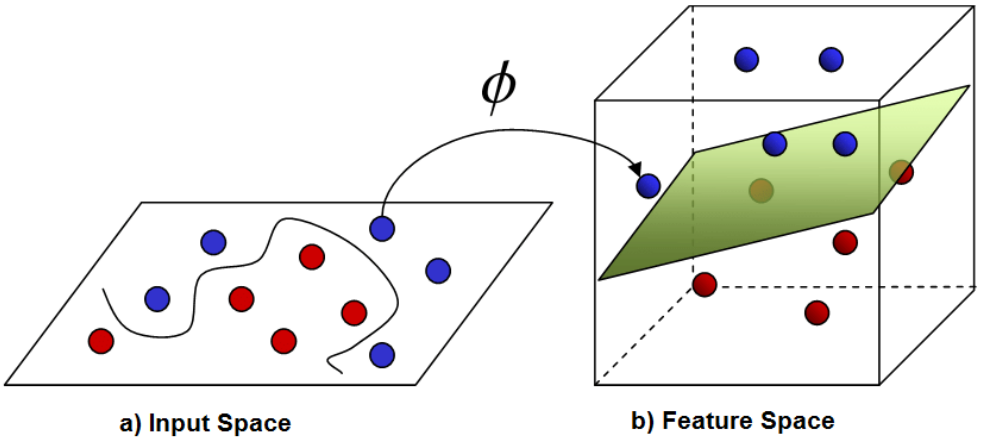

Fig. 2.1 Support vector machine. From [Vocaturo et al., 2019].

As mentioned earlier, a support vector machine's goal is to build a hyperplane or sequence of hyperplanes in a high dimensional space that can be used for regression and classification. The hyperplane that successfully achieves an ideal separation is the one with the largest distance from the closest training data sample of any class (also known as the margin). This is so that the classifier's generalization error can be reduced [Suthaharan, 2016]. It is typical for data that need to be discriminated not to be linearly separable, even if the initial challenge may have been described in a space with a finite number of dimensions. It was suggested that the finite-dimensional initial space be mapped onto a much higher-dimensional space to facilitate separation. The SVM schemes' mappings make it simple to compute the dot products in terms of the variables in the original space. This is accomplished by defining the mappings as a kernel function chosen to fit the problem. This helps to limit the computational load to manageable levels [Suthaharan, 2016].

2.2.3 Ensemble Methods

Ensemble modeling is a method in which numerous different models are combined to address a specific task. This can be accomplished by employing various modeling techniques or multiple training datasets. The ensemble model then adds the results of each model's forecast to get a final prediction for occurrences that have not yet been seen. The goal of adopting ensemble models is to reduce the variation of the predictions and the amount of generalization error by applying a divide-and-conquer strategy.

When the predictions are combined, more accurate and resilient models almost always result. This is because there is no longer a requirement for the high degree of fine-tuning necessary for single-model solutions. The models used in the combination process are typically taken from the same algorithm family. Decision tree-based ensemble methods are the most commonly used in the literature [Ardabili et al., 2019].

2.2.4 Decision tree

A decision tree is a nonlinear supervised classifier that takes the form of a tree and consists of a succession of decision stages completed until a final class is selected. Each block in Figure 2.2, known as a decision node, represents a question asking whether a particular characteristic is greater or lower than a value. The responses "yes" and "no" partition a node into two distinct subnodes, and there is no overlap between them. The entire training set is connected to the decision tree's first node, the root node. A leaf node is a node found at the end of a branch and includes data from the same class as the parent node.

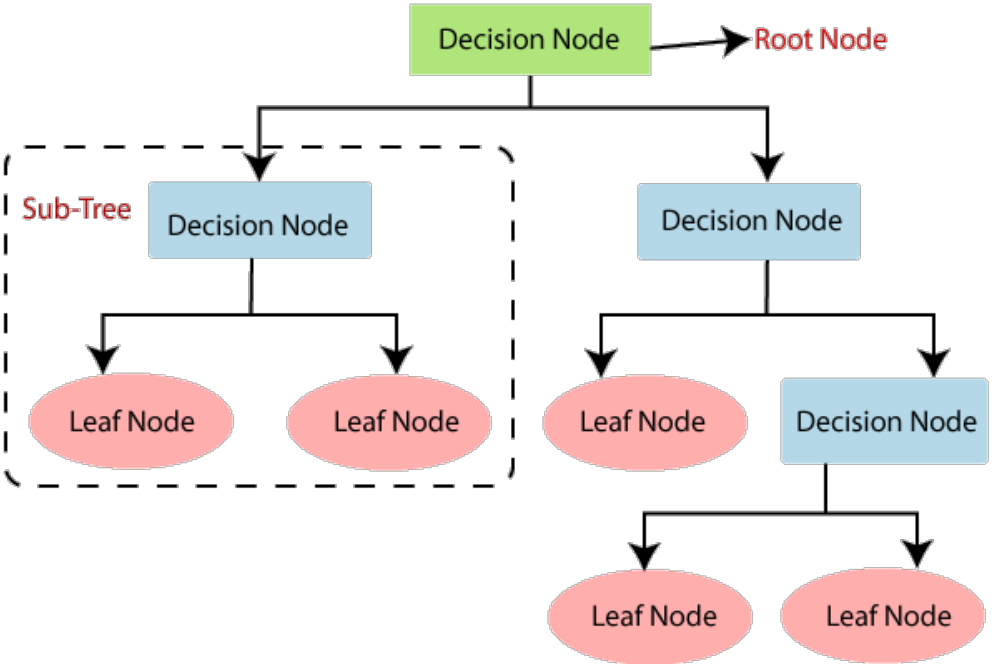

Fig. 2.2 Decision tree. From [Janikow, 1998].

In a general sense, the goal of decision trees is to partition the data in such a way that, following the partitioning, the data originating from the various classes are separated as much as possible into their respective child nodes. In other words, the purity of the data in the child nodes increases compared to that in the parent nodes as the tree is traversed. "ID3," "C4.5," and "CART" are examples of criteria that are typically applied when separating data at each decision node [Izza et al., 2020]. The Gini index is used to generate binary splits. A lower value of Gini indicates a higher degree of homogeneity. Assuming that the proportion of individuals belonging to class $k-th$ in the data set D is P (where k = 1, 2,..., |k|), the Gini index of the data set D can be calculated as follows:

$$\begin{aligned} \text{Gini}(D) &= \sum_{k=1}^{|N|} \sum_{k' \neq k} P_k P_{k'} \\ &= 1 - \sum_{k=1}^{|N|} {P_k}^2 \end{aligned} \tag{2.16}$$

This indicates the probability of randomly selecting two different samples from the data. D stands for "data set." If we assume that a feature can take on one of the potential values of V $(a^1, a^2, ..., a^V)$ and use it to decide how to divide the data set D, we can anticipate that there will be V possible splitting points. The Gini index of a feature a is defined as follows:

$$\text{Gini}(D) = \sum_{k=1}^{V} \frac{|D^v|}{D} \text{Gini}(D^v) \tag{2.17}$$

where $\frac{|D^v|}{D}$ is the weight assigned to the dividing point a. As a result, in the set of features A, the optimal splitting a^* is the one that results in the minimum Gini index after splitting.

$$a^* = \min(\text{Gini}(D,a)), \quad a \in A \tag{2.18}$$

The following summarizes the algorithm used to grow decision trees.

1. The method starts with the root node, where all training data is stored.
2. The Gini index is calculated for each feature and value, which is done for each node. Two descendant nodes are generated following feature selection and the associated splitting value that produces the lowest Gini index.
3. Declare a node to be a leaf if all of the ancestor data belong to the same class, the cardinality of the ancestor data is low enough, or the depth of the branch reaches a specific value. In that case, proceed to the previous step.

The decision tree algorithm can process data of both numerical and categorical types. Therefore, it is effective even when applied to large datasets [Izza et al., 2020]. However, there is a risk of overfitting when using decision trees. This problem can be alleviated by setting a maximum depth for the tree. However, this may result in a higher error rate due to bias. Random forest is another approach that can be taken to address the overfitting problem.

2.2.5 Random Forest

The random forest algorithm involves the construction of many decision trees during training. When several reasonably uncorrelated models are worked together, they can outperform any one of them working alone. Unlike a single tree, which can be excessively deep and has a propensity to have a low bias and a high variance, several trees working together can average the different findings, reducing the variance, and increasing the bias. In practice, there is usually a significant reduction in variance. When new information is available, the output of a random forest is the decision tree that receives the most support from the total population. On the other hand, a random forest has a lower risk of overfitting than a decision tree [Ghojogh et al., 2019].

The formation of random forests is based on developing a wide variety of decision trees. Decisions that go into building a decision tree are based on selections based on random replacements made from the entire data set. In doing this, we ensure that the data used in each decision tree are unique. Because decision trees are susceptible to the data used for training, even relatively small changes could result in significantly different tree architectures. Additionally, the feature selection process adds another element of randomization to the random forest. In a decision tree, the optimal feature for each node is selected from a subset of the entire set of features that have been chosen at random. In the absence of random feature selection, particular predictive traits would be selected in many base trees, causing these trees in the forest to become strongly associated with each other.

In general, decision trees, also known as weak learners, are used to complete random forest training by maximizing their parameters at each split node j through:

$$\theta_j^* = max(I_j) \tag{2.19}$$

Then, when each decision tree has been trained successfully and independently, in the case of classification issues, all "weak" forecasts are merged into a single forest prediction via an averaging process as follows:

$$p(c|v) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|v) \tag{2.20}$$

Where T denotes the total number of decision trees, v denotes an instance of an attribute, and $p(c|v)$ denotes the posterior ensemble probability distribution of every instance of the attribute of a discrete class [Denil et al., 2014]. In other words, by returning the entire class distribution, the classification trees produce a probabilistic output.

The concept of bagging is combined with the creation of random forests. Bagging is an acronym for aggregating boot straps [Ghojogh et al., 2019]. It is an ensemble algorithm that seeks to improve the accuracy and resilience of machine learning (ML) algorithms. In most cases, Bagging is used on less capable learners to enhance overall efficiency. Furthermore, it helps avoid overfitting while reducing variance [Breiman, 2001].

When there are n training samples, each of which has x features, a random forest, rather than training the model with a feature space equal to nx, creates a random subspace of features for each training space and then trains the model using that feature space (see Fig. 2.3). The feature space is chosen at random once more before the next iteration, at which point it is used to train the model. The

term "Random Forest" refers to the selection of subspaces at random to generate random feature spaces that resemble forests. Instead of searching for an essential feature to split a node, the best feature of the randomly selected subset of features is utilized to split the tree by building the node.

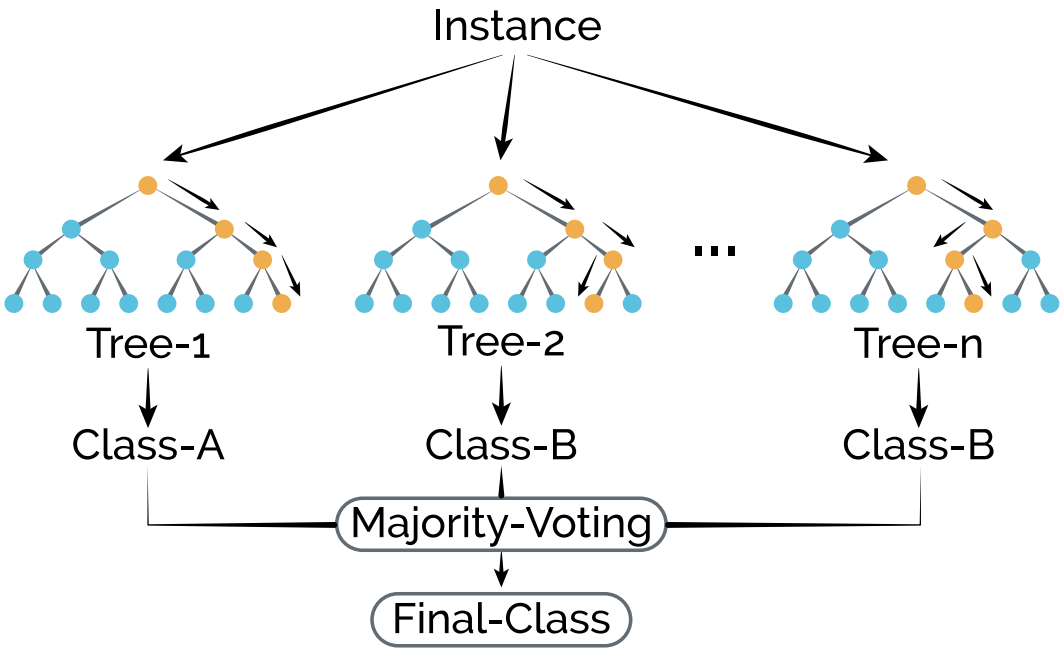

Fig. 2.3 Diagram of a random forest.

2.2.6 Extreme gradient boosting

Extreme gradient boosting (XGBoost) is an alternate version of gradient tree boosting. The approach of tree ensemble boosting, known as gradient tree boosting, works by combining several less effective classifiers into a single more effective one. Beginning with a basic learner, an iterative training process is used to create a powerful learner. Both gradient boosting and XGBoost operate according to the same fundamental principle [Friedman, 2001]. However, most of their differences lie in the specifics of their respective implementations. Better performance can be obtained with XGBoost by regulating the complexity of trees by applying various regularization techniques [Chen et al., 2016].

Let $(x1, y1), (x2, y2), ..., (x_n, y_n)$ represent a set of inputs and outputs. The tree ensemble algorithm employs K additive functions to predict the result, representing data at each decision node. The sum of the predictions for each function produces the expected result, as shown in the following equation.

$$\hat{y}_i = \sum_{k}^{K} f_k(x_i), \quad f_k \in F \tag{2.21}$$

As a result, the objective function is approximated given a set of parameters θ by minimizing the following regularized function:

$$obj(\theta) = \sum_{i}^{n} l(\hat{y}_i, y_i) + \sum_{k}^{K} \Omega(f_k) \tag{2.22}$$

Here, the first term $l(\hat{y}_i, y_i))$ is the training of the loss function, which calculates the difference between the expected and actual output. The training of the loss function can be quantified using the Mean Squared Error (MSE) for regression or the Logistic loss for classification [Chen et al., 2016], which are calculated as:

$$\begin{aligned} \text{mse} &= \sum_{i}^{n} (y_i - \hat{y}_i)^2 \\ \text{Logistic Loss} &= \sum_{i}^{n} [y_i log(1 + e^{\hat{y}_i}) + (1 + y_i log(1 + e^{\hat{y}_i}))] \end{aligned} \tag{2.23}$$

The second term Ω is the regularization term, which penalizes the complexity of the model to avoid overfitting. The regularization term in XGBoost is given by:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w^2|| \tag{2.24}$$

Where T is the leaf count and the second term is the leaf score L_2 norm [Bektaş et al., 2010]. The model trains additively, optimizing for one tree at a time during training. Let $\hat{y}_i^t$ be the prediction value in iteration t, and the additive process is as follows:

$$\begin{aligned} \hat{y}_i^0 &= 0 \\ \hat{y}_i^1 &= f_1(x_i) = \hat{y}_i^0 + f_1(x_i) \\ \hat{y}_i^2 &= f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i) \\ &\vdots \\ \hat{y}_i^t &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t x_i \end{aligned} \tag{2.25}$$

The tree added at each step maximizes the objective function and can be rewritten as follows.

$$\begin{aligned} obj^{(t)} &= \sum_{i} l(\hat{y}_i, y_i) + \sum_{k}^{K} \Omega(f_k) \\ &= \sum_{i} l[(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))] + \Omega(f_t) \end{aligned} \tag{2.26}$$

The objective function can be further condensed into a scorer function using a second-order approximation, as detailed in [Chen et al., 2016]. The score function is then applied to determine the quality of the tree structure. To avoid overfitting, shrinking is implemented in XGBoost. The shrinkage variable, commonly known as the learning rate, scales the weight of the features by a factor of η. Furthermore, XGBoost enables row and column subsampling, two techniques used in Random Forest to control bias and variance [Ghojogh et al., 2019].

The single most important factor in the overall success of the algorithm is its ability to scale performance. The algorithm performs 10 times faster on a single machine. It can scale to handle billions of instances even in memory-constrained or dispersed situations. Several substantial algorithmic and systemic improvements have significantly improved XGBoost's capacity to scale. These contributions include a novel tree learning approach for sparse data handling and a theoretically supported weighted quantile sketch procedure that allows the management of instance weights in approximation tree learning. In approximation tree learning, XGBoost can handle sparse data thanks to these approaches. Due to parallel and distributed computation, learning can occur more quickly, allowing faster model exploration. More importantly, XGBoost uses non-core functionality to its advantage [Chen et al., 2016].

2.2.7 Light Gradient Boosting Machine

Due to their efficiency, accuracy, and interpretability, boosted trees have recently gained prominence. However, the emergence of big data in recent years has presented challenges for the boosted trees. The information gained from each potential split point is estimated by a standard boosted tree considering all of the data's features. As a result, using boosted trees to work with large volumes of data takes a long time. The Light Gradient Boosting Machine (LightGBM) was created to solve this problem. The LightGBM algorithm employs two novel techniques, exclusive feature bundling and gradient-based one-sided sampling (GOSS), which differentiate it from traditional boosted trees. Both approaches aim to increase accuracy. The result is an algorithm that learns 20 times faster than previous boosted models while keeping the accuracy close to the same [Ke et al., 2017].

Information gain can be improved with larger gradients; hence, it is crucial to maintain it when downsampling. This is so that the theory underlying GOSS may be applied, which states that samples with larger gradients will increase the amount of information gained. Minor gradients will be eliminated randomly, and training may result in an improvement in estimation that is more precise than uniform random sampling while still maintaining the same desired sample rate. [Ke et al., 2017].

The fact that LightGBM uses a leaf-wise tree development strategy (see Fig. 2.4) rather than a level-wise tree growth approach is another way in which it differs from traditional tree-based models. The likelihood of overfitting is minimized when the level-wise growth technique is used.

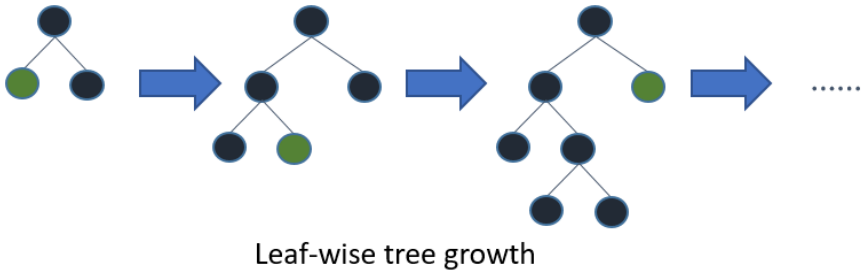

Fig. 2.4 Leaf-wise tree growth. From [Venkata Jagannath, 2017].

Although less adaptable, the leaf-wise approach often reduces loss, and its adaptability makes it a suitable choice for handling large data. Additionally, LightGBM processes data using histogram-based algorithms as opposed to pre-sort-based methods. When histogram-based methods are used, the cost of computing the gain of individual splits is diminished. The algorithm will divide features with continuous values into discrete bins and use the bins to generate feature histograms while training. Therefore, the time required to carry out the work after the histogram has been created will be proportional to the number of bins, which will be less than the entire quantity of data. [Ke et al., 2017].

2.3 Feature learning from neural networks

A feature learning algorithm, also known as representation learning, can learn to recognize and separate the underlying explanatory elements embedded in the data. The main benefit of this approach is that the learning algorithms become less dependent on feature engineering, allowing faster development of new applications. [Bengio et al., 2013].

Probabilistic graphical models and neural networks are two distinct but related areas of research in feature learning [Bengio et al., 2013]. The first approach is used to identify latent random variables that characterize a distribution. In contrast, the latter approach extracts abstract features from the data using a computational graph.

2.3.1 Neural Networks

To understand how the human brain functions and how neurons interact with each other, the researchers developed a type of ML model known as a neural network (NN) [Albawi et al., 2017].

Neural network algorithms can learn hierarchical representations of data using nonlinear information processing [Learning, 2020]. The basic computational components known as nodes or neurons make up a NN. A neuron receives information along the edges coming into it, multiplies that information by the weights associated with those edges, and then uses the weighted sum of that information to apply a nonlinear function called an activation function to produce an output. The vector equation 2.27 can mathematically depict the process.

$$y(x) = f(w \circ x + b) \tag{2.27}$$

The input vector, the weight vector, the bias of the neuron, multiplied by the element, the activation function, and the output of the neuron are all denoted by the symbols $x, w, b, \circ, f$, and y, respectively. Typical activation functions include the sigmoid function (S), the tanh function ($Tanh$), and the rectified linear units ($ReLU$) [LeCun et al., 2015].

$$S(x) = \frac{1}{1+e^{-x}} \tag{2.28}$$

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.29}$$

$$ReLU = max(0, x) \tag{2.30}$$

All neurons in one layer are connected to neurons in the next layer by a set of directed edges. The component has a weight assigned to each vertex. The input layer, which is the initial layer, processes the data. The output layer, the last layer, calculates the NN output. The phrase "hidden layers" refers to all still-existing layers together. The structure of the layers, as shown in Figure 2.5, supports a hierarchical arrangement, since the information passes from the input layer to the output layer.

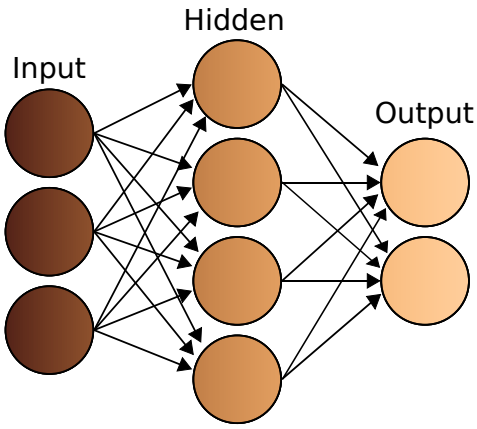

Fig. 2.5 Artificial neural network diagram.

The input and output data are shown to the network during training. A loss function is used to simplify learning. This function evaluates the gap between the output of the network and the desired output. The most popular loss function used to solve a classification problem is the categorical cross-entropy loss function (CE). Equation 2.31 provides the formula for calculating CE, taking into account the total number of observations (n), the measured value (y_i), and the projected value ($\hat{y}_i$). The error or residual is the difference between the actual and predicted values. For multiclass classification situations where each class receives a distinct one-hot encoded value, CE estimates the average difference between the actual and predicted probability distributions for all classes and is the default loss function [Garavaglia et al., 1998].

$$L = -\sum_{i=1}^{n} y_i * log(\hat{y}_i) \tag{2.31}$$

2.3.2 Neural networks training

The learning task is transformed into an optimization job (error minimization) when the NN parameters are changed to achieve the best results in minimizing the loss function. The optimization technique used to train neural networks is known as gradient descent. The procedure known as gradient descent includes calculating the gradients of the loss function associated with network parameters, such as weights and biases. The technique to calculate gradients is called backpropagation and is based on the chain rule of derivatives [Rumelhart et al., 1986]. In this case, the gradient is a measurement that shows how much the loss value alters in response to a small change in one of the network parameters. Equation 2.32 states that the learning rate (γ), a scalar variable, is used to update the parameters (θ) in a direction opposite to the gradient. Analyzing the practice data allows for iterative execution of the

technique. The epoch refers to a cycle through the training data. The parameters gradually reach their ideal values with each iteration, reducing the loss function.

$$\theta = \theta - \gamma \frac{\partial L}{\partial \theta} \tag{2.32}$$

Computing the loss and gradient throughout the entire dataset while dealing with a large data set will take too long and be computationally inconvenient. Because of this, stochastic gradient descent (SGD), a version of gradient descent, is commonly used in real-world computing. The parameters are adjusted once the loss function has been computed for one batch of the SGD data, separated into smaller groups known as batches. Other popular variations are RM- Sprop, AdaGrad, and Adam [Ruder, 2016]. Some of these variants use an additional parameter, called decay, whose function is to progressively reduce the learning rate as the parameters get closer to their ideal values.

When training neural networks (NNs), overfitting is often a problem. Overfitting occurs when a model tries to fit the noise in the training data. This usually occurs as a result of using a model that is more complex than is required. When overfitting, the model works well with the data was trained on, but does poorly in unknown data. Several methods can be used to prevent overfitting. When early stopping is used, a small training data sample is first held and used as a validation set. The value of the loss function in the validation set is compared to the value in the training set after each iteration of the training process. If the loss in the training set decreases but increases in the validation set, overfitting is noted and the model training procedure can be stopped. Another tactic commonly used in DL is dropout. In the dropout learning approach, a preset percentage of links of the NNs are randomly deleted after each training epoch [LeCun et al., 2015].

2.3.3 Universal approximation theorem

The universal approximation theorem states that a neural network (NN) can arbitrarily approximate any continuous function H as long as it has enough neurons and at least one hidden layer [Hornik et al., 1989]. This was first shown to work for feedforward networks with a hidden layer activated by a sigmoid function, but it has now been shown to work for all non-polynomial activation functions [Schäfer et al., 2006].

The theorem does not specify the number of neurons required or the conditions under which the network must operate to minimize the gap between the function H and the network approximation $\hat{H}$. The traditional method for optimizing network parameters is still gradient descent, but important parameters such as the number of neurons and architecture need to be engineered [LeCun et al., 2015].

Since any problem can be transformed into a function, NNs can approximate a wide range of scientific problems, and understanding the function itself is not necessary [Leshno et al., 1993].

2.3.4 Recurrent neural networks

A subtype of NN called a recurrent neural network (RNN) is particularly created to handle sequential input. RNN processes the input sequence one element at a time while keeping track of a hidden state vector that serves as a memory for the data already processed. They learn to selectively recall relevant information, allowing them to spot dependencies that span several time steps. This enables them to make forecasts while considering both current and past data. The model automatically discovers all this information without prior knowledge of the data's cycles or temporal dependencies. RNN can analyze sequences of various lengths and eliminate the need for a fixed time window. Furthermore, it is possible to indicate the number of states that grow an RNN at a rate proportionate to the number of nodes [Grossberg, 2013].

The most crucial component of an RNN is a state vector in the hidden units. The sequence items that preceded them are stored in the memory of this vector. A feedback connection in NN binds the hidden neurons over time. The network as a whole has this relationship. The current sequence element, x_t, and the hidden state of the time step preceding it, S_{t-1}, are sent to RNN as input in time t. Subsequently, the activation function h_t is used to calculate the network output, while the hidden state S_t is updated. As a result, given $t' \leq t$, the current output h_t is based on all the inputs x'_t that preceded it. The weight matrix between the input layer and the hidden layer in a standard RNN is U. The letter W represents the weight matrix for the repeated transition from one hidden state to the next. V represents the weight matrix for the change from hidden to output, and b_s,b_h biases. Equations 2.34 contain a summary of the calculations that were performed.

$$\begin{aligned} S_t &= (U_{x_t} + W_{S_t-1} + b_s) \\ h_t &= \sigma(V_{St} + b_h) \end{aligned} \tag{2.33}$$

The softmax function, commonly used as the activation function for the output layer of a multiclass classification job, is represented by σ in 2.34. This function aims to maximize the chance that the classification is accurate. The softmax function ensures that the result is between 0 and 1 and that the total of these outputs is equal to 1.

The RNN unfolds during training, and copies of the model are created at each time step. RNN may be considered a multilayer NN and can be taught similarly to backpropagation. This technique

for training RNN is called backpropagation over time, or BPTT [Werbos, 1990]. The use of BPTT would enable NNs to develop long-range dependencies over any length of the sequence. To save the appropriate data in memory, the training algorithm should be able to learn and fine-tune the weights.

In the real world, RNN are notoriously difficult to train. Even when important inputs and outputs are separated by just 10 steps. It is now well accepted that the standard RNN cannot be taught to recognize dependencies over long timescales [Bengio et al., 1994; Hochreiter et al., 2001]. The error gradients must be backpropagated in numerous time steps to train a RNN using BPTT. The error is backpropagated because the error gradient is repeatedly raised by the same number. As a result, the gradients either grow to large sizes or gradually decrease to zero. These problems are described as "exploding gradients" and "vanishing gradients," respectively. The model learning process may not converge or take too long in situations like these. The size of the repetition edge weight and the specific activation function are used to define the precise nature of the problem. Vanishing gradients are more likely to occur when the sigmoid activation function (Equation 2.28) is used and the weight magnitude is less than one. On the other hand, if the weight magnitude is greater than one and the ReLU activation function (Equation 2.30) is applied, explosive gradients are more likely to occur [Pascanu et al., 2013].

Long short-term memory

The Long Short-Term Memory (LSTM) architecture, which was published in [Hochreiter et al., 1997], was motivated by the problem of vanishing gradients. LSTM networks have become popular because they are substantially more successful than conventional NNs in learning long-term dependencies. The LSTM method can learn relationships that span any length of time. Using an architecture known as the LSTM unit in place of a normal neuron, the LSTM can address the issue of fading gradients. A network of less complex nodes linked in a specified way forms the basis of a LSTM unit. The main components of LSTM are as follows:

1. A primary unit continuously connected to a weight unit is known as a constant error carousel (CEC). The expression shows that the recurrent connection is a feedback loop with a time step of 1. The internal state known as CEC activation is a repository of previously received information.

2. Input Gate: A multiplicative component that protects against the corruption of data stored in the CEC due to disturbances caused by irrelevant inputs.

3. Output Gate: A multiplicative unit that protects against damage caused by other units due to data stored in the CEC.

The CEC's input and output gates regulate access. The input gate learns when to pass the new information on to the CEC during training. Data is not allowed inside the input gate as long as it has zero value. The output gate learns when to allow CEC information to pass in a similar way. Information or activation is locked inside the memory cell when both gates are closed (activation close to zero). The recurrent edge with unit weight makes it possible for the error signals to flow through numerous time steps without running into the issue of vanishing gradients.

2.3.5 Convolutional Neural Networks

A subclass of neural networks called convolutional neural networks (CNN) was created to employ various arrangements of spatial information [LeCun et al., 1995]. They are made to operate using matrices as input, ranging from volumetric images and feature maps from previous layers to even higher dimensions [LeCun et al., 2015], as well as 1D signals in 2D and 3D images and videos. The stride, a predetermined step size, moves a convolution kernel over the input [Springenberg et al., 2014]. The convolution kernel produces a feature map in the next layer at each step. It balances and adds the input values before passing them through a non-linearity such as a sigmoid or ReLU activation function [Nair et al., 2010]. Each kernel that processes the input creates a unique feature map.

A digital picture represents visual data in a two-dimensional form that contains a sequence of pixel values arranged in a grid-like structure. For example, each pixel in a digital image signifies the brightness of a color, and the structure of a digital image is similar to that of a grid. The convolution layer can detect local features in specific picture regions by executing a dot product between the set of learnable parameters known as a kernel and the feature map. These regions can be distinguished from each other. The term "local receptive field" refers to the window of pixels that connect to each neuron in the buried layer.

Most of the time, the convolutional kernel size will remain constant during training while changing weights. This suggests that as a CNN is trained, its kernels will become detection methods for specific input features. Because they travel over the input, these feature extraction techniques are also spatially invariant.

Each concealed neuron will ultimately be able to assess its specific receptive field, according to the theory [LeCun et al., 2015]. Any CNN design has a hyperparameter that can be configured to change the size of the local receptive field. After the initial connection, the stride length adjusts the receptive field to scan all input pixels. Each hidden neuron is a link in the local receptive field of a particular input layer. As a result, the kernels based on the spatially adjacent subsets of the feature map of the

layer underneath it are convolved to compute the activation units of the convolution layer. This implies that if the input is slightly altered, the activation of the units will also be slightly altered to the same extent.

The pooling layers can be used for downsampling [Yamashita et al., 2018]. As a result, a network must learn fewer features, with some translation invariance. Convolution and pooling are both somewhat analogous processes. It could be considered as a window that moves around the input it receives. At each step, it achieves or exceeds the maximum value of the window for a maximum grouping operation or the average value for an average grouping operation.

The input layer identifies low-level features such as edges and curves, whereas successive convolutional layers generate high-level features. Because of this, the construction of a CNN should not be restricted to at most just one or a few convolutional layers. A network that applies more convolutions to the input data can extract the features that, according to the dataset, decide the output with more precision [Goodfellow et al., 2016]. Therefore, a comparably deep architecture consisting of numerous layers is necessary to capture the various degrees of abstraction.

Fully connected layers can often be found at the end of CNN. Although dense layers and connectivity are frequently used interchangeably [Huang et al., 2017], the latter describes densely linked convolutional networks. The final layer of a fully connected neural network typically contains the same number of output neurons as the overall number of classes [Yamashita et al., 2018]. The values of these output neurons can be thought of as the probabilities of the various classes to which the input may belong when fed into the softmax activation function. This interpretation is feasible given that the softmax activation function takes several classes into account [Kouretas et al., 2019]. A standard CNN design is shown schematically in Figure 2.6.

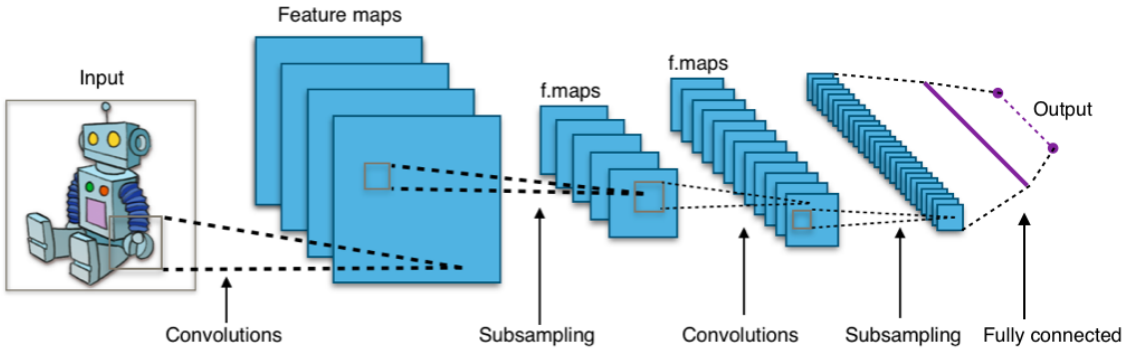

Fig. 2.6 Convolutional neural network diagram. From [Aphex, 2015]

2.3.6 Deep learning

In practically every area of data-driven research, machine learning (ML) and deep learning (DL) are applied. This research topic is concerned with developing sophisticated algorithms that can "learn" how to solve problems such as how people might develop their skills in particular fields. Once sufficiently "trained" to identify relevant features in datasets similar to the one under consideration, machine learning (ML) algorithms, more frequently referred to as "models," can classify data or make predictions about new target values. The phrase "Deep" refers to the number of layers or processing units the network possesses. Most of the ML tasks are now performed at the highest level by DL. In particular, in the fields of image processing and computer vision [Voulodimos et al., 2018].

DL uses learning-based algorithms to generate rules to interpret large data sets independently; this considerably boosts the algorithm's learning efficiency. The ability of NN to automate the process of feature creation and selection is the primary factor contributing to their efficacy [LeCun et al., 2015]. During the last decade, research in the field of ML has gained momentum on a global scale, propelling the development of a large number of sophisticated algorithms that are data-driven [Krizhevsky et al., 2012], allowing for easy adaption to a larger range of challenges [Schmidhuber, 2015]. As a result, there is an urgent need for intelligent algorithms capable of streamlined clinical procedures to help clinicians guide diagnosis and treatment. It has become the main driving force behind artificial intelligence research in the last decade, outperforming humans in a variety of challenging tasks such as facial recognition [Sun et al., 2014], natural language processing [Collobert et al., 2011], self-driving cars [Bojarski et al., 2016], and medical diagnosis [Xiong et al., 2018b]. Furthermore, DL is widely used for the analysis of biomedical data sets in the disciplines of neurology [Qi et al., 2017], pulmonology [Kamnitsas et al., 2017], most notably, cardiology [Matias et al., 2021; Aljanabi et al., 2018] and sleep apnea detection [Mostafa et al., 2019].

The outstanding performance of DL in computer vision tasks has attracted the attention of researchers interested in exploring their potential in medical imaging and pathological images [Castiglioni et al., 2021]. Furthermore, DL has allowed new avenues to be explored in medical image analysis by making it possible to discover representations automatically rather than depending on hand-made features specific to the problem, which require a certain level of domain knowledge [Bury et al., 2021].

Residual Neural Networks

Iterative learning is a technique used by neural networks (NNs) to learn by training a data set and getting feedback on how it performs. The backward propagation of the feedback occurs during

backpropagation [Linnainmaa, 1976; Schmidhuber, 2015]. The weights of each layer are then adjusted based on their influence on the training error after the learning signal has been propagated backward through several layers. When the networks are sufficiently deep, a typical vanishing gradient problem occurs [Goodfellow et al., 2016]. Multiplications with values below 1 are often performed in this process. In deeper networks, these multiplications occur more frequently, making training more difficult because the learning feedback vanishes after numerous multiplications with values below 1. As a result, deeper networks are more difficult to train.

A neural network can skip one or more layers by making shortcut connections. This results in skipping levels in the gradient flow, allowing the gradients to propagate back across the network more quickly during backpropagation. A problem called, The degradation problem [He et al., 2016], occurs when adding additional layers to a sufficiently deep network, preventing the training error from growing. Instead of estimating functions $H(x)$, one can learn residual functions $F(x) = H(x) - x$ using shortcut connections. The same result can be obtained by learning $H(x)$ or the residual $F(x)$, then adding x. Studies showed that the network would learn the residual more quickly [He et al., 2016].

Because the weights of a network are altered at every training step, the output distribution of a weight layer is susceptible to change. This problem is called the internal covariate shift [Ioffe et al., 2015]. It is harder for a layer to learn a function because the input distribution is altered at each iteration. Batch normalization can decouple these layers by standardizing the input of subsequent levels. The optimization landscape is smoothed when batch normalization is applied, which aids in the training of NNs [Santurkar et al., 2018].

EfficientNet

A revolutionary family of architectures called EfficientNets was created to find scalable compounded network structures. Compounding scaling is a technique to effectively balance input resolution, network depth, and network width. The convolution process in the EfficientNets models can be separated into a point-wise convolution across channels and a depth-wise convolution on single feature maps [Chollet, 2017]. This method allows for a significant reduction in the number of multiplications that must be estimated [Sandler et al., 2018]. Inverted residual blocks, a distinctive feature of EfficientNets, are given a low-dimensional feature map from the previous block. It must be triggered linearly [Sandler et al., 2018] to keep all its information. The data from the bottleneck is then incorporated into a higher-dimensional space after the space has been expanded using a point-wise convolution. From this, it may be inferred that the information manifold is situated in a low-dimensional area of the activation

space. A ReLU activation function can be used to continue the depth-wise convolution and still add the required complexity [Sandler et al., 2018]. A pointwise convolution is used to project the filtered feature map back into a low-dimensional environment (see Fig. 2.8). This layer allows for the use of ReLU much like the one before it. To enable gradient flow, a bypass connection is built between the bottlenecks.

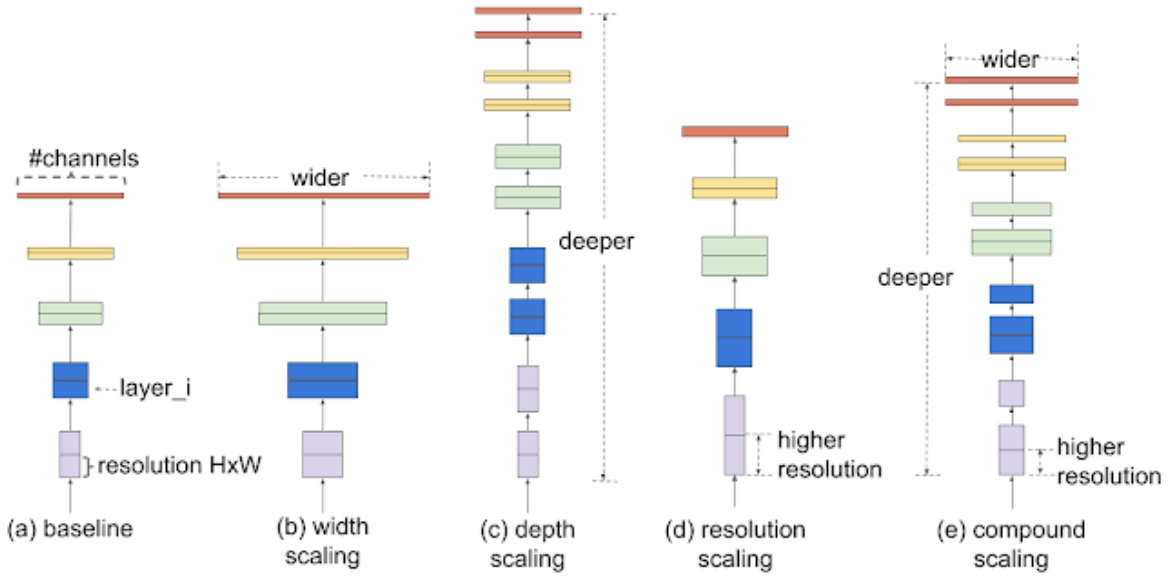

Fig. 2.7 EfficientNet scaling comparison. (a) Baseline network; (b)-(d) Examples of conventional scaling that enhances one of the network's widths, depth, or resolution dimensions. (e) method of compound scaling. From [Tan et al., 2019]

The core of EfficientNet is the compound scaling approach. Balances the number of layers, also called the depth of a network, the number of neurons per layer, and the input resolution of a network in a time-efficient way. Deeper networks are believed to be better at generalizing to new data and learning richer, more complicated features [Tan et al., 2019]. However, deeper networks contain diminishing gradients, making training more difficult. The shorter path will expand longer than the longer route even with shorter links and will encounter the same problem.

In most cases, the width of the layers will be scaled down when working with smaller models. The input resolution is the last parameter to be scaled using the compound scaling procedure (see Fig. 2.7). It seems to be the case that a higher resolution will be beneficial to the network for the simple reason that there will be more information present. The impacts of changing these hyperparameters eventually fade. Therefore, all three must be offset using the following equation.

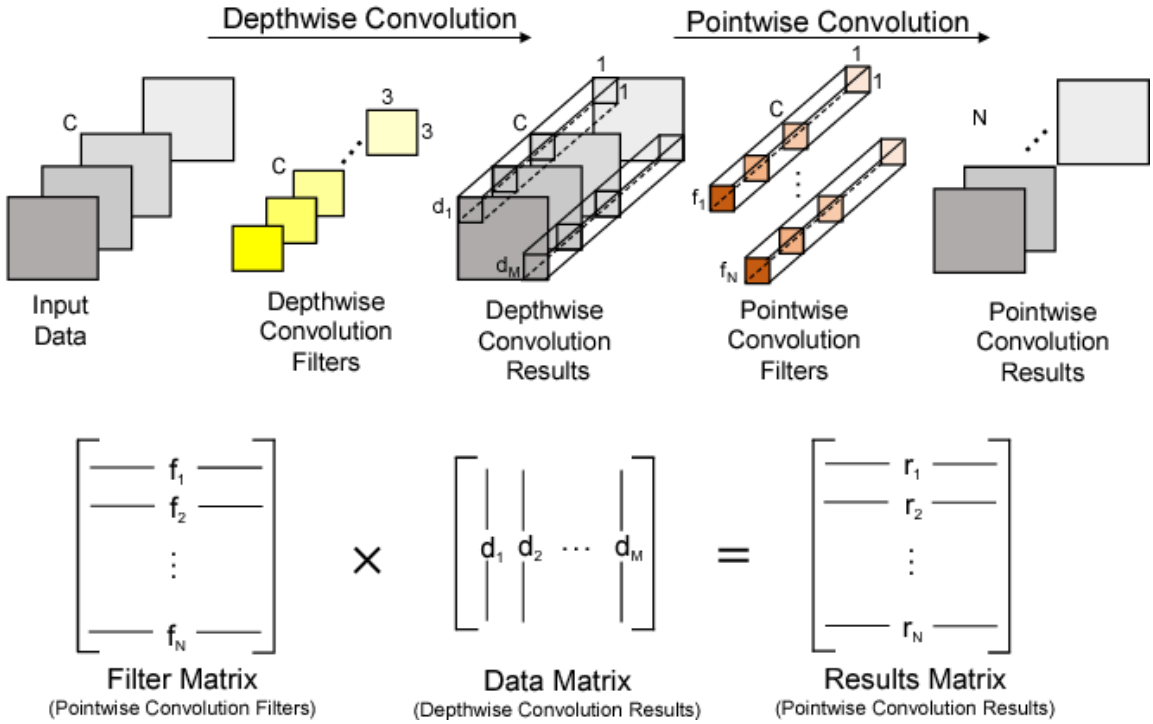

Fig. 2.8 Depthwise and Pointwise Convolution comparison. From [Tan et al., 2019]

$$\alpha \times \beta^2 \times \gamma^2 \approx 2 \tag{2.34}$$

Where depth = α^ϕ, width = β^ϕ and resolution = γ^ϕ as well as $\alpha, \beta, \gamma \geq 1$. ϕ represents the scaling factor. It represents the computing resources available for up-scaling, while α, β, and γ distribute additional resources between resolution, width, and depth [Tan et al., 2019].

2.3.7 EfficientNetV2

EfficientNetV2 is a revised and improved version of EfficientNet that was developed by Google in 2021 [Tan et al., 2021a]. Compared to earlier models, it has a faster training time and a higher parameter efficiency. The EfficientNetV2 architecture can train up to 6.8 times faster than most state-of-the-art models while being less deep. Training can be enhanced even more by gradually increasing the image size during the training process; however, doing so typically reduces accuracy. To compensate for this loss of accuracy, an adaptively adjusted regularization strategy (such as dropout [Wager et al., 2013] and data augmentation [Shorten et al., 2019]) has also been added. This strategy ensures that the network can achieve rapid training and high levels of accuracy [Tan et al., 2021a].

EfficientNetV2 is the architecture used in both Chapter 3 and Chapter 4. This NN consists of 479 layers and is used as input images of size 224×224 pixels. The last fully connected layer was adjusted to perform the classification between three classes in Chapter 3 and two classes in Chapter 4.

2.3.8 Grad-CAM

The decision-making process of a CNN can be explained using a method known as Gradient-weighted Class Activation Mapping (Grad-CAM). The method produces a map, which can then be overlaid over the original image to produce a heat map. The cornerstone of this method is the assumption that deep CNNs can learn to represent information hierarchically [Khan et al., 2020]. The early layers are in charge of extracting low-level features like oriented edges, while the subsequent layers integrate these features to create higher-level ones. The final feature map is considered to include the most detailed representation of the input image [Selvaraju et al., 2017; Islam et al., 2020]. The weight of the newly active final feature map can be determined by network gradients [Selvaraju et al., 2017]. This weighting is only used when the gradient is positive compared to the projected class. This is because positive gradients represent positive feedback. As a result of this procedure, a heat map is created; the heat map is a consequence of the gradients and the activations. This map indicates where relevant features are found and how essential these sites are to each other. This map is then adjusted to the input size of the network and placed on the input image. This is the most prevalent circumstance when the final convolutional layer contains many feature maps. These feature maps are given weights and averaged before the interpolation procedure begins. The result is an image placed on a heat map showing the most crucial places for network activation.

Figure 2.9 illustrates a summary of the method. The input is first processed by CNN, followed by task-specific calculations to create a numerical rating for the image and a class of interest (for example, "dog"). Except for the expected class (cat), which has a gradient equal to 1, all other classes have zero gradients. This information is used and backpropagated to the rectified convolutional feature maps of interest, which are then merged to generate the granular Grad-CAM location (blue heatmap), which shows where the model must look to reach the specified decision. Finally, the heatmap is multiplied pointwise using guided backpropagation to provide a guided Grad-CAM representation.

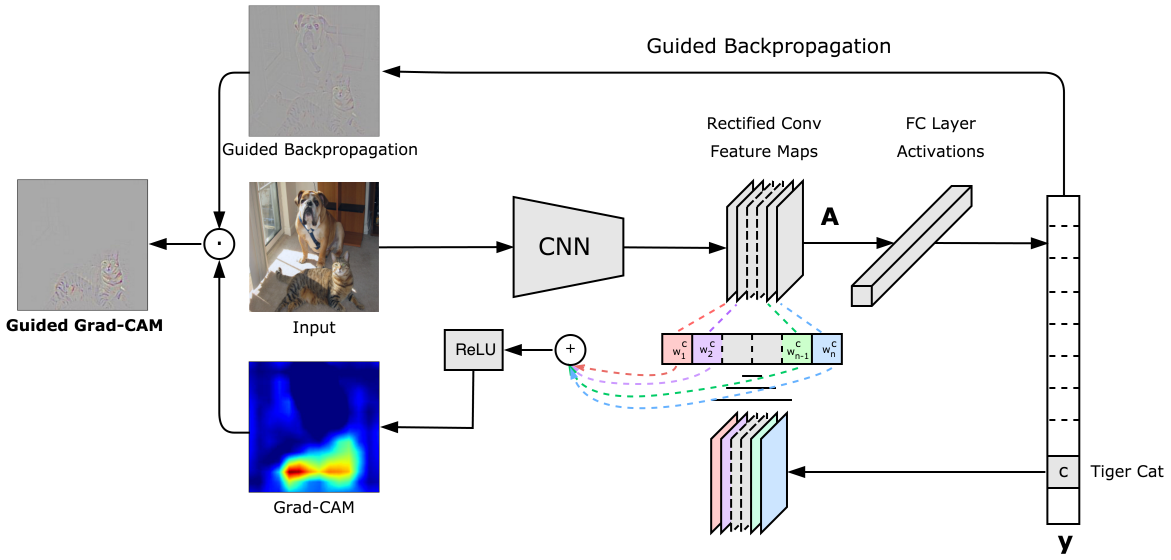

Fig. 2.9 Gradient- weighted Class Activation Mapping (Grad-CAM). From [Selvaraju et al., 2017]

2.4 Signal processing

2.4.1 Pan-Tomkins Algorithm

In Chapter 3, we focus on analyzing the R-to-R interval (RRI) between heartbeats; therefore, R waves must be extracted from the electrocardiogram (ECG) data. RRI analysis begins with the detection of QRS complexes to segment R waves. Pan and Tomkins [Pan et al., 1985] proposed an algorithm for QRS detection in real time that uses an adaptive threshold with an average error rate of approximately 1%. This threshold automatically adjusts its settings to fix periodic changes in the ECG signal. Additionally, it analyzes information on the slope, amplitude, and width of QRS complexes. The Pan-Tomkins Algorithm is summarized in Figure 2.10 and is described as follows:

1. For initial processing, a low-pass filter and a high-pass filter are applied to the ECG signal. The equation below can be used to determine the low-pass filter.

$$y(n) = 2y(n-1) - y(n-2) + x(n) - 2x(n-6) + x(n-12) \tag{2.35}$$

The high-pass filter is defined as follows:

$$y(n) = y(n-1) - \frac{1}{32}x(n) + x(n-16) - x(n-17) + \frac{1}{32}x(n-32) \tag{2.36}$$

2. Using the equation below, the signal is differentiated after filtering to learn more about the signal slope.

$$y(n) = \frac{1}{8}(2x(n) + x(n-1) - x(n-3) - 2x(n-4)) \tag{2.37}$$

3. Then, to emphasize the higher frequencies, the signal is squared point-wise to make them positive.

$$y(n) = x(n)^2 \tag{2.38}$$

4. After the data is squared, the method performs a moving window integration to obtain the waveform feature information.

$$y(n) = \frac{1}{N}(x(n-(N-1)) + x(n-(N-2)) + \cdots x(n)) \tag{2.39}$$

where n is the time step and N is the size of the moving window.

The temporal position of the QRS is then determined from the integrated waveform. Peak detection uses two adjusted thresholds. The highest of the two thresholds is used to compare the R peaks of the feature signal. The lower threshold is utilized for a search-back method, in which the algorithm must go back in time to find a lost peak if no peak is identified within a certain period. A peak is a local maximum that can be identified by looking for changes in signal direction within a predetermined period. When a new peak is located, it is classified as a noise peak or a signal peak depending on whether it exceeds the high threshold. It is regarded as a noise peak if it does not exceed. The thresholds are automatically modified to drift over the background noise according to the signal and noise peaks. Further research is done on the integration waveform and filter signals to determine separate threshold values at various processing points, strengthening their detection resistance. To be recognized as a QRS complex, a peak must appear as one in both integrated and filtered waveforms.

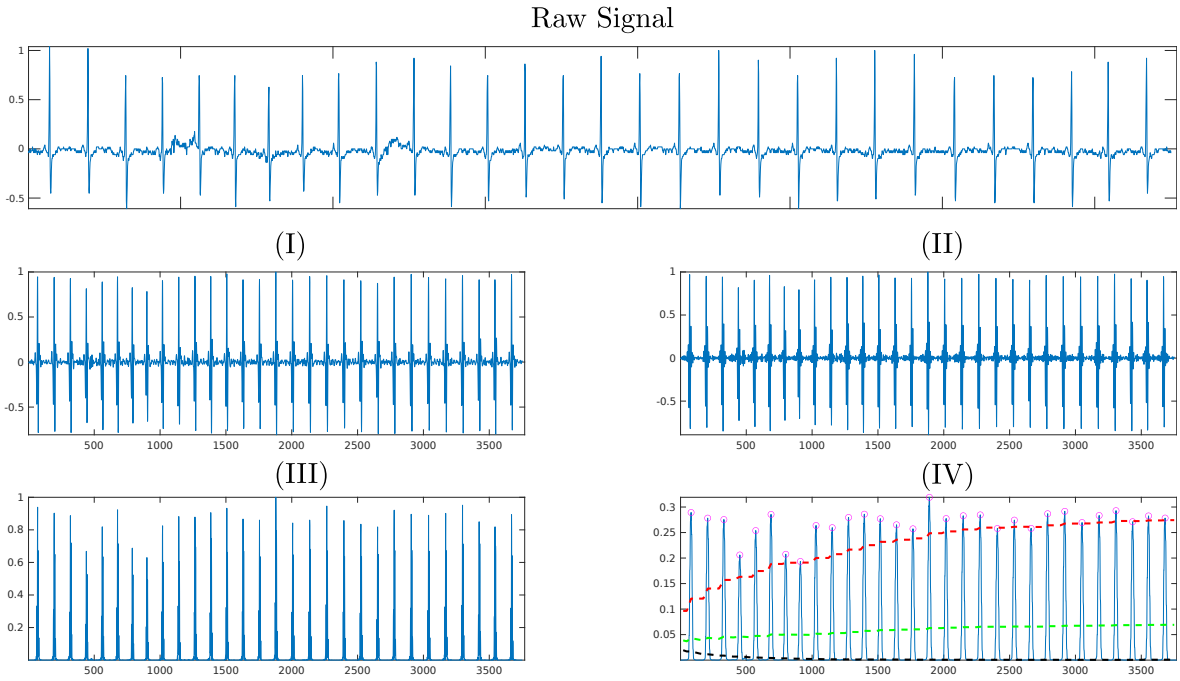

Fig. 2.10 Steps of the Pan and Tomkins Algorithm. (I) The band pass filter is applied. (II) The derivative filter is applied. (III) The signal is Squared. (IV) the moving window integration is applied (Black represents the noise level, Green represents the adaptive threshold, red represents the signal Level, and red circles represent the segmented QRS).

2.4.2 Recurrence plot

The periodicity of a signal can be quantified with the help of a recurrence plot, which is a two-dimensional representation of the recurrent states present in the trajectory of a dynamical system. Furthermore, this type of plot can also detect transitions in the system, such as when it moves from a periodic regime to a chaotic regime [Portes et al., 2019; Marwan et al., 2007].

For a set of time series data with N points, let the state vector at the data point $k = 1,\ldots,N$ be represented as $x(k) \in^n$, where n is the dimension of $x(k)$. The recurrence plot is a matrix $N \times N$ defined by the pairwise distance of all states along the trajectory of the system: $R_{ij} = \|x(i) - x(j)\|$, where $\|\cdot\|$ is the Euclidean norm [Eckmann et al., 1995]. For a given (i, j)-th cell in the recurrence plot, the darker the plot (i.e. smaller R_{ij}), the closer (recurrent) two states $x(i)$ and $x(j)$ are in the state space. Figure 2.11 is shown as an example; in this case, darker parallel diagonal lines indicate periodicity in the state trajectory of a system, since R_{ij} decreases with closer proximity of two states. Recurrence plots are usually defined with a particular threshold choice (e.g., $R_{ij} = 1$ if $\|x(i) - x(j)\| \leq \varepsilon$), which are not considered to prevent loss of information.

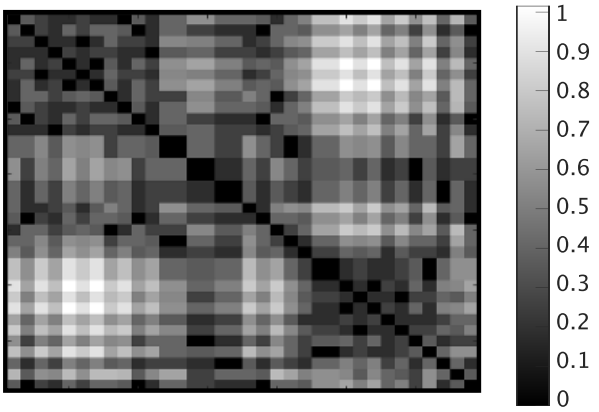

Fig. 2.11 Recurrence plot of the RRI of a 30s sample).

2.4.3 Short-time fourier transform

The discrete Fourier transform (DFT) is a linear transformation of a finite discrete signal x of length N, into N complex coefficients $c_k, k \in 0,1,\ldots,$ N 1 each describing a *sin* wave of frequency $\omega = 2\pi k/N$. The DFT is defined as.

$$c(k) = \frac{1}{N}\sum_{k=0}^{N-1} x(n)e^{-i2\pi kn/N} \tag{2.40}$$

The frequency of a nonstationary signal can be examined using the short-time Fourier transform (STFT) [Mitra et al., 2006], by using a moving window of length M over a signal and computing the DFT of the data. For all R samples, the window crosses the initial signal. To prevent spectral ringing, most window functions taper off towards the edges. The signal attenuation at the window boundaries is made up of overlapping the window segments if a nonzero overlap length L is given. A matrix that includes the magnitude and phase for each time and frequency point is added along with the DFT of each windowed segment. The number of columns in the STFT matrix is determined by the following.

$$k = [\frac{N_x - L}{M - L}] \tag{2.41}$$

where N_x represents the length of the original signal. For centered and two-sided transforms, the number of rows in the matrix is equal to N_{DFT}, the number of DFT points; for one-sided transforms, it is equal to $N_{DFT}/2+1$. $x(n)$ and the symbols represent the floor function.

The STFT matrix is given by $X(f) = [X_1(f)\,X_2(f)\ldots X_n(f)]$. According to which the mth component of this matrix is:

$$x_m(f) = \sum_{n=-\infty}^{\infty} x(n)g(n-mR)e^{-j2\pi f_n} \tag{2.42}$$

where $X_m(f)$ designates the DFT of the data centered around time mR. $g(n)$ designates the windows function of length M and R designates the overlap size between consecutive DFTs given by the difference between the window size M and the overlap size L. The squared magnitude of the STFT yields the spectrogram of the power spectral density of the function (see Fig. 2.12).

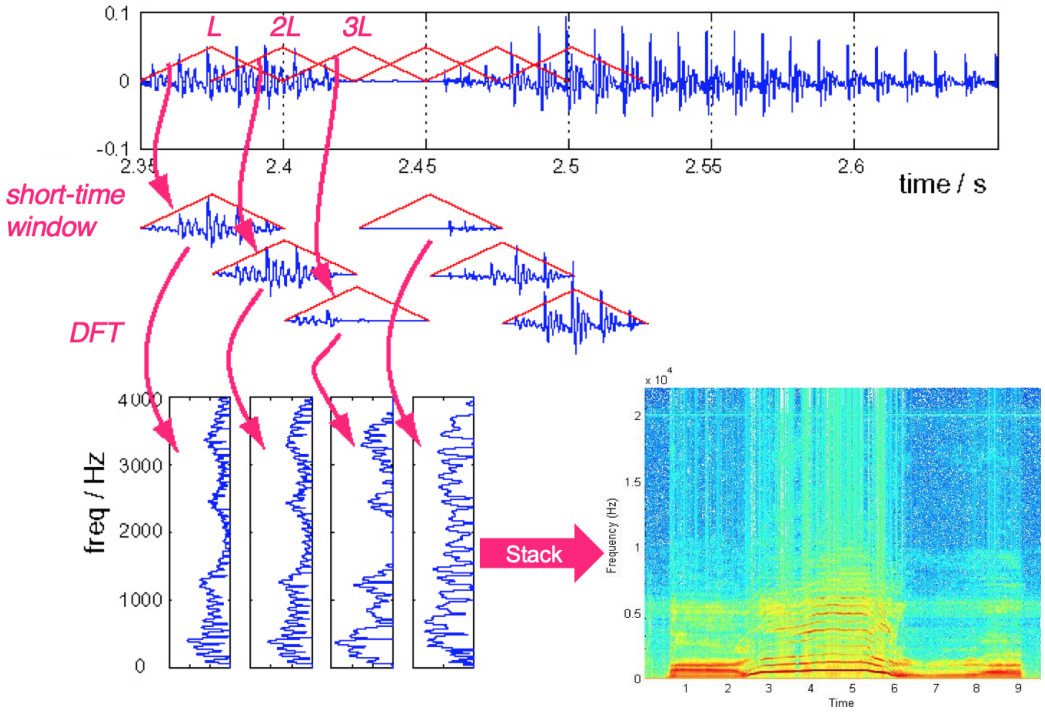

Fig. 2.12 Spectogram from a STFT. From [Bryan Pardo, 2020].

2.4.4 Cubic spline interpolation

A cubic spline, frequently used in numerical curve fitting, uses cross-validation between pairs of nearby samples to assess the degree of smoothing and estimates the absent data based on the value of the spline. This method fits a "smooth curve" to the known data. This method is used to mathematically fit a curve to the data consistently. Splines can be of any degree, although the cubic form is the most prevalent. The cubic spline technique guarantees that the interpolant is not only always differentiable on the interval, but also maintains a continuous second derivative because a general cubic polynomial comprises four constants. This is conceivable because there are four constants in a general cubic polynomial [Keys, 1981].

Given a function $f(x)$ defined in $[a,b]$, a nodes: $a = x_0 < x_1 < \ldots < x_n = b$ can be identified. A function must satisfy the following criteria to qualify as a cubic spline interpolant.

1. $S(x)$ denoted a cubic polynomial, represented by $S_i(x)$ in $[x_i, x_{(i+1)}]$, $0 \leq i \leq n$ 1.
2. $S_i(x_i) = f(x_i) = y_i$, and $S_i(x_{(i+1)}) = f(x_{(i+1)}), 0 \leq i \leq n$.
3. $S_{i+1}(x_i + 1) = S_i(x_i + 1)$, for each $i = 0, 1, 2, \ldots, n2$.
4. $S'_{i+1}(x_i + 1) = S'_i(x_i + 1)$, for each $i = 0, 1, 2, \ldots, n2$.
5. $S''_{i+1}(x_i + 1) = S''_i(x_i + 1)$, for each $i = 0, 1, 2, \ldots, n$ 2.
6. And satisfying one of the boundary conditions:
   - $S''(x_0) = S''(x_n) = 0$ (Natural condition).
   - $S''(x_0) = f'(x_0)$ and $S'(x_n) = f'(x_n)$ (Clamped condition).

2.5 Evaluation procedures

Performance evaluation is a crucial component of the training, validation, and testing phases of a machine learning (ML) method. This establishes the level of output quality and the impact of the input data. Unfortunately, imbalanced data sets are often used in real-world applications for classification tasks. These classes, known as the minority and majority classes, indicate that some classes have fewer samples than others. This asymmetry poses a substantial challenge when using machine learning (ML) to solve classification problems. When conventional ML is used, it can bias the predictions in favor of the majority class [DeBrusk, 2018]. Therefore, the generalization may not be correct. Depending on the objective and characteristics of the dataset, it is possible to use a variety of measures.

2.5.1 Data Split

As part of the model evaluation procedure, the data can be divided into test and training sets. Two new groups are formed as a result of this process. Employing training and test sets when building and validating a model is essential. The model evaluation procedure allows the model to be trained and evaluated on different data samples. It also does a better job of extrapolating unknown information than other methods. One problem in evaluating the model with a simple split is introducing a significant degree of variance.

2.5.2 Cross-validation

It is common practice to divide the data set into training and testing parts before evaluating classification tasks. After training on the first, the machine learning (ML) algorithm is applied to the test data set to provide performance metrics to evaluate the algorithm's performance. For ML methods, access to insufficient test and training data is a frequent problem. As a result, overfitting may occur when studying these methods. A common solution for this is to use X-Fold cross-validation. In cross-validation, a data collection is divided into X portions, each of which is used as a test data set, while the others are combined to serve as training data. The performance metrics for all validation processes are then averaged. When comparing ML algorithms, no single indicator can be used for all cases because they all have strengths and weaknesses.

The cross-validation strategy begins by generating several train-test splits (see Fig. 2.13). Next, it determines the accuracy of the test set for each split and then averages the results of these calculations [Refaeilzadeh et al., 2009]. The following is a list of the stages involved in the cross-validation technique.

1. Separate the initial data set into N sections, each of the same size.
2. A test set should be created using the first part and a training set should be created by combining the other parts.
3. Determine the accuracy of the test set.
4. Iterate the previous two stages N times, each time selecting a new component to serve as a training set.
5. Determine the ultimate accuracy by calculating the average accuracy of each individual.

Although cross-validation is N times slower than the train-test split, it has higher out-of-distribution accuracy. N is assumed to have a value of 10 in most cases. Cross-validation strategies can be improved by repeating the same steps more frequently. For example, perform 1,000 iterations of 10-fold cross-validation.

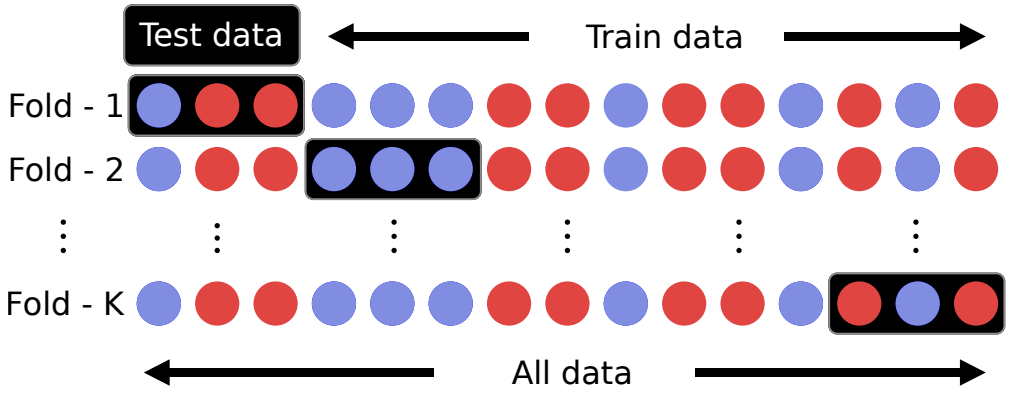

Fig. 2.13 k-fold Cross-validation.

2.5.3 Confusion Matrix

A confusion matrix is the most fundamental way to evaluate the accuracy with which a classification task has been completed. Compared to true classes, it tells us how many predicted classes were outputted correctly and how many were outputted incorrectly. The form of the confusion matrix is a grid $A \times A$, with A representing each of the classes. For example, for binary classification, the confusion matrix takes the following form.

		True		
		Positive	Negative	
Predicted	Positive	TP	FP	↔ PPV
	Negative	FN	TN	↔ NPV
		↕	↕	
		Sen	Spc	

where the positive and negative values represent:

- TP, or True Positive, indicates that a positive target has been accurately predicted.
- TN, or True Negative, indicates that a negative target has been accurately predicted.
- FP stands for "false positive," which means that a negative target was incorrectly predicted.
- FN, which means "false negative," means that a positive target was incorrectly predicted.

Several other classification metrics, including accuracy, sensitivity, specificity, precision, and the F1 score, can be derived from the confusion matrix.

Accuracy

Calculating the percentage of correct predictions, often known as accuracy, is one of the most common metrics used to evaluate a model. Classification accuracy considers the number of correct predictions from all the predictions. This type of measure is the easiest to understand and calculate.

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total prediction}} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.43}$$

However, when working with imbalanced datasets, a standard accuracy would be biased towards the majority class.

Sensitivity

Also known as recall, represents the percentage of true positive cases that are accurately predicted.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{2.44}$$

Specificity

Represents the percentage of true negative cases that are accurately predicted.

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{2.45}$$

Precision

Also called the Positive Predictive Value (PPV) represents the percentage of correctly detected positive cases out of all real positive cases.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.46}$$

f1 score

The f1 score is the harmonic mean of precision and sensitivity and provides a more accurate measurement of the number of cases incorrectly classified.

$$\text{f1 score} = 2\frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{2.47}$$

2.5.4 Receiver operating characteristic curve

A visual tool for rating classification models is the receiver operating characteristic (ROC). The ROC graph illustrates the sensitivity along the Y axis and the false positive rate (1-Specificity) along the X axis. A random model will result in the ROC curve having a diagonal line. The model performs more accurately the higher the line grows towards the upper left corner. Figure 2.14 shows an ROC plot as an example. The Area Under ROC Curve (AUROC) is also used to evaluate classification models. The range of possible values for AUROC extends from 0.5 to 1, with 1 indicating the ideal prediction capacity and 0.5 indicating that there is no predictive capacity. Although a result between 0.700 and 0.800 is acceptable, a value greater than 0.800 is considered outstanding.

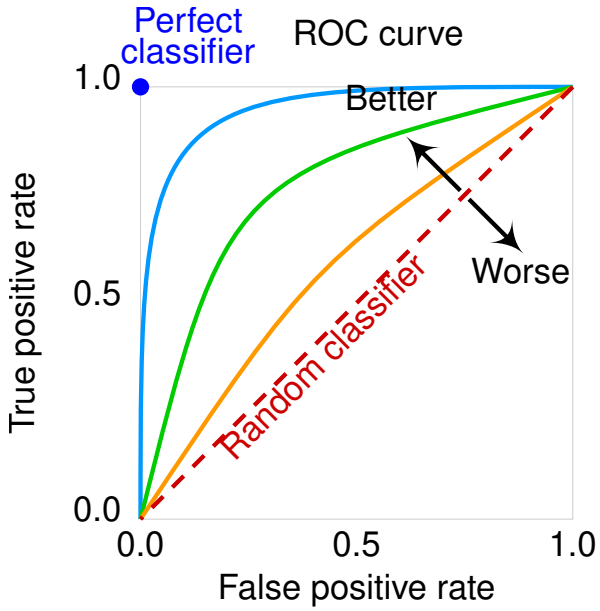

Fig. 2.14 Receiver Operating Characteristic curve interpretation. From [Martin Thoma, 2018]

2.5.5 Precision-Recall curve

When classes are uneven, the precision-recall curve is a useful indicator of the success of a prediction. Precision in information retrieval refers to how relevant the results are, whereas recall refers to how many true results are returned. The sensitivity is illustrated along the X axis and the precision along the Y axis of the Precision-Recall curve. The precision-recall curve represents the trade-off between precision and recalls at various thresholds. Good recall and precision are represented by a large area under the curve. A large recall is associated with a lower false negative rate, whereas a large precision is associated with a lower false positive rate. The high area under the precision-recall curve (AUPRC) indicates that the model produces both accurate (high precision) and most positive (high recall) results.

Chapter 3

Early warning of Atrial Fibrillation

This chapter begins with an overview of the cardiovascular system, starting with heart physiology. It expands multiple concepts to understand heart arrhythmias, focusing on atrial fibrillation (AF), which is the most prevalent arrhythmia worldwide. From the anatomy and function of the heart to the pathophysiology and treatment of AF, we end the section with a review of the literature on the detection and prediction of AF. Finally, we present our contribution: an innovative method for the detection of early warning of AF (adapted from article[1]).

3.1 Introduction & literature review

3.1.1 Biological background

The human heart is an organ responsible for the circulation of blood throughout the body, providing oxygen and nutrients to all cells, and removing waste products resulting from typical cell activities. There are four distinct chambers within the heart. The right atrium and right ventricle are located on the right side of the heart, whereas the left atrium and left ventricle are located on the left side. A continuous partition divides the left side of the heart from the right side, with an interatrial septum that separates the left and right atriums and an interventricular septum that separates the left ventricle and the right ventricle (see Fig. 3.1). The atria and ventricles are separated from each other by the atrioventricular septa. Subsequently, blood is transferred from the atria to the ventricles through the two atrioventricular orifices, openings in the atrioventricular septa. Two atrioventricular valves, which open and close at regular intervals together with each heartbeat, are responsible for controlling the

[1] Marino E. Gavidia, Hongling Zhu, Arthur N. Montanari, Jesus Fuentes, Cheng Cheng, Sergio Dubner, Martin Chames, Yinuo Jiang, Shengjun Zhang, Hai-Tao Zhang, Xin He, Basi Teng, Guohua Wan, Ye Yuan, Xiaoyun Yang, and Jorge Goncalves. *Early Warning of Atrial Fibrillation*. In review, 2022

openings. The tricuspid valve is the valve that connects the right atrium to the right ventricle, while the mitral valve is the valve that connects the left atrium to the left ventricle. The pulmonary arteries (PA) and pulmonary veins connect the right and left atriums of the heart to the pulmonary circulation, respectively. The left ventricle and the RA are connected to the rest of the body through the aorta and the cavas of the superior and inferior vena, respectively. The pulmonary artery and the aorta have valves that prevent blood from flowing back into the right ventricle and left ventricle [Buckberg et al., 2018].

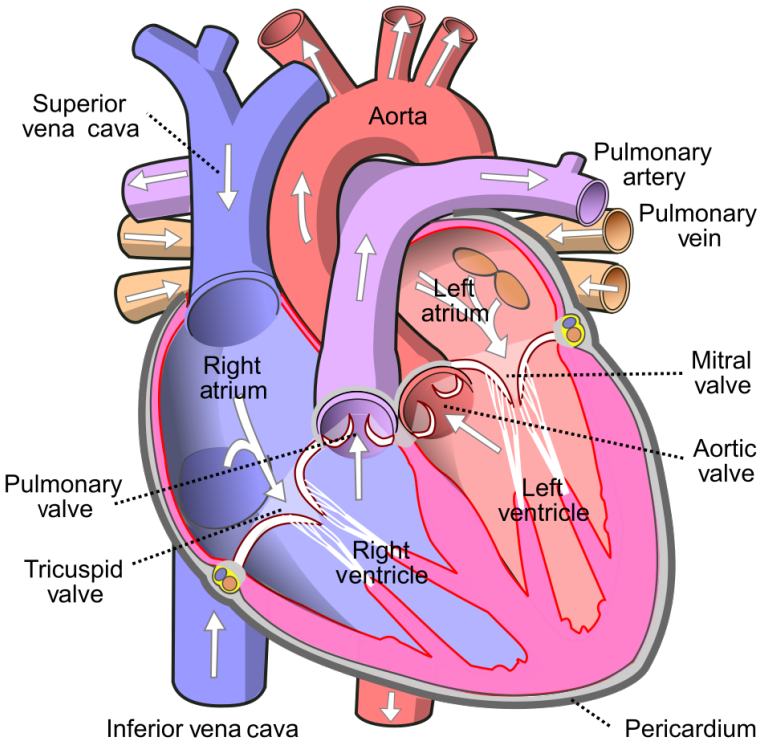

Fig. 3.1 The heart anatomy. From [Betts et al., 2013].

The electrical system of the heart is responsible for controlling both the rate and rhythm of the heartbeat. With each beat, an electrical signal is sent from the top to the bottom of the heart, causing the heart to contract. During heart activity, the four chambers are contracted to a specific rhythm, which coordinates the effectiveness of the input and output of blood flow in a single cycle [Steinberg et al., 2015; Tung et al., 2016]. These contractions can be broken down into two primary phases: diastole and systole. The first phase of the heart cycle is called atrial diastole. During this phase, the atrial chambers relax, allowing oxygenated blood to reach the left atria through the pulmonary veins and deoxygenated blood to enter the right atria through the vena cava. The atria will then enter systole when the atrial

chambers contract to force blood through the atrioventricular valves and ventricles. Ventricles must be in a diastolic state during the systolic and diastolic activities of the atria so that blood can pool in the atria. Subsequently, the ventricles contract, causing blood to flow from the left ventricle to the rest of the body through the aorta, the right ventricle, and the lungs through the pulmonary arteries. At rest, the heart of a healthy adult beats anywhere from 60 to 100 times per minute, which is called the heart rate (HR) and is measured in beats per minute (bpm) [Lantelme et al., 2002].

3.1.2 Electrocardiogram

An electrocardiogram, more commonly called an ECG, is a graph of voltage versus time produced by a method that involves placing electrodes on the body of a patient to record the electrical activity of the heart for a certain time. This method is also known as an electrocardiograph. These electrodes are capable of detecting electrical changes in the skin that occur as a result of depolarization of the heart muscle during each heartbeat. At each point in the cardiac cycle, ECG can be used to determine the total amount of electrical depolarization of the heart and the direction of this depolarization. In an electrocardiogram (ECG) [Blackburn et al., 1960], we may observe electrical changes that control heart contractions. In a typical ECG consisting of 12 leads, the patient's limbs and chest each have ten electrodes attached. These electrodes are attached to the chest skin at the appropriate locations (see Fig. 3.2). Six views are obtained from the leads' placement by Einthoven's triangle [Crawford et al., 2012]. Standard leads I, II, and III are the initial three views. The placement of the two sensors on the skin's surface (positive and negative) completes the circuit; they are also known as "bipolar leads."

In an ECG, one complete cycle corresponds to one heartbeat. The components of the electrocardiogram cycle present during a complete cycle are the P wave, the QRS complex, and the T wave, respectively (see Fig. 3.3). The P-wave can be found at the beginning of each cycle, which indicates that the atria have depolarized. The QRS complex follows immediately, which appears as three waves closely related to each other and signifies depolarization of the ventricles. This complex can be recognized by its appearance (Q, R, and S waves). The T-wave represents the repolarization of the ventricles of the heart. The PR interval is the time elapsed between a pair of P waves and a pair of R waves.

Similarly, the length of time that elapses between the S wave and the T wave, also known as the ST interval, shows the amount of time that elapses between ventricular depolarization and repolarization. This can be seen on an electrocardiogram (ECG). The R peak-to-R peak interval, or the RR interval, is the time that elapses between the peaks of the R wave in two consecutive ECG cycles. It is also known

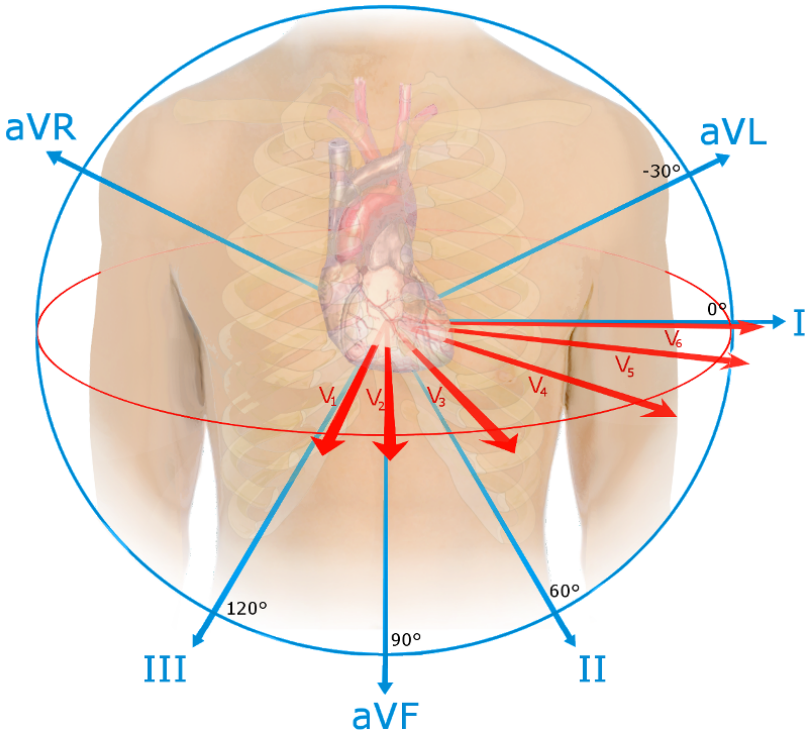

Fig. 3.2 Spatial orientation of ECG leads. From [Npatchett, 2015] .

as the R-to-R interval. The duration of this time is expressed in milliseconds. Measurement of the RR interval, or the length of time that elapses between each pair of heartbeats, is one way to determine an individual's heart rate.

An ECG can identify areas of the heart muscle that lack oxygen and areas of tissue that have already passed away. Examination of ECG data, particularly the waves and interval patterns produced, is critical to assessing whether the heart performs its functions normally. Furthermore, the interpretation of an electrocardiogram has a wide variety of applications in the real world, such as the detection of heart disease [Aljanabi et al., 2018], diagnosis of sleep apnea [Ramachandran et al., 2021], and the biometric identification [Gutta et al., 2015].

Electrocardiograms can be measured using a wide range of currently accessible approaches. These are In-the-person, on-the-person, and off-the-person [Silva et al., 2015]. In-the-person measurements are taken directly from the subject being measured. Measurement devices are introduced into the subject's body in one of two ways: surgical implantation or the intake of a pill-shaped system. It is common practice to save these strategies for chronic patients and difficult clinical situations. On-the-person. This group is the birthplace of most measurement methodologies. In procedures of this kind, the signal is obtained by deploying a device directly attached to the body's surface. Standard examples include bedside monitors and Holter machines. Both types of equipment require a total of 12 leads

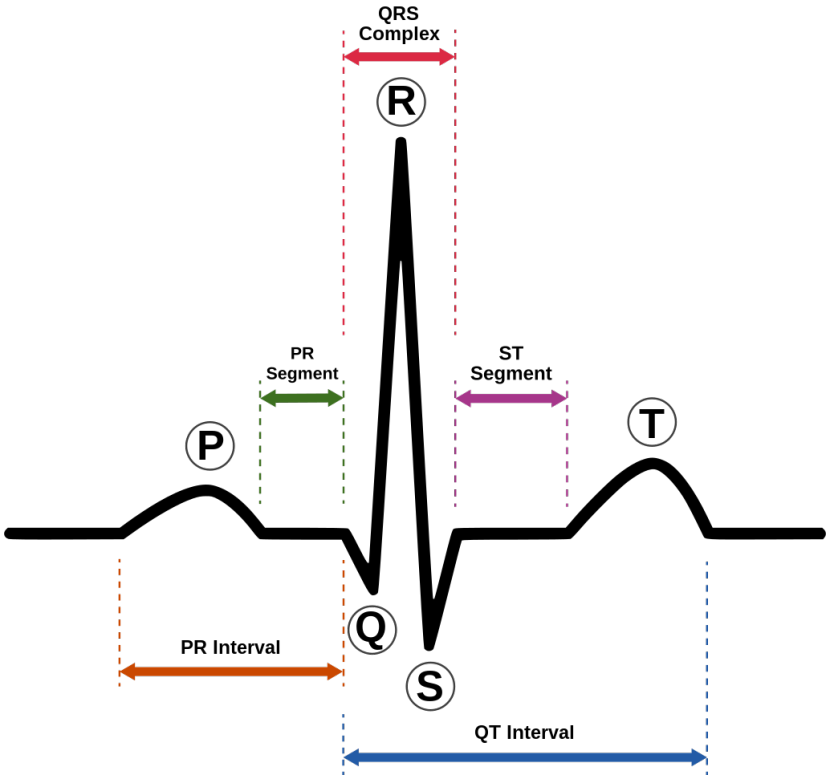

Fig. 3.3 Representation of the QRS complex. From [Anthony Atkielski, 2007] .

and the connection of electrodes to the patient's arms, legs, and chest to obtain an accurate reading. Off-the-person is obtained using a technique that uses sensors that are included in everyday items, such as smartwatches and smartphones. In recent years, there has been a tendency to increase the number of applications that use such approaches.

3.1.3 Type of Arrhythmias

Two terms used in electrophysiology are rate and rhythm. The term rate refers to the speed at which the heart beats and the term rhythm refers to the pattern of variation in the way the chambers of the heart contract. A normal sinus rhythm, often known as (SR), is used to describe a healthy rhythm, while cardiac arrhythmias result from an irregular beat. ECG measurements demonstrate that the cardiac rhythms adhere to a particular pattern. This pattern can be recognized as the subwaves P, QRS, and T, and each of these subwaves has a temporal length in humans that varies very little relative to each other. The first step in classifying arrhythmias is determining where the abnormal rhythm occurs in the body. The second step is to identify the mechanism responsible for the rhythm disturbance. When someone has an arrhythmic condition, their heartbeat can be abnormally fast (also known as tachycardia) or abnormally slow (also known as bradycardia) compared to the usual state. Fibrillations are abnormal

heart rhythms that can occur when heart contractions are completely disorganized, causing the heart to flutter rather than having a slow or fast HR.

Therefore, arrhythmias are classified according to the rate of heart rhythm, the process that produces them, and the length of each beat. For example, bradycardia, also known as bradyarrhythmia, is a condition in which the heart beats slower than 60 beats per minute. On the other hand, tachycardia refers to rapid heartbeats that occur at a speed of more than 100 beats per minute. Another way to classify arrhythmias is by the location of the irregular rhythmic beating that defines the disorder in the heart. Supraventricular arrhythmias (SVT) begin in the atria or upper chambers of the heart. Arrhythmias that begin in the ventricles, the lower chambers of the heart, on the other hand, are referred to as ventricular [Desai et al., 2022].

After the aberrant component of the electrocardiogram has been identified based on the subwaves that show abnormal patterns, the arrhythmia can be named as follows. Atrial arrhythmia: It can be identified by a rate greater than 140 beats per minute and by the presence of normal QRS complexes and irregularly shaped P waves, provided that the P waves are visible and not obscured by the T wave that preceded them. Ventricular arrhythmia: is caused by disturbances in the normal conduction pathways of the ventricles as a result of disorganized electrical activity in those chambers. These are the most common types of arrhythmia [Crawford et al., 2012; Kocheril et al., 2009; Stroobandt et al., 2016]:

- Atrial Flutter: It originates in the atrium, typically in the lower atrium close to the AV node. The flutter rate could be between 250 and 350 per minute. Another distinction between atrial fibrillation and flutter is that individual heartbeats may not show all P waves.

- Atrial fibrillation: is a type of arrhythmia caused by frequent electrical impulse leakage in the atrium. The atrium rate is extremely fast (350-600 bpm). Therefore, the patient's pulse is erratic.

- Premature Atrial Contractions (PAC): These are atrial discharges that cause the atrium to constrict, but do not cause ventricle contractions. It can be identified by P waves that have an uneven shape.

- Premature ventricular contractions (PVC): These are extra beats that come from the walls of the ventricles and cause them to contract erratically and out of proportion to each other. Furthermore, if the ventricular muscles have suffered long-term structural damage, the highly frequent occurrence of these contractions over an extended period could develop cardiomyopathy. This might be the case if the heart has been under sustained stress.

- Ventricular tachycardia (V-Tach): when there are three or more PVC visible in a row at a pace of 100 beats per minute or more. Because there is not enough time for the ventricles to fill, heart output is significantly reduced. This irregular heartbeat could also cause a heart attack.
- Ventricular fibrillation (V-Fib): often known as cardiac arrest, is a risky arrhythmia that can result in sudden death. After lowering blood pressure, cardiac output and blood pressure quickly reach zero. Rapid electrical discharges in the myocardium result in cardiac muscle fibrillation. One of the most common causes of sudden cardiac arrest is the V-Fib rhythm. Almost always, the presence of significant heart disease is required. Ventricular fibrillation can be caused by coronary artery disease, acute myocardial infarction, myocardial ischemia, and third-degree AV block with a slow ventricular response.

3.1.4 Atrial Fibrillation

During atrial fibrillation, irregular electrical patterns disturb the regular electrical impulses of the sinoatrial node. This leads to chaotic and disorganized electrical activities, causing the atriums to become uncoordinated, quiver, and fibrillate. Electrical stimulation of the sinoatrial node controls the cardiac cycle in a normal and healthy heart. After passing through the atrioventricular node, the electrical pulse reaches the ventricle of the heart. The atrioventricular node slows the electrical impulses just before ventricular contraction, so there is enough time for the atria to empty during each cycle. Heartbeats that are in normal sinus rhythm (SR) are the consequence of electrical activations that travel through the atrium and induce coordinated contractions [Glomset et al., 1952].

Patients diagnosed with atrial fibrillation can have resting heart rates ranging from 100 to 200 beats per minute, compared to typical resting heart rates during SR ranging from 60 to 100 beats per minute. All these erroneous signals flood the AV node with electric impulses that cause the ventricles to start beating quickly (see Fig. 3.4). Despite this, the ventricles cannot beat as quickly as the atria because the AV node cannot send the signals to the ventricles at the same rate they arrive. Due to this, the atria and ventricles of the heart will no longer beat in a coordinated manner [Davies et al., 1972], and the heart will not be able to adequately eject blood into the ventricles, which ultimately ventricular output to the rest of the body.

In most cases, atrial fibrillation begins as brief, intermittent episodes and then gradually develops into more frequent episodes that last longer. AF can be diagnosed as: [Allessie et al., 2001].

1. Paroxysmal AF: This occurs when an atrial fibrillation event lasts no more than 7 days and resolves on its own or with cardioversion.

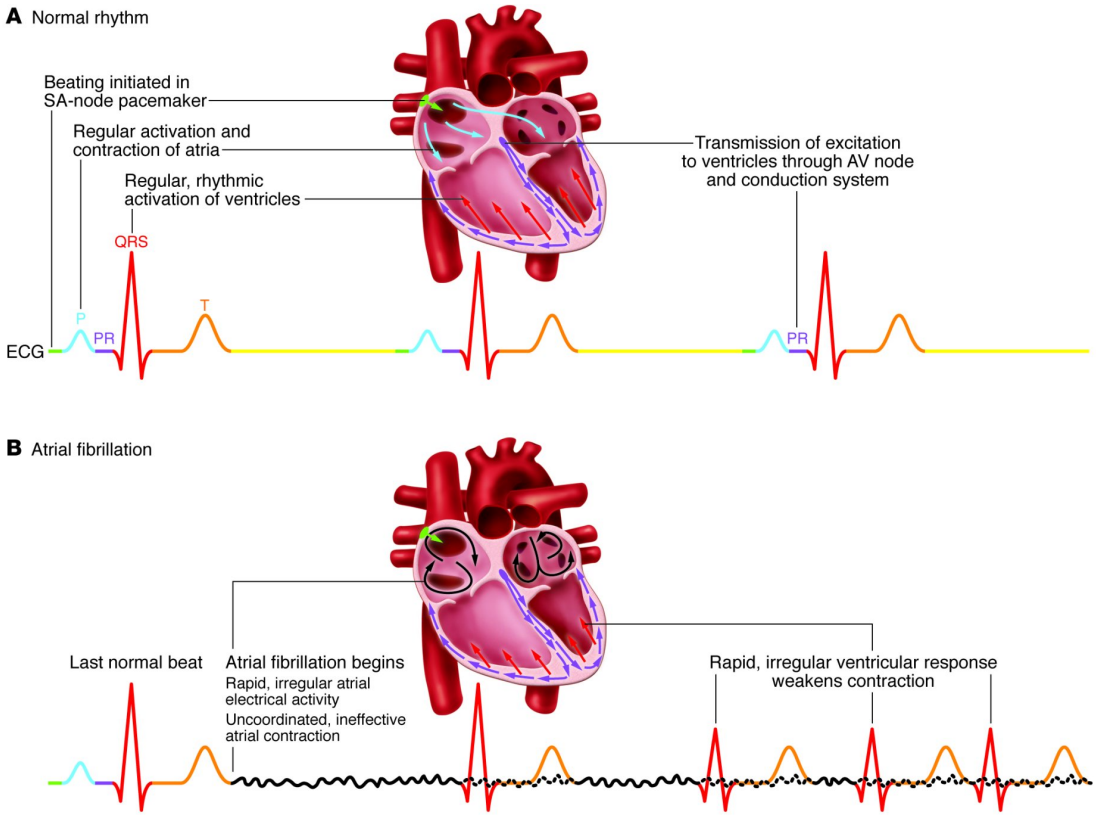

Fig. 3.4 Heart rhythm. Normal rhythm (top). Atrial fibrillation (bottom). From [Wakili et al., 2011].

2. Persistent AF: This occurs when an atrial fibrillation event lasts more than 7 days but resolves naturally or with cardioversion within a year.

3. Long-standing persistent AF: This occurs when a person has experienced atrial fibrillation for more than a year, but the restoration of sinus rhythm is still the main objective of the doctor and the patient.

4. Permanent AF: This occurs when the doctor and the patient decide that the long-lasting rhythm of AF is permanent.

Atrial fibrillation (AF) can also be called asymptomatic AF when there are no symptoms, or symptomatic atrial fibrillation when there are symptoms such as palpitations, shortness of breath, chest tightness, lethargy, weakness, dizziness, or fainting. Symptoms can range from mild to severe, making it hard to do everyday activities. [Crawford et al., 2012].

The concern with atrial fibrillation is that it can develop into an illness that does not have symptoms and can go unnoticed for years if it does not cause complications and it is often discovered by accident during routine medical checks [Friberg et al., 2014], and a stroke may be the first sign of this

condition [Barbarossa et al., 2014]. According to research [Steinberg et al., 2015; Kerr et al., 2005], a significant number of patients experience a transition from paroxysmal to chronic AF in a single year. This makes treatments increasingly ineffective as the severity of the disease increases.

3.1.5 Pathophysiology of atrial fibrillation

The pathophysiological mechanisms responsible for atrial remodeling can result from heart disease or situations that encourage the development of atrial fibrillation (AF). When atrial fibrillation develops, it creates irregularities that promote AF, further increasing the heart's susceptibility to the induction and maintenance of AF. Remodeling causes electrical and structural changes that, in turn, create conduction disturbances. These disturbances lead to an increased propensity for triggers of AF, such as re-entry-prone substrates [Wijffels et al., 1995]. It is common practice to refer to this auto-reinforcing and self-perpetuating aspect of AF using the term "AF begets AF" [Chaldoupi et al., 2009]. The structural and electrical remodeling of the atrium is what medical professionals call the process by which AF worsens with time [Nattel et al., 2014].

Figure 3.5 shows the conceptual framework for the start, maintenance, and progression of AF [Heijman et al., 2014]. AF onset may appear relatively early in patients with a large genetic predisposition. Remodeling caused by AF helps keep the arrhythmia going and helps AF get worse (Fig. 3.5A). The genetic background alone is not enough to make most people susceptible to AF. More changes caused by the disease can make the heart more vulnerable and make it easier for paroxysmal episodes to begin AF. Some people with paroxysmal AF can develop forms of AF that last longer over time (Fig. 3.5B). Because substrate and trigger are combined, some patients have the first episode of AF that lasts more than 7 days. If the underlying disease worsens or the doctor leaves the patient with AF, AF can become permanent (Fig. 3.5C).

Structural remodeling

The term "structural remodeling of the atria" refers to the changes that occur, adaptive or maladaptive, in the cellular architecture of the heart. The structure of the atrium can be remodeled at both the organ and tissue levels, and the process depends on time and etiology. Studies have shown that the dimension of the atrium is a significant predictor of the persistence of AF-maintaining reentry under specific functional settings [Ausma et al., 1997]. This is the main hallmark of organ-level remodeling, characterized by enlargement of the atria [Zou et al., 2005]. Age and genetics are two examples of these types of stressors. Concomitant conditions, such as heart failure, hypertension, diabetes, and

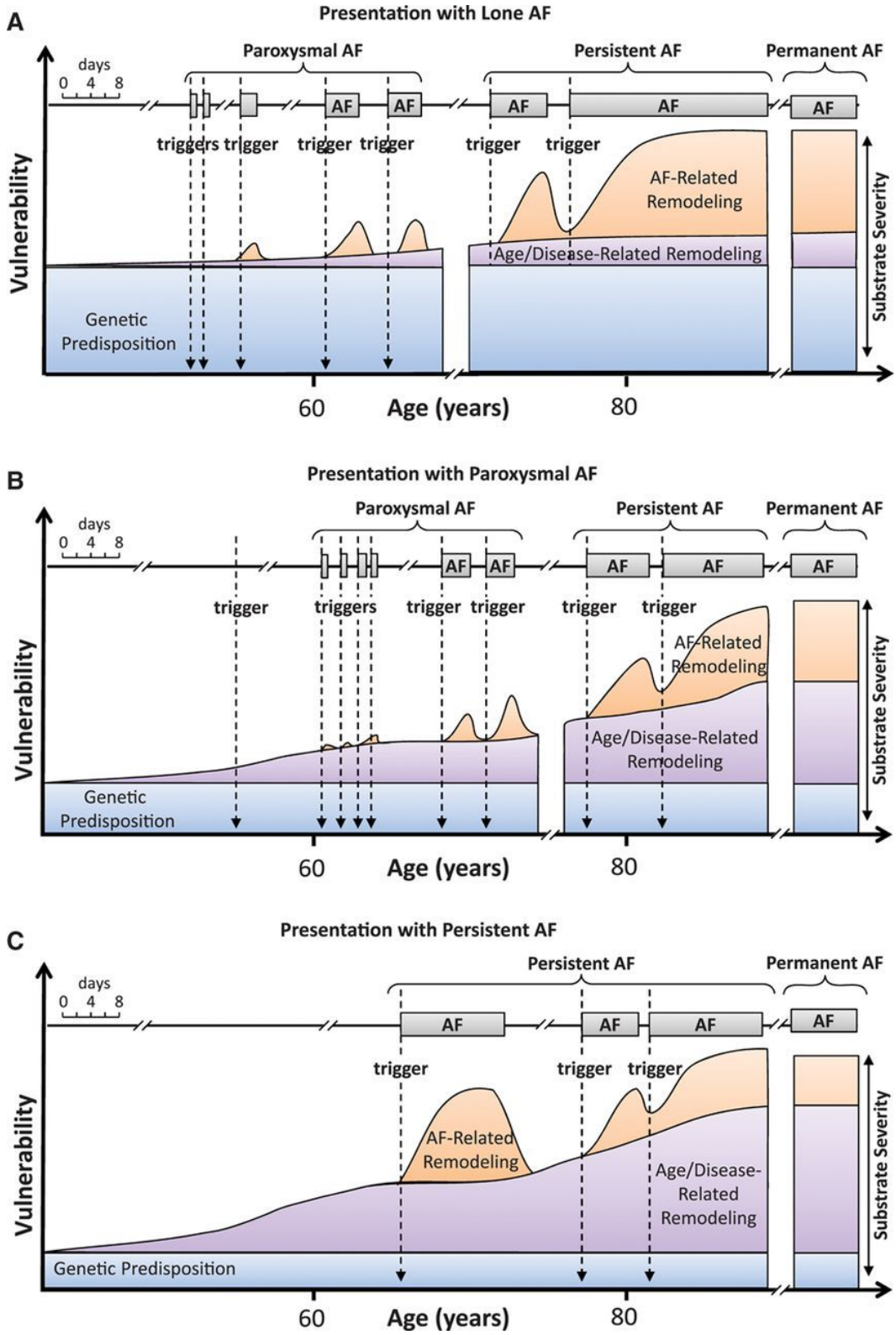

Fig. 3.5 The maintenance and progression of AF as a function of A) AF genetic predisposition. B) Paroxysmal AF. C) Persistent AF. From [Heijman et al., 2014].

obesity, can also be a factor. Cellular hypertrophy, myolysis, dedifferentiation, fibrosis, apoptosis, and mitochondrial and sarcoplasmic reticulum dysfunction are some significant alterations at the tissue level. In particular, many consider fibrosis the defining characteristic of atrial fibrillation [Burstein et al., 2008]. This is because it appears to be a common endpoint for a wide range of conditions that promote AF, and studies have shown that it may be able to predict recurrences potentially [Oakes et al., 2009].

Due to the inability of the fibrotic tissue to expand or contract, myocardial compliance decreases, hindering cardiac contraction [Baudino et al., 2006]. This is due to the increased stiffness of the atrial wall that occurs due to the development of fibrosis. According to several studies, fibrotic regions experience a decrease in oxygen supply, which can cause the death of cellular tissue [Sabbah et al., 1995]. Fibrosis makes atrial fibrillation more likely by disrupting the continuity of the fiber bundles, decreasing the number of intercellular connections, and generating local conduction disruptions [Burstein et al., 2008]. Because the creation of fibrosis is originally a repair mechanism in response to tissue damage or cardiomyocyte death, the loss of more cardiomyocytes in the fibrotic regions can further perpetuate the formation of fibrosis [Camelliti et al., 2005].

The development of fibrosis is distinguished by an abnormally high production of extracellular matrix, which is responsible for the pathological growth of connective tissue [Zou et al., 2005]. Several secreted factors are pro-fibrotic [Kupfahl et al., 2000], although the processes that regulate extracellular matrix remodeling are not yet fully understood. Among these are angiotensin II, transforming growth factor-1, platelet-derived growth factor, and connective tissue growth factor, all of which have profibrotic effects on their own and when combined [Chaldoupi et al., 2009]. Both types of fibrosis that have been found, reparative fibrosis, which replaces dead cardiomyocytes, and reactive interstitial fibrosis, which splits muscle bundles, are known to alter the electrophysiology of the atria [Weber et al., 1992]. In postmortem studies that included atrial biopsies taken from various locations in the right and left atria, the degree of fibrosis and adipose tissue in the atrial indicated a significant correlation with the development of AF in all tissue samples (see Fig. 3.6).

Angiotensin II is caused by the type I receptor, which is responsible for increasing connective tissue formation (also known as fibroblast proliferation), cardiomyocyte hypertrophy, and programmed cell death (also known as apoptosis) [Burstein et al., 2008]. Phospholipase C is also more active when angiotensin type I is activated. An enzyme called phospholipase C is responsible for breaking a signaling phospholipid protein known as PIP2 into diacylglycerol and IP3, an inositol phosphate signaling molecule. Protein kinase C is an enzyme responsible for controlling cardiac contractility,

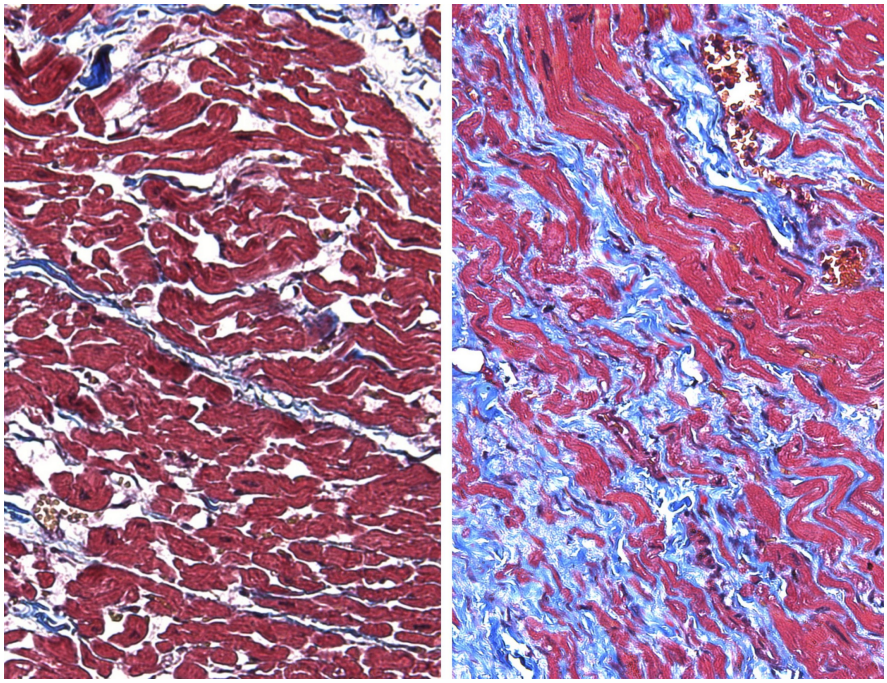

Fig. 3.6 Light microscopy from patients without AF (left) and with AF (right). From [Schotten et al., 2016].

pathological growth responses, and pump function [Newton et al., 2016]. When diacylglycerol is present, it activates protein kinase C. IP3 causes intracellular Ca2+ release. Diacylglycerol and IP3 have been shown to trigger the remodeling process [Garcia et al., 2017]. However, activation of angiotensin type II inhibits mitogen-activated protein kinases. Generate antiproliferative effects, both negative changes caused by angiotensin type I [Hunyady et al., 2006]. The equilibrium or lack thereof of the two receptors acting as counterregulatory agents may have significant consequences. Transforming growth factor 1, which functions as the main mediator of angiotensin II, is secreted by cardiomyocytes and fibroblasts. This affects cells that produce angiotensin II (autocrine) and cells close to it (paracrine). Angiotensin II also reciprocally increases transforming growth factor-1, increasing the levels of components that promote fibrosis and creating positive feedback [Rosenkranz, 2004]. These factors promote selective atrial fibrosis and atrial conduction heterogeneity [Attisano et al., 2002]. They also increase fibroblast activation and collagen deposition [Verheule et al., 2004].

The production of more fibroblasts and their differentiation are both encouraged by platelet-derived growth factors. The differentiation process of fibroblasts occurs when fibroblasts change their phenotype to become myofibroblasts, which are in a more proliferative condition [Darby et al., 2007]. Platelet-derived growth factors also go through auto-phosphorylation, a process in which its receptors are self-activated by activating the tyrosine kinase enzyme. This allows platelet-derived growth factors to initiate signaling through mitogen-activated protein kinases and phospholipase C pathways in a

manner analogous to transforming growth factor-1 and angiotensin II. Platelet-derived growth factors appear to be the underlying factor for fibroblast hyperresponsiveness exclusively in the atria [Burstein et al., 2008]. The connective tissue growth factor is an important downstream effector of transforming growth factor 1 and is coordinated with the change of growth factor 1 in places with active fibrosis promotion [Chuva De Sousa Lopes et al., 2004]. The connective tissue growth factor also contributes similarly to remodeling. Fibroblasts are activated directly by connective tissue growth factor, which is additionally elevated by angiotensin II and transforming growth factor-1. This makes the atria more sensitive to fibrotic remodeling than the ventricles.

Electrical remodeling

Electrical remodeling is characterized by a long-lasting change in the electrophysiological characteristics of the atria as a direct result of a modification of the order in which electrical activation occurs. Any or all ion channels, pumps, and exchanges on the cardiac cell membrane can be remodeled, affecting the electrophysiological properties of the atrial chambers. Alterations in ionic currents and aberrant distribution of the connexin hemichannels of the gap junction that electrically connect cardiomyocytes are the primary elements that contribute to electrical remodeling. In cardiomyocytes, sodium channels (Na +), potassium (K +) and calcium (Ca2 +) regulate the inflow and outflow of sodium ions (Na +), potassium (K +) and calcium (Ca2 +) to generate action potentials, which in turn drive cell contraction. Electrical coupling of action potentials between neighboring cardiomyocytes is mediated by gap junction ion channels such as connexin 40. These ion channels are located primarily in the heart atria and the conduction system. Most alterations in ionic currents are caused by a down-regulation of L-type Ca2+ currents [Nattel et al., 2012], an up-regulation of rectifier background K+ currents and constitutive acetylcholine-regulated K+ currents [Pandit et al., 2005], and a reduction in Na+ current [Ellinor et al., 2010]. Down-regulation of the L-type Ca2 + current resulting from the increase in intracellular Ca2 + resulting from sustained tachycardia leads to a shorter duration of action potentials. Atrial refractoriness decreases as a result of repolarizing abnormalities that are caused by changes in K + currents. A decrease in the Na+ current is also responsible for slowing the conduction velocity and reducing the reentry wavelength. Alterations in the gap function, such as downregulation and an increasingly heterogeneous distribution, may further contribute to AF-induced remodeling [Igarashi et al., 2012].

3.1.6 Risk factors of Atrial fibrillation

Age is the main risk factor for atrial fibrillation (AF); those under 50 years of age are significantly less likely to experience the disease. Atrial fibrillation is more common after the sixth decade of life, with incidence rates ranging from 0.5% in people over 50 years of age to more than 9% over 80 [Schnabel et al., 2015]. The masculine sex is another risk factor for incident AF [Claes et al., 2012]. In addition to increased left atrial pressure, atrial ectopic activity, atrial fibrosis, and slower rate of electrical transmission within and between the atria, arterial hypertension is also associated with left ventricular hypertrophy, inadequate left ventricular contractions, increased left atrial pressure, and atrial fibrosis [Lantelme et al., 2002]. Significant risk factors for atrial fibrillation include congestive cardiac failure, myocardial infarction, and mitral stenosis. These illnesses frequently co-occur and have a complicated relationship with each other. They are mutually helpful and result in a worse prognosis since they share many underlying predisposing factors and pathophysiology mechanisms [Linker et al., 2018; Santhanakrishnan et al., 2016]. Hyperthyroidism, obstructive sleep apnea (OSA), chronic obstructive pulmonary disease (COPD), and chronic renal disease are non-cardiovascular diseases that can be risk factors for AF [Buch et al., 2003; Schnabel et al., 2015; Zhao et al., 2018; Selmer et al., 2012]. People's lifestyle choices such as smoking [Baber et al., 2011], drinking [Larsson et al., 2014], engaging in rigorous activities [Chamberlain et al., 2011], and being overweight [Neefs et al., 2019] are additional risk factors for atrial fibrillation.

3.1.7 Treatment and management of atrial fibrillation

When deciding on the appropriate treatment for atrial fibrillation, patients have several factors to consider, including the type of atrial fibrillation (AF), their age, the presence of other medical disorders, and the severity of their symptoms. Controlling heart rate (also known as rate control), restoring normal cardiac rhythm (also known as rhythm control) and preventing blood clots are the three main focus of the most common preventive measures today to reduce the probability of an individual suffering a stroke. Patients must return to normal sinus rhythm (SR) and remain in SR as long as clinically feasible using rhythm control measures. People with persistent or permanent atrial fibrillation, which is typically diagnosed by routine electrocardiography, are at increased risk of having a stroke, a systemic embolism, being hospitalized, and passing away [Uittenbogaart et al., 2018; Botto et al., 2021]. Physicians can reduce this risk using the main therapy modalities of antiarrhythmic medications and electrical cardioversion.

Antiarrhythmic medication

When it comes to the treatment of atrial fibrillation (AF), medication is generally the first line of defense. Anticoagulants, which reduce the risk of stroke, antiarrhythmic drugs, which suppress irregular heart rhythms, and beta-blockers, which slow the rate of heartbeats, are the three main drugs for the treatment of AF. Anticoagulants reduce the risk of stroke. Antiarrhythmic drugs suppress irregular heart rhythms. Additionally, drugs that treat arrhythmias can prevent the condition from happening in the first place before it even begins. The purpose of rate control is to lower the patient's heart rate to a typical healthy level when they are at rest. This is achieved using drugs such as digoxin, beta-blockers, and Ca2+ channel blockers. Anticoagulants, often called blood thinners, are a component of preventive treatment for people with AF. This is because the condition is associated with an increased risk of stroke. When selecting the appropriate antithrombotic drug for a patient, it is essential to take into account a variety of aspects, including the patient's medical history, as well as the patient's probability of experiencing bleeding and the presence of risk factors for stroke [January et al., 2014].

Patients often continue to take antiarrhythmic medications after the successful initial restoration of sinus rhythm (SR). This is because the use of these medications helps patients maintain SR and minimizes the recurrence of AF. Suppose that rhythm restoration is unsuccessful, cannot be held for extended periods, or is not acceptable due to the patient's specific requirements. In that case, the rate control technique is used instead. However, if rhythm restoration is successful, it can be sustained for extended periods. When medication is not an option, or it is impossible to properly treat the problem, a pacemaker can be surgically implanted in the chest to control and lower the heart rate [January et al., 2019].

Electrical cardioversion

Electrical cardioversion is a technique in which a direct electrical shock is applied to the heart using a defibrillator. This shock helps restore the heart rhythm to normal. This method is recommended when there is a pressing need to act quickly or when patients are experiencing acute symptoms [DeSilva et al., 1980].

Catheter ablation

Catheter ablation of atrial fibrillation is an invasive procedure designed to remove defective electrical pathways leading to AF. To do this, small scars are formed in the involved cardiac tissue. An energy-emitting probe is implanted at the tip of a catheter to create these scars. This probe, which supplies the

ablating energy, can damage the tissue by applying radiofrequency energy (heat generated by medium frequency AC) or cryothermic energy (severe cold). Figure 3.7 illustrates both procedures. To reach its working zone, this catheter passes through the femoral vein, the inferior cave vein, the right atria, and the left atria.

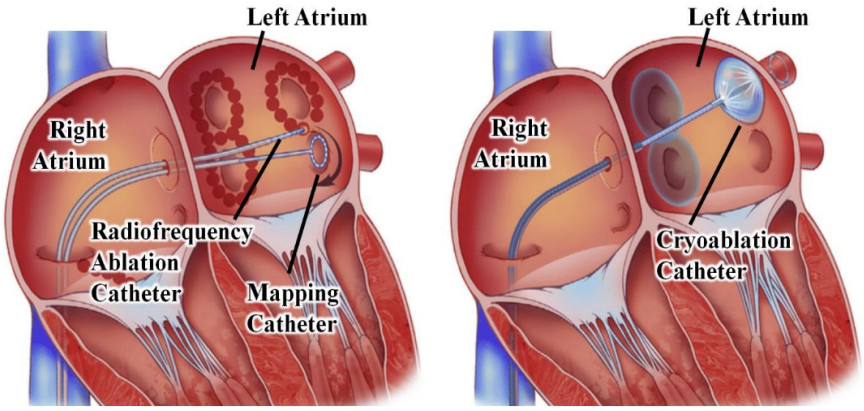

Fig. 3.7 Catheter ablation using radiofrequency ablation (left) and cryoablation (right). From [Calkins et al., 2017].

A comprehensive investigation has been carried out to precisely locate the parts of the atrium that need to be targeted by the catheter ablation process. The ostium of each of the four pulmonary veins was found to be a significant trigger for atrial fibrillation [Haissaguerre et al., 1998]. Since then, isolating the pulmonary veins, sometimes called pulmonary vein isolation (PVI), has become the technique for ablation. To do this, scars are formed in an elliptical pattern throughout the ostium of each vein. At some point in the procedure, the heart surgeon should be careful not to disconnect the vein. In contemporary medicine, there is also a practice that involves wrapping the ostium of the pulmonary vein on one side with a single wider elliptical line [Burstein et al., 2008].

When atrial fibrillation progresses from paroxysmal to persistent atrial fibrillation, the AF drivers extend through both atriums beyond the PVs [Brooks et al., 2010]. Identifying the position of the drivers is a difficult challenge to solve [Hansen et al., 2018; Stiles et al., 2018]. Research has indicated that the percentage of patients who require additional PV ablation has increased from 31% (in patients with paroxysmal AF) to 55% (in patients with chronic AF) to 75% (permanent AF). These studies came to an important conclusion: almost 62% people with persistent or permanent AF have AF drivers in the right atria [Lim et al., 2017; Miller et al., 2017]. Furthermore, atrial remodeling was discovered in both atrial chambers of the patient with chronic or permanent atrial fibrillation [Prabhu et al., 2017]. Therefore, improved patient-specific visualization and reconstruction of atria geometry can potentially guide the targeting of the driver regions and improve the elimination of AF substrates [Hansen et al.,

2015; Zhao et al., 2012]. This is because enhanced patient-specific visualization and reconstruction of the atria geometry allow for better targeting of driver regions.

However, despite the encouraging effects that PVI has on suppression of AF compared to pharmaceutical methods, the prevention of successful AF recurrence is still not at its optimal level. The success rates of patients with persistent AF are much lower than those with paroxysmal AF [Parkash et al., 2011]. Clinical trials in multiple centers reported a 40% to 70% AF termination rate one year after ablation for patients with paroxysmal AF [Morillo et al., 2014; Cosedis Nielsen et al., 2012].

3.1.8 Earlier methods for detection of atrial fibrillation

A growing number of researchers have classified ECG data using convolutional neural networks (CNN) to help detect arrhythmias. To automatically distinguish different segments of ECG, a CNN technique was described [Acharya et al., 2017]. The method used an 11-layer deep CNN with a four-neuron output layer, each representing a different type of ECG. This study applied two sets of ECG signals. The first set of signals lasted two seconds, and the second set lasted five. Both types of signals were classified using the identification of the QRS complex. For the initial data, the model reached accuracy, sensitivity, and specificity values of 92.50%, 98.09%, and 93.13%, respectively. ECG signals were classified in [Kiranyaz et al., 2015] using a one-dimensional CNN.

The ECG classification process was divided into sections for feature extraction and classification using CNN. Based on the results, the recommended method surpasses the previous techniques for categorizing ECG beats using the MIT-BIH database. To discriminate between different heartbeat patterns, [Rai et al., 2018] used a wavelet neural network (WNN). In this study, the ECG information from the signals from the MIT-BIH arrhythmia database was classified [Goldberger et al., 2000]. A WNN was used to classify the signals and compute the wavelet coefficients for QRS. The obtained findings showed a 98.08% accuracy. The same database was utilized in [Kaur et al., 2012], however, the signal characteristics used were the wavelet coefficients and the QRS complex. Factor analysis with and without orthogonal rotation was used to reduce these features. The MIT-BIH arrhythmia database was then classified with an accuracy of 99.06% using a linear discriminant analysis (LDA) classifier. In [Li et al., 2017], ECG signals were classified using a 1D CNN. The proposed CNN model contained five layers: input and output layers, two convolution levels, two downsampling layers, and one completely connected layer. This model retrieved the properties of the ECG signals, which labeled the signals as regular, left-bundle branch block, right-bundle branch block, premature atrial contraction, or early ventricular contraction. The study team used the MIT-BIH arrhythmia database,

which contains 48 half-hour segments of two-channel ambulatory ECG recordings of 47 individuals who were the subject of investigations by the BIH arrhythmia laboratory between 1975 and 1979. According to the findings, the recommended method has a classification accuracy of 97.5% [Li et al., 2017], outperforming several popular ECG classification methods.

For the classification of ECG data of any duration, two Deep Neural Network (DNN) designs were presented in [Zihlmann et al., 2017]. The first architecture consists of deep CNN aggregation features using a temporal average. For feature extraction and convolutional layers for feature temporal aggregation, the second approach uses LSTM layers. The performance evaluation of these networks was performed using the Physionet AF classification data set. Combining such networks increased the classification accuracy of ECG signals to 82.1%. The same dataset was used in [Andreotti et al., 2017] to classify short fragments of ECG signals. This study compared a CNN approach and a state-of-the-art feature-based classifier. Both algorithms were trained using Physionet data. The feature-based classifier has a 72.0% accuracy with a five-fold cross-validation in the training set and a 79.0% accuracy in the concealed test set. CNN received scores of 72.1% in the test set and 83.0% in the expanded database. A final score of 79.0% was obtained using the latter approach.

In [Rajkumar et al., 2019] a CNN and DL algorithm were employed in to classify the ECG data. To train and evaluate CNN, the researchers used the MIT-BIH database. The features were automatically retrieved and compared with various activation functions and epoch counts using a CNN. The results show that, with a 93.6% accuracy, the Exponential Linear Unit (ELU) activation function has the highest accuracy. ECG signals were classified using branching programs (BR) and (NN) in [Mansouri et al., 2019]. The first technique used an autoregressive (AR) classification algorithm to extract AR features. The coefficients of the AR model offer a signal approximation that can be applied to automatically extract features. ECG data were classified using a binary decision tree, and each heartbeat was classified using four coefficients of the AR model. The second method classified signals using a NN with two layers and six neurons in each layer, using the same four AR coefficients as input. For the BR and NN models, the categorization of the ECG signals was 86.35% and 88.57%, respectively. RhythmNet was created in [Xiong et al., 2018a] to classify ECG signals, including AF. The Rhythm-Net architecture is a 21-layer 1D convolutional recurrent neural network that uses three recurrent layers to scan ECGs of various durations and identify arrhythmia events in lengthy recordings after starting with 16 convolution layers to extract properties from ECG waveforms. Large convolutional filters were used to rapidly learn signal properties, including P waves and QRS complexes, while the MIT-BIH arrhythmia from Physionet was used for training. An accuracy of 82%

was achieved. To divide the ECG signals into eight categories, [Ullah et al., 2020] created a 2D CNN model. Four convolutional layers and four pooling layers make up this model, which was built to extract characteristics from the spectrograms. The 1D ECG time series data is transformed into 2D spectrograms using a short-time Fourier transform. This method achieved a classification accuracy of 99.11% using the MIT-BIH arrhythmia data set.

Principal component analysis (PCA), local discriminant analysis (LDA), and a probabilistic neural network (PNN) classifier are combined in a feature reduction technique suggested by [Wang et al., 2013] to classify arrhythmias based on ECG beats. With a sampling rate of 360 Hz, this technique extracted 200-valued ECG beat samples from ECG signals. The PNN was then trained to distinguish different types of ECG beats. The general classification accuracy of the proposed method was 99.71%. Using PCA, LDA, and Independent Component Analysis (ICA) applied singly in Discrete Wavelet Transform (DWT) sub-bands in [Martis et al., 2013], arrhythmias were detected to decrease the number of dimensions. After that, the PNN and support vector machine (SVM) classifiers received input from the reduced-dimensionality characteristics. The results showed that the ICA features and a PNN classifier worked better than PCA and LDA. PNN specifically achieved mean values of 99.97%, 99.83%, 99.21%, and 99.28% for sensitivity, specificity, positive predictive value (PPV), and precision with ten-fold cross-validation. A dynamical model of ECG signals was used to offer a new classification method for ECG data in [Vafaie et al., 2014]. In this method, the signals are classified using a fuzzy classifier. The simulation results, 93.34% of the ECG signal, could be correctly classified. The performance of this classifier was improved using a genetic algorithm, increasing the prediction accuracy to 98.67%. A categorization algorithm for arrhythmias was created using morphological and dynamic data in [Anwar et al., 2018]. Each heartbeat was sent through a Discrete Wavelet Transform (DWT) to extract morphological characteristics. To capture the nonlinear dynamics of the RR intervals, the Teager Energy Operator was applied to the RR interval as a dynamic feature (TEO). Furthermore, redundancy was reduced by employing an ICA of the DWT subbands.

3.2 Contribution

3.2.1 Early Warning of Atrial Fibrillation

Maintenance of sinus rhythm (SR) in patients can alleviate some of the symptoms of atrial fibrillation (AF) and further prevent atrial remodeling that can increase the risk of future episodes [Prystowsky, 2000]. As a result, early prediction of AF episodes in patients with paroxysmal AF can prevent

emergency room visits and associated expenses associated with healthcare. However, the correct identification of patients with a high risk of developing AF and its prediction of early warning shortly are significant problems in the clinical scenario [Wilson et al., 2020]. Therefore, in this chapter, we developed a method to anticipate and provide early warnings of the imminent onset of AF, using only data that can be collected by simple wearable devices, such as smartwatches.

Automated detection of AF regimes from recorded time-series data is a well-studied subject in the literature. Recent studies based on machine learning and neural networks have achieved a precision greater than 99% in the classification job (see Section 3.1.8). Despite this, there is no clear way to predict the onset of atrial fibrillation [Aljanabi et al., 2018]. Several methods have been proposed to assess the risk of atrial fibrillation over a period ranging from months to years [Attia et al., 2019; El Moaqet et al., 2017; Biton et al., 2021; Raghunath et al., 2021]. However, none of these methods can produce real-time forecasts. Research done in the past to predict the onset of AF can be divided into two main categories. The first category consists of methods that can be applied to one or more ECG leads and involve the extraction of features from premature atrial contractions, ectopic atrial or ventricular beats, and spectral analysis [Langley et al., 2001; Zong et al., 2001; Thong et al., 2004; Yang et al., 2001; Aytemir et al., 1999; Clavier et al., 2002; Blanche et al., 2013; Boon et al., 2016; Boon et al., 2018; Alcaraz et al., 2015]. These methods can be performed on single or multiple ECG leads. The second category deals with the analysis of R-to-R intervals (RRI), which often involves hand-crafted feature extraction and power spectrum analysis [Chesnokov, 2008; Tateno et al., 2001; Sarkar et al., 2008; Dash et al., 2009; Lee et al., 2013; Mohebbi et al., 2012; Guo et al., 2021]. Or a combination of the two, often comprising the analysis of the morphological features of the ECG and RRI in conjunction with convolutional neural networks (CNN) and automatic feature extraction algorithms [Attia et al., 2019; Erdenebayar et al., 2019; Cho et al., 2018; Tzou et al., 2021; Jalali et al., 2019; Hannun et al., 2019; Jo et al., 2021; Mahmud et al., 2020; Costin et al., 2013; Kim et al., 2016; Shen et al., 2016; Li et al., 2018b; Ebrahimzadeh et al., 2018]. These hand-made features have major disadvantages, such as being computationally expensive, having the ability to contribute to human bias, and having to require long-term samples, making them unsuitable for real-time monitoring applications [Matias et al., 2021].

In general, despite these recent advances, there are still important limitations. The main one is that all of these methods provided “predictions of AF at ’onset.’ In other words, the data window used for the prediction comprises the time series signal sampled up to the onset of AF. Therefore, it provides

zero early warning (Fig. 3.8, middle). Standard detection algorithms use AF data (Fig. 3.8, right). This chapter presents a method for early warning prediction of AF long before onset (Fig. 3.8, left).

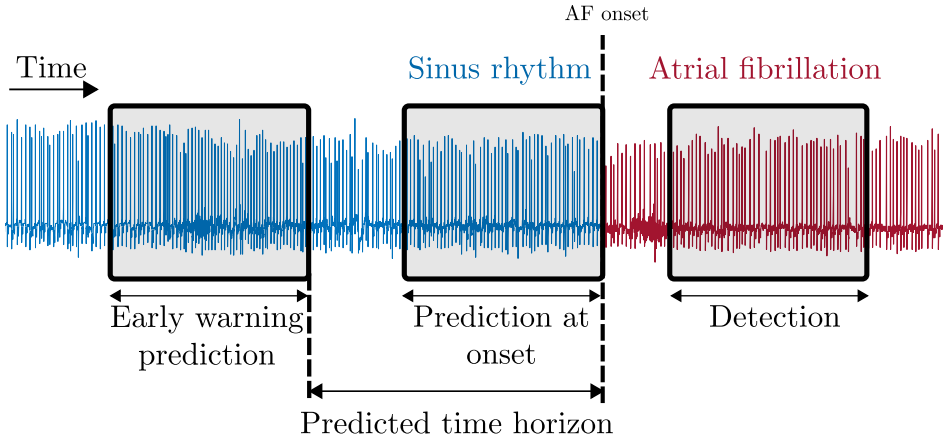

Fig. 3.8 Detection versus prediction. Early-warning AF prediction (left), AF prediction at onset (middle), and AF detection (right). All methods are based on time-series windows sampled at different instants concerning AF onset.

A limitation of ECG data is that long-term recordings are difficult to acquire, requiring the inconvenient and impractical use of Holter devices or patches. Although simple smartwatches and smart bands can continuously capture RRI during the daily life of individuals, existing methods using RRI tend to require large time-series windows of up to 30 min length for prediction at onset [Guo et al., 2021; Mohebbi et al., 2012; Boon et al., 2016]. This limits its application for short-term predictions. Moreover, when the RRI duration was reduced to less than 30 min in previous methods, poor prediction accuracy was obtained [Matias et al., 2021].

We present a retrospective study that develops a method, based on artificial intelligence (AI), for the early warning of AF from moving RRI windows with short-term duration, entitle WARN (Warning of Atrial fibRilatioN). To avoid the limitations mentioned above, this study proposed training a deep CNN using RRI samples that are 30 seconds long and a flexible temporal horizon for our predictions. As a result, we did not require feature selection and used RRI samples that are noticeably shorter than those of previous research. Furthermore, our method does not require ECG data; instead, it uses RRI signals that can be sampled by inexpensive and easy-to-use pulse signal recorders, such as smartwatches or smart fitness bands. This makes the method significantly more convenient. Patients could continue to use these devices, paving the way for real-time monitoring algorithms that could learn and monitor long-term cardiac dynamics.

3.3 Methods

3.3.1 Biomedical datasets

The original data was obtained from Tongji Hospital, part of the Huazhong University of Science and Technology in Wuhan, China. It consists of a 12-lead long-term ECG Holter of 595 patients, and each ECG is recorded in SR at baseline. Furthermore, each ECG contains at least one AF event. The beginning and end points of each episode of AF were labeled by qualified cardiologists from Tongji Hospital. The records have an average duration of between 22.2 ± 2.2 hours, have a sampling frequency of 128 hertz, and have a resolution of 12 bits.

Additional data from patients with AF from different health centers were considered: Clinica y Maternidad Suizo Argentina (53 patients with 24 h ECG) and the open access database Atrial Fibrillation Prediction Database (AFPDB) from Physionet [Moody et al., 2001]) (75 patients with 30 min ECG). This was done to externally validate the performance of WARN on "out-of-distribution".

Exclusion criteria

Records that did not have an episode of SR or AF were excluded. Furthermore, records that began with AF were excluded because the ECG section preceding AF cannot be segmented. Only episodes of AF that lasted 10 min or more were considered. Finally, records with severe noise artifacts before the onset of atrial fibrillation (by determining whether the proportion of missing R peaks within a sliding window of 5 min exceeded 15%).

Following the exclusion criteria, the remaining 350 records were used for this study. The cohort was divided into two groups in chronological order between 2014 and 2019. The first 80% (280 patients) were used for the training/cross-validation of the model (252 for training and 28 for validation), while the latter 20% (70 patients) were used for testing (Table 3.1).

Table 3.1 Characteristics of the patients.

Characteristic	Training Cohort (total = 280)	Test Cohort (total = 70)
Age $<$ 65 (mean)	115 (55 years old)	31 (54 years old)
Age $\geq$ 65 (mean)	165 (73 years old)	39 (73 years old)
Male	163	26
Female	117	44

The Physionet database consists of 50 healthy controls (SR) and 25 patients with AF (with ECGs separated into 30 min just before the onset of AF and 5 min immediately thereafter). This cohort excluded five patients with a duration of AF of less than 1 min. Furthermore, two healthy control records showed severe distortion AF resulting from artifact noise and were therefore eliminated. Consequently, there were 20 records to predict AF and 20 records randomly selected from healthy participants to serve as controls. In general, 48 patients were considered for external validation; 8 were from the Clinica y Maternidad Suizo Argentina, while the remaining 40 were from the Physionet database.

Ethical considerations

To safeguard the privacy of the patients, the data was anonymized. The data collection teams of each center were responsible for sample collection and anonymization. Only age and sex information was provided for this study. The study design was examined and excluded from a comprehensive evaluation by the Institutional Review Board of the Huazhong University of Science and Technology (approval number: TJ-IRB20220423) and approved by the Ethics Review Panel of the University of Luxembourg (approval number: ERP 22-057 RTMonitor). All data were collected according to the principles of the Declaration of Helsinki [Association et al., 2009].

3.3.2 Pipeline of WARN

The WARN algorithm for early detection of AF comprises two stages. First, WARN is trained to recognize three cardiac rhythms: SR, AF, and Pre-AF (the instances just before the onset of AF). Second, this model progressively analyzes the RRI data to track the probability of an imminent transition to AF. When this probability exceeds a predetermined threshold, an alert is triggered.

3.3.3 First stage of WARN

The first stage of WARN is summarized in Figure 3.9, and begins by segmenting each ECG recording into three classes: SR, Pre-AF, and AF (Fig. 3.9a). The onset and end points of each episode of AF were labeled by experienced cardiologists in the hospital. The SR segment corresponds to the complementary segment of the labeled AF segments.

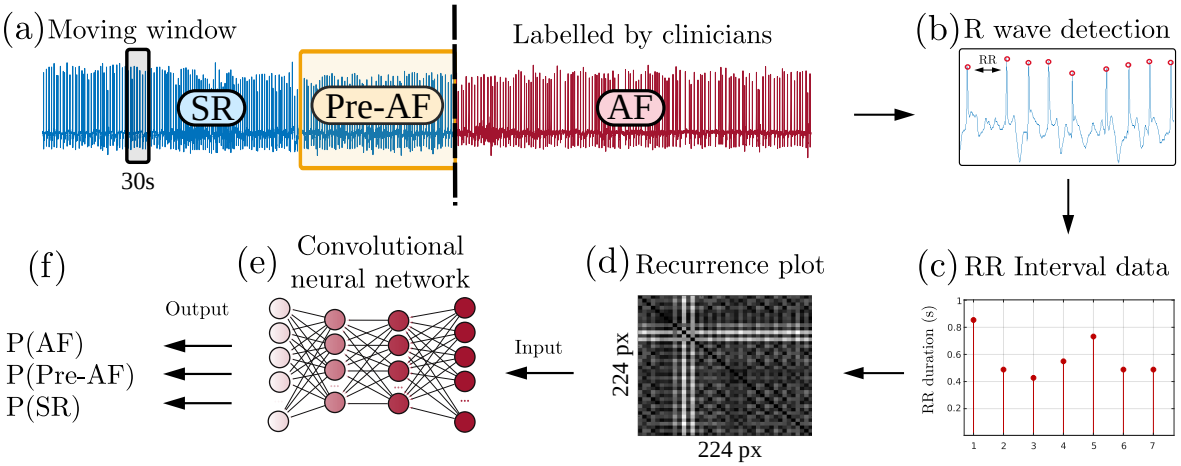

Fig. 3.9 Pipeline of the first stage of WARN. (a) Every ECG record is divided into three classes: SR, Pre-AF, and AF. (b) R peaks are discovered in the ECG data using a 30-second sliding frame. (c) The R peaks generate the RRI signal. The RRI signal is used to create a recurrence graphic. e) The recurrence plots are used as inputs to train a deep CNN. (f) Network output consists of the probability that sampled data belongs to each of the three classes (SR, Pre-AF, and AF).

The pre-AF segmentation process consists of five steps, each of which can be seen in Fig. 3.10 for a patient who serves as a representative example.

1. First, beginning with the time point where the doctors indicated the beginning of AF, a sliding window is built to extract data of 5 min long ECG data with 30 s overlap; this sliding window travels backward in time.

2. Second, a second sliding window is used within each five-minute frame to obtain more granular samples of 30 seconds every five seconds.

3. Third, the detection of R waves is performed for each 30-second window using the Pan-Tompkins algorithm (see Section 2.4.1 for more details about this algorithm). Once the R waves are detected, the difference between the heartbeats (i.e. from R to R peaks) is calculated to generate the RRI signal.

4. Fourth, the coefficient of variation of RRI is calculated for every 30-second window and the related histogram is constructed for every 5-minute window.

5. Fifth, the evolution of the frequency distribution is tracked until the median is less than 0.7. The threshold of 0.7 is chosen as the intercept point between the frequency distributions of the coefficient of variation for the AF and SR regimes (Fig. 3.11) of patients in the training set. This threshold is used to assess heart variability. When the median of the distribution of

the coefficient of variation of a given window of 5 min is below this threshold, it shows low variability. Therefore, it can be related to SR [Lan et al., 2020]. The Pre-AF segment is labeled in this fifth and final stage, comprising the beginning of this last window before the onset of AF.

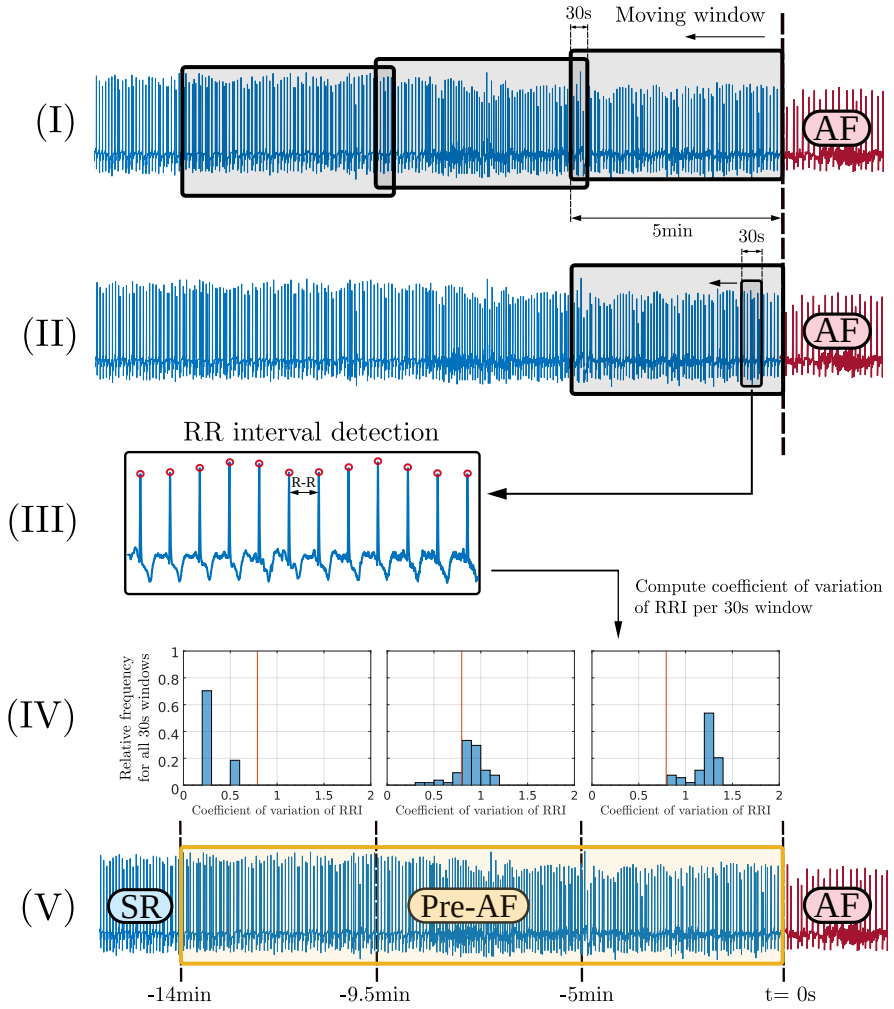

Fig. 3.10 Process of pre-AF labeling for a sample patient. Starting from the onset of atrial fibrillation and moving backward, a sliding window is formed to retrieve ECG data of 5min with 30 s of overlap. (II) A second sliding window is formed for each 5-minute window to retrieve 30-second samples every 5s. (III) R waves are identified, and RR intervals are calculated for each 30 s window. (IV) The coefficient of variation of all 30-second windows within a 5-minute window is computed. The procedure is then repeated for each 5-minute interval to build an RRI coefficient of variation histogram. (V) When the median of the histogram of RRI is less than 0.7, the Pre-AF section is segregated from the start of the last window until the onset of AF. In this instance, Pre-AF lasts for 14 min before the onset of AF.

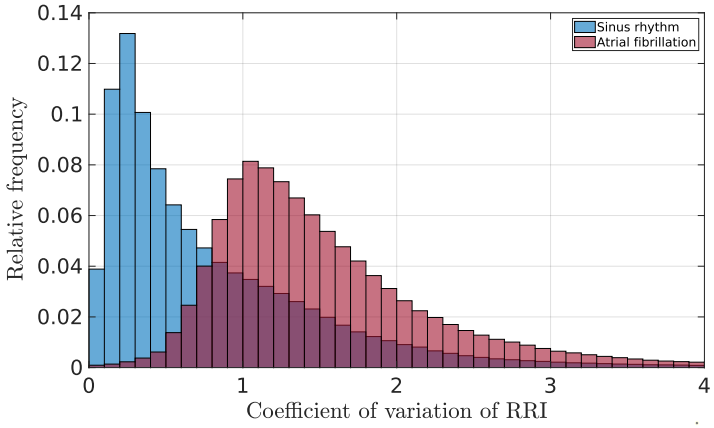

Fig. 3.11 Distributions of the coefficient of variation of the RRI for all patients from the training set, split by SR and AF regimes (as labeled by clinicians).

Unlike the SR segments, which are often characterized by low RRI variability, Pre-AF is characterized by high RRI variability [Lan et al., 2020]. Furthermore, the Pre-AF segments vary in duration from patient to patient [Narin et al., 2018] and within multiple onsets of AF for the same patient as a result of morphological and electrical cardiac alterations [Pławiak, 2018].

After segmenting the ECGs, samples of 30 seconds are retrieved and transformed into RRI data by first extracting the R waves (Fig. 3.9b) with the Pan-Tompkins algorithm (see Section 2.4.1 for more details on the algorithm). The RRI is then calculated by subtracting the time from R to R waves (Fig. 3.9c). Subsequently, the 30-second RRI samples were converted into recurrence plots (Fig. 3.9d) by following the algorithm described in Section 2.4.2.

Recurrence plots are a visual representation of RRI signals in this context. In the literature, it has been shown that they can be used to discover dynamical transitions in time-series data [Trauth et al., 2019; Portes et al., 2019] and to find hidden periodicities in physiological signals [Sun et al., 2008; Carvalho et al., 2018; Zhang et al., 2021]. For example, the SR and AF states are represented by recurring spatio-temporal patterns of certain rhythmicity, which can be recognized using recurrence plots [Censi et al., 2000].

Figure 3.12 shows the recurrence plots for various regimes (SR, Pre-AF, and AF). When the SR and AF regimes are compared, the number of states with high recurrence decreases on average (Fig. 3.12b), showing that the system's periodicity reduces. Pre-AF segments emphasize a dynamical transition with a large number of "cross-shaped" low-recurrence regions caused by RRI signal intermittency, that is, the alternating between periodic (SR) and non-periodic (AF) regimes. Despite a considerable difference in the average recurrence plots in Fig. 3.12b compared to Fig. 3.12a, there are hidden trends in the time series data that would be impossible to detect in real-time using conventional statistical

approaches. As a result, we hypothesize that a deep learning model may detect such patterns, including the brief Pre-AF state before the onset of AF.

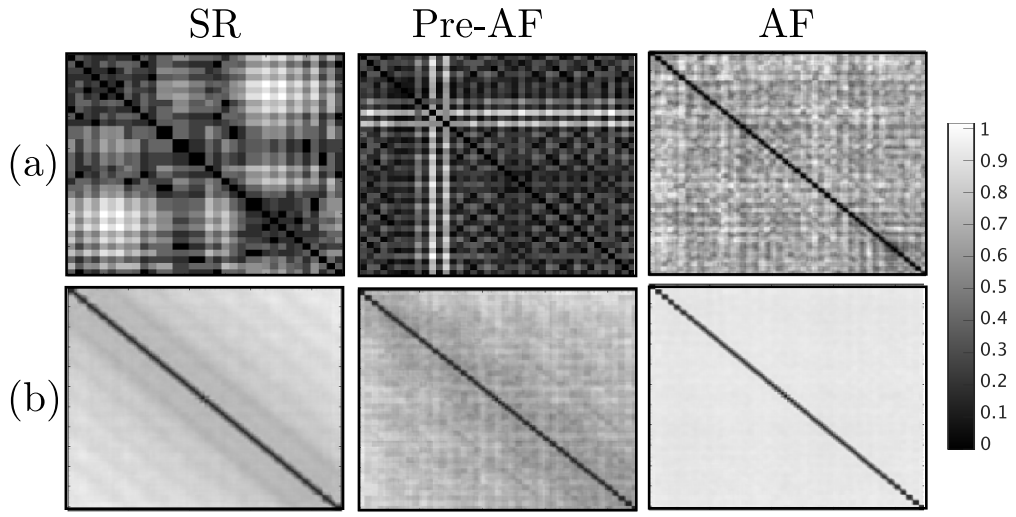

Fig. 3.12 Recurrence plots for each segment (SR, Pre-AF, and AF) created from (a) a single 30-s sample and (b) the average of all recurrence plots generated for a representative patient.

EfficientNetV2 is used as the deep learning model to detect patterns in the recurrence plots. The specifications of this deep learning model are described in Section 2.3.7. As shown in Fig. 3.9e. The recurrence plot serves as input to the network; the plots have a dimension of (224 x 224) pixels. The network is then trained and cross-validated in random samples of the recurrence plot from 280 patients. Categorical cross-entropy was used as the loss function, ADAM was used as the optimizer, and stochastic gradient descent was used as the objective function optimizer. The data is divided into three classes to handle the sample imbalance between the three regimes (SR, Pre-AF, and AF) during training. The data was resampled, and the loss function was weighted according to the ratio 3/1/2 for the SR, Pre-AF, and AF samples, respectively. The training was terminated when the validation loss did not decrease for 8 consecutive epochs. Finally, the output of the network is the probability that the input belongs to one of the three regimes (see Fig. 3.9f).

Model optimization

We investigated the effect of the sample window length used to construct the recurrence plot from the RRI data on the performance of the method. Starting from 10 to 5 min we calculated the average accuracy to predict individual samples using the EfficientNetV2 (Table 3.2). The best result was found with a window length of 30 seconds, which was previously reported in [Erdenebayar et al., 2019]. Performance changes are caused by trade-offs between the number of samples generated and

the window duration. The broader the window, the fewer samples are obtained for training, which decreases WARN's ability to generalize the data adequately. On the other hand, a shorter window length can result in information loss [Kraemer et al., 2018].

Table 3.2 Optimal length of the sampling window.

Length (seconds)	Samples ($\times 10^6$)	Accuracy
10	2.3	0.70
30	0.8	0.74
60	0.4	0.72
120	0.2	0.69
300	0.1	0.66

After limiting the sampling window to 30 seconds, we compared the performance of stage 1 of WARN with that of two other network benchmarks: 1-D CNN (see Section 2.3.5) and LSTM (see section 2.3.4), both of which are commonly employed for arrhythmia detection and prediction [Oh et al., 2018; Somani et al., 2021]. The suggested WARN model outperformed the benchmark networks (Table 3.3), with an average validation accuracy of 0.74 and good data generalization (a modest standard deviation of 0.03 in all 10 folds). Finally, the best model (model 1) was chosen for the performance analysis of the 70-patient from the unseen test set.

Table 3.3 10-fold cross-validation accuracy after training WARN and two benchmark networks.

Model	1	2	3	4	5	6	7	8	9	10	Average
EfficientNetV2	0.78	0.71	0.73	0.73	0.73	0.71	0.75	0.75	0.76	0.71	0.74
LSTM	0.64	0.61	0.63	0.61	0.61	0.58	0.60	0.63	0.57	0.60	0.60
1D-CNN	0.71	0.68	0.67	0.71	0.69	0.67	0.68	0.67	0.70	0.70	0.69

In addition, we compared the performance of stage one of WARN using ECG data instead. To reduce the dimensionality of the 12-lead ECGs, only one lead (Lead II) was considered. Since it is commonly used for diagnostic purposes and maximizes the P waves amplitude [Crawford et al., 2012].

We considered two different 2D representations for this comparison; ECG spectrogram, by applying the STFT with 128-sample segments and a periodic Hann window (see Section **??** for more details), and recurrence plots from ECG data.

The performance improved when using spectrograms and recurrence plots of the ECG data by 5.12% (see Table 3.4). This was expected since the ECG data is sampled at a higher frequency, and information on the P-waves is still contained. In contrast, the RRI data do not contain such information. However, the RRI data is widely accessible and easier to measure.

Table 3.4 Performance for different input representations.

Length (seconds)	Accuracy
Recurrence plot of RRI	0.74
Recurrence plot of ECG	0.78
Spectrograms of ECG	0.78

Interpretability of the neural network output

We implemented the gradient-weighted class activation mapping method (Grad-CAM) described in Section 2.3.8 to gain a deeper understanding of the NN decision-making process. We use the model that was trained on the spectrograms of the raw ECG data because it allows easier interpretability, such as the relevance of the magnitude of particular frequency bands in the spectrograms. The raw ECG plots and their corresponding spectrograms are shown in Figure 3.13; these were taken from a patient who served as a representative example for the test set. We took data from each of the three cardiac states (SR, Pre-AF, and AF), then generated a heat map using Grad-Cam. The activation of the NN was observed in both the low- and high-frequency bands for the SR samples (Fig. 3.13a). The activation of the Pre-AF samples was observed mainly in the low- to middle-frequency range (Fig. 3.13b). Finally, it is possible to observe activation taking place at high bands during AF (Fig. 3.13c).

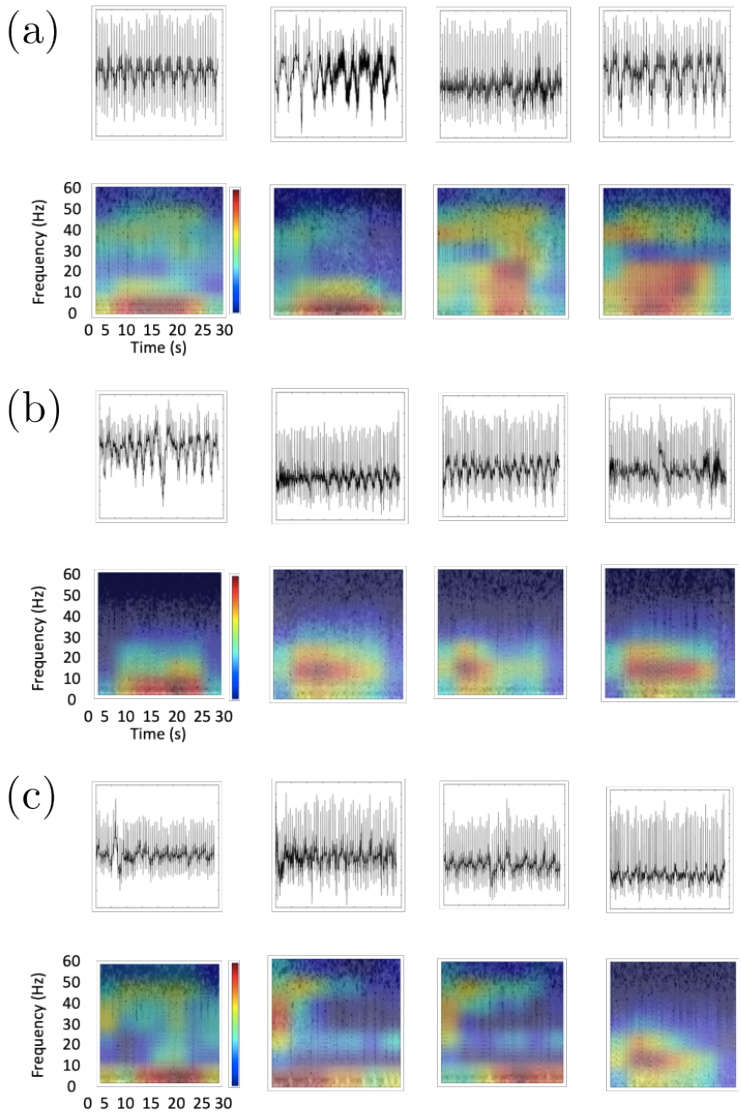

Fig. 3.13 Raw ECG plot (top), Gradient-weighted Class Activation Mapping (bottom) of the corresponding spectrogram from samples of a representative patient: SR (a), Pre-AF (b) and AF (c)

If we consider averaging the activation heat maps of multiple samples from multiple patients, we can better understand where the activation occurs. Figure 3.14 shows that there is no clear pattern between the SR and Pre-AF samples. However, for Pre-AF samples, reduced activation can be observed between the 10 and 20 Hz range. For the AF samples, the activation occurs mainly in a high-frequency band of the spectrogram.

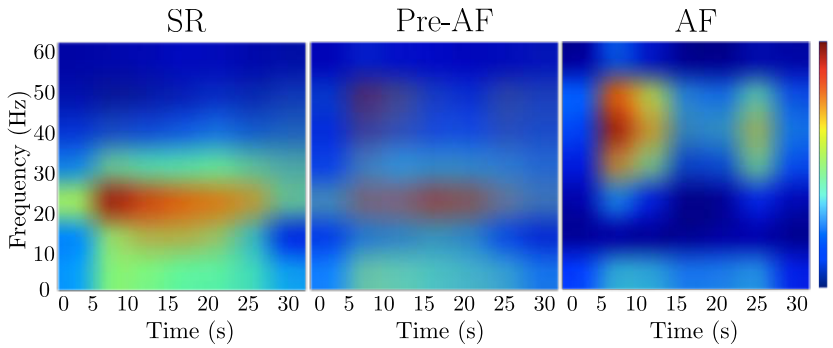

Fig. 3.14 Average heat map activation using Grad-CAM from spectrogram samples labeled as SR, Pre-AF, and AF segments of all patients.

3.3.4 Second stage of WARN

The second stage of WARN's algorithm computes the probability of a patient switching to AF in the near future from the outputs of the trained CNN. Define the probability of danger as $P(\text{danger}) = P(\text{Pre-AF}) + P(\text{AF})$, which represents the probability that a sliding window is in either Pre-AF or AF states. From the input time-series RRI data, the algorithm sequentially generates a sliding window of 30s every 15s. For each new window, the recorded 30s time-series are converted into a recurrence plot and fed to the CNN for classification. Figure 3.15a illustrates the probability of danger computed by WARN for a representative example. Since the probability of danger has very high variability, we implemented a non-anticipative moving average window to filter this high-frequency noise and smooth the output (Fig. 3.15b). A binary early-warning indicator ("danger" or "no danger") also requires the selection of a particular threshold of the probability. These two hyperparameters, the moving average window length, and probability threshold, can be optimized to maximize different performance metrics, depending on the needs of particular patients.

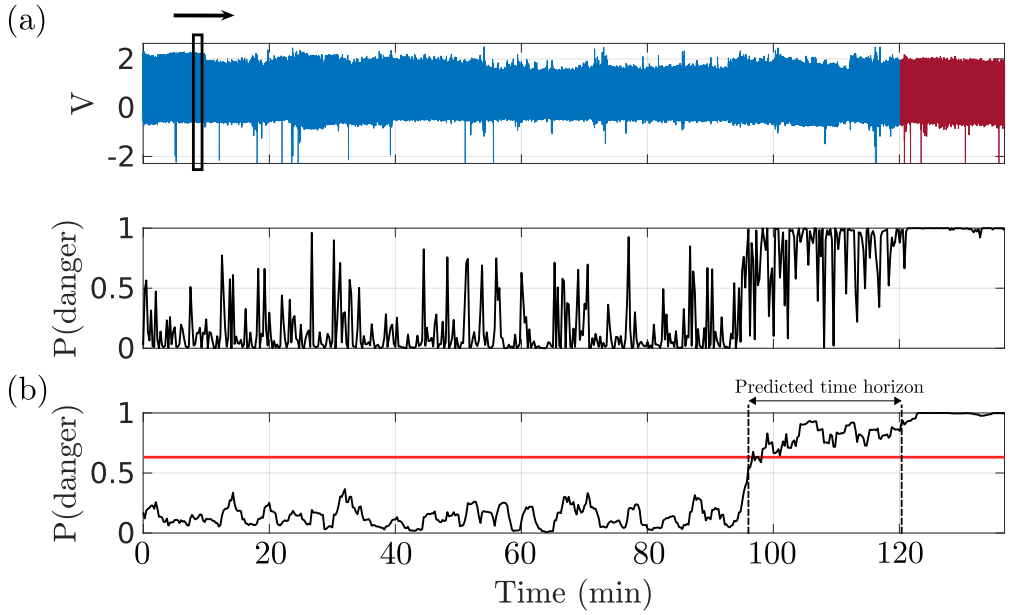

Fig. 3.15 Second stage of WARN: an early warning indicator. (a) WARN computes the probability of danger as a function of time for a representative patient in the test dataset. The sample images are generated by sampling a sliding window of 30*s* every 15*s*, and the danger probability is calculated for each sampled window. (b) The average danger probability is computed with a non-anticipative moving average window of 7 samples to smooth out the probability variation. The red line is a 0.57 threshold that will trigger an alert before the onset of AF.

The performance of WARN was evaluated on a time series of 60min of sequential data before AF onset and also far from AF (in SR), selected patient wise. This is given by the fact that more than 70% of Pre-AF segments duration are shorter or equal to one hour (see Fig.3.16). The selection of samples far from AF is performed randomly, at least 2 hours before AF, to guarantee that the median value of the coefficients of variation of the RRI signals (computed over the selected 60min sample) is close to the median value associated with the SR distribution computed over all patients (Fig. 3.11). Therefore, the data were evenly divided into 60 min samples in order to preserve class balance and allow fair comparisons within the confusion matrix.

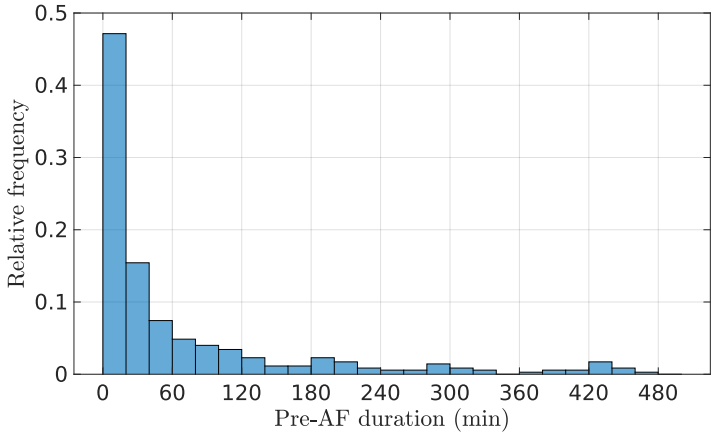

Fig. 3.16 Distribution of Pre-AF length for all patients.

Parmeters optimization

For the validation data (corresponding to the cross-validation with the highest performance in the CNN), as expected, it is not possible to simultaneously maximize all performance measures. For example, Figs. 3.17a-c show that the maximum predicted time horizon until AF onset (that is, the instant of the first early warning until AF onset, see 3.15b)) is achieved at low thresholds (Fig. 3.17b,c). However, the accuracy is very low for those values (bottom of Figs. 3.17a). Likewise, when the accuracy is maximized at 88.3%, the predicted time horizon is relatively short.

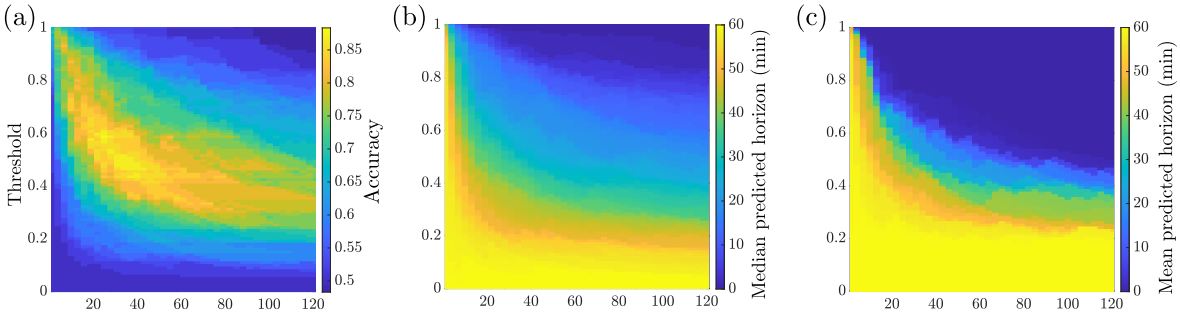

Fig. 3.17 Performance of WARN on the validation set. (a) Model accuracy, (b) mean, and (c) median of the predicted time horizon before AF onset as a function of the probability threshold and the size of the moving average window that smooths the probability of danger of RRI data.

To achieve a trade-off in the validation set, we searched for hyperameters where the accuracy, sensitivity, specificity, and F1 score are all greater than 80%, and mean and median predicted time horizon are above 30min. There was a total of 34 hyperparameters satisfying this criterion. We picked the smallest moving average window, since this leads to lower computation and memory usage in smart devices, and, among those, the one that maximizes accuracy, F1 score, and mean and median predicted time horizon. The resulting hyperparameters are a moving average window size of 7 samples

(corresponding to a 1.5min window) and a threshold of 0.57. This produced an accuracy of 86.7%, F1 score of 87.5%, sensitivity of 93.3%, and specificity of 80%. The mean (median) predicted time horizon until onset of AF is 31.4min (36.3min). Figure 3.15b illustrates the moving average of the probability of danger and the threshold for a representative patient. In this example, an early-warning signal is alerted to the patient by WARN with 22min in advance of the AF onset. These hyperparameters can be adjusted according to different objectives by giving more importance to, for example, sensitivity, specificity or size of the predicted time horizon.

3.4 Results

3.4.1 Performance on test data

WARN was evaluated in the test data set as a predictor of the onset of atrial fibrillation (AF). Figure 3.18 provides a summary of the results. The threshold selection for a fixed moving average window depends on the patient's needs. For example, the threshold may be higher and more precise for a healthy user to avoid false positives. Reduce false negatives for a patient at risk by lowering and increasing the threshold sensitivity. Figure 3.18 shows the trade-offs between sensitivity and specificity for three threshold values (0.57 as default, 0.73, and 0.88). Figure 3.18a shows the mean and median values of the anticipated time horizon for the beginning of AF for all patients, that is, the time from the first early warning to the onset of AF (Figs. 3.8 and 3.15b). With an accuracy of 82.7% and a threshold of 0.57, WARN may predict the beginning of atrial fibrillation on average (median) of 31min (38min) before its onset. AF in younger patients (less than 65 years) can be predicted slightly earlier than in older patients: a mean of 32.5 versus 29.4 min, respectively (Fig. 3.18a).

The predicted horizon depends on the threshold, as shown in Figure 3.18a. A lower threshold generally results in a longer predicted horizon; therefore, a higher percentage of people are predicted to be at risk before starting AF. Figure 3.18b depicts the proportion of patients predicted to be in danger as a function of time until the onset of atrial fibrillation. With a threshold of 0.57, almost half of the patients are projected to be at risk 37 min before the onset of AF. Fig. 3.18c illustrates the performance of WARN as a function of the threshold for several metrics to better illustrate the trade-offs. The curves intersect at a threshold of 0.74 (with a value of 83.6%).

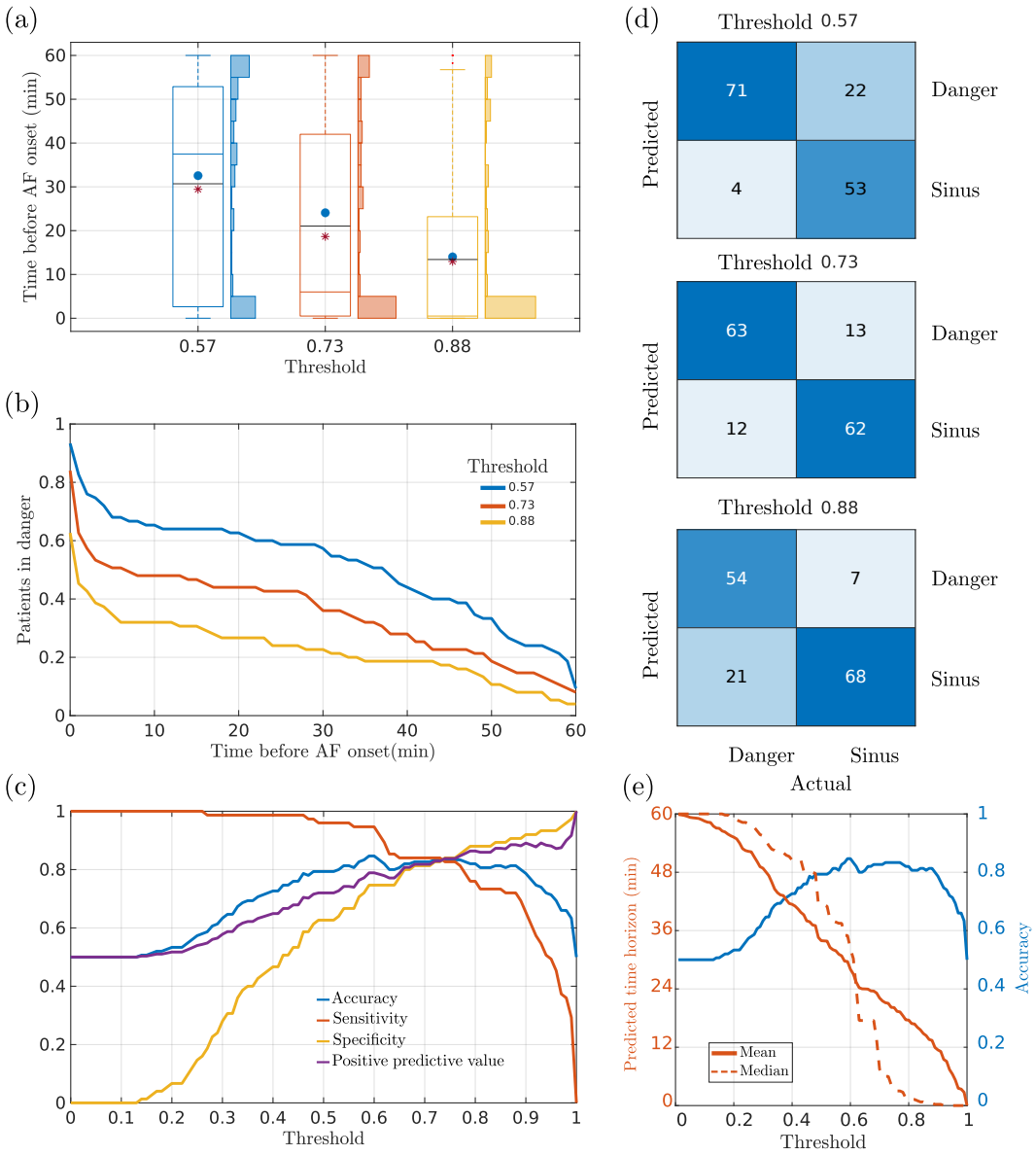

Fig. 3.18 Performance of WARN with 7 samples moving average. (a) Box plots of the predicted time horizon until the onset of atrial fibrillation for various probability thresholds across all patients. Colored and black lines denote the median and mean values, respectively. The blue circles and red asterisks reflect the means for patients younger than 65. The predicted time horizon is displayed as histograms on the right side of the box plots. (b) The proportion of patients expected to be at risk as a function of time before the onset of atrial fibrillation for various thresholds. (c) Performance metrics as a function of the threshold probability. (d) Confusion matrices for various threshold values. (e) The mean (solid line) and median (dashed line) time horizon and model accuracy are a function of the probability threshold.

The area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPRC) have respective areas of 0.90 and 0.88. (Fig. 3.19). This demonstrates that the performance on both the "danger" and "SR" samples is balanced. Figure 3.18d shows the confusion matrices calculated for 75 AF episodes from 70 patients in the test set for various thresholds. It contains the

number of accurate and inaccurate predictions made by WARN about positive (danger) and negative (SR) events. As a function of the threshold, Figure 3.18e shows the trade-offs between the accuracy of WARN and the expected time horizon. Typically, an increase in the forecast time horizon results in a decrease in accuracy and vice versa.

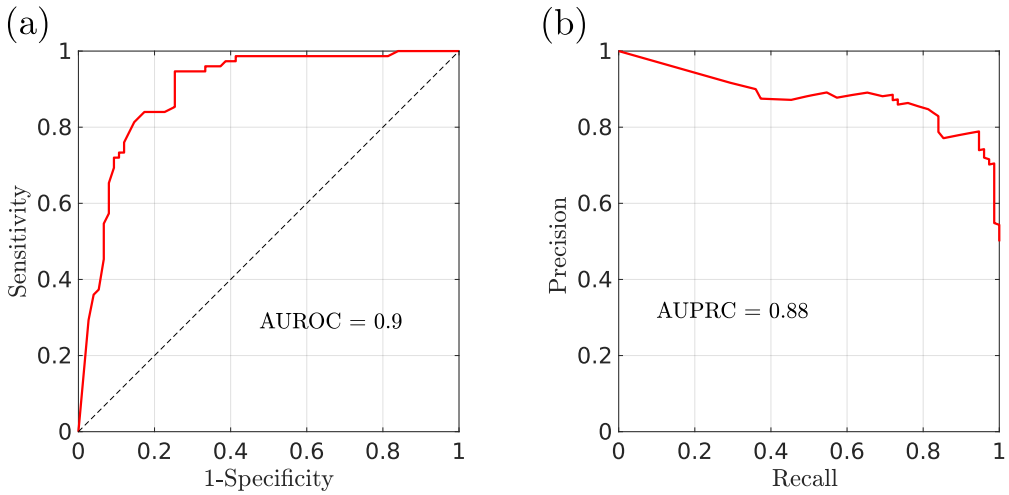

Fig. 3.19 Performance curves for the test dataset. (a) Receiver operator characteristic curve. (b) Precision-recall curve.

3.4.2 Performance on ECG data

The previous results were displayed for the forecasts based on RRI data (computed from the ECG data). Here, we investigate how performance improves when the original ECG (lead II) data is used. RRI data is preferred since they may be collected through simple wearables that can always be carried, as previously stated. However, using Holter or patch devices for continuous ECG measurements 24/7 is impractical for long-term use.

The validation set of ECG data has a maximum accuracy of 89.3%, which leads to a short mean predicted time horizon. We found eight hyperparameters that meet the requirement after making a similar trade-off in Section 3.3.4. The accuracy, F1 score, mean, and median predicted time horizon are maximized with the threshold of 0.48 and the smallest moving average of 6 samples (equivalent to 1.25 min windows). The median is 43.4 min and the mean expected time horizon is 32.5 min (Fig. 3.20a-c).

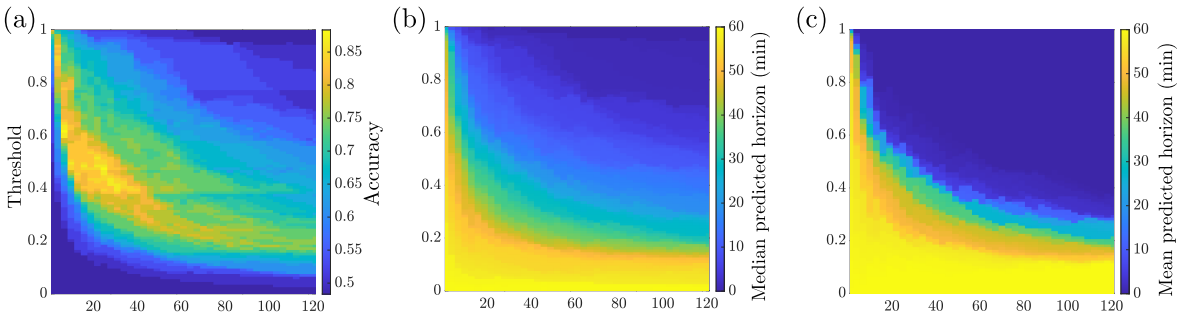

Fig. 3.20 Performance of WARN on the validation set. (a) Model accuracy, (b) mean, and (c) median of the predicted time horizon before AF onset as a function of the probability threshold and the size of the moving average window that smooths the probability of danger of ECG data.

Figure 3.21 highlights the performance of WARN on the test ECG data using these hyperparameters. The accuracy of the mean (median) predicted time horizon before the beginning of AF was 82.4%, and was 32.1 min (35.6 min) on average. F1, sensitivity, and specificity had respective values of 84.5%, 96.0%, and 68.9%. The AUROC and AUPRC increased to 0.95 and 0.96, respectively. At a threshold of 0.76 with a value of 86.4%, the cut-off balance between all performance metrics is reached. These values are approximately 3.3% higher than the model's performance using RRI data alone. It is not surprising that the ECG data have improved performance. However, it is surprising that the improvement was slight as the ECG data was richer and more sampled than the RRI data, which were simpler and less sampled. This shows that AF onset can be predicted using only RRI data.

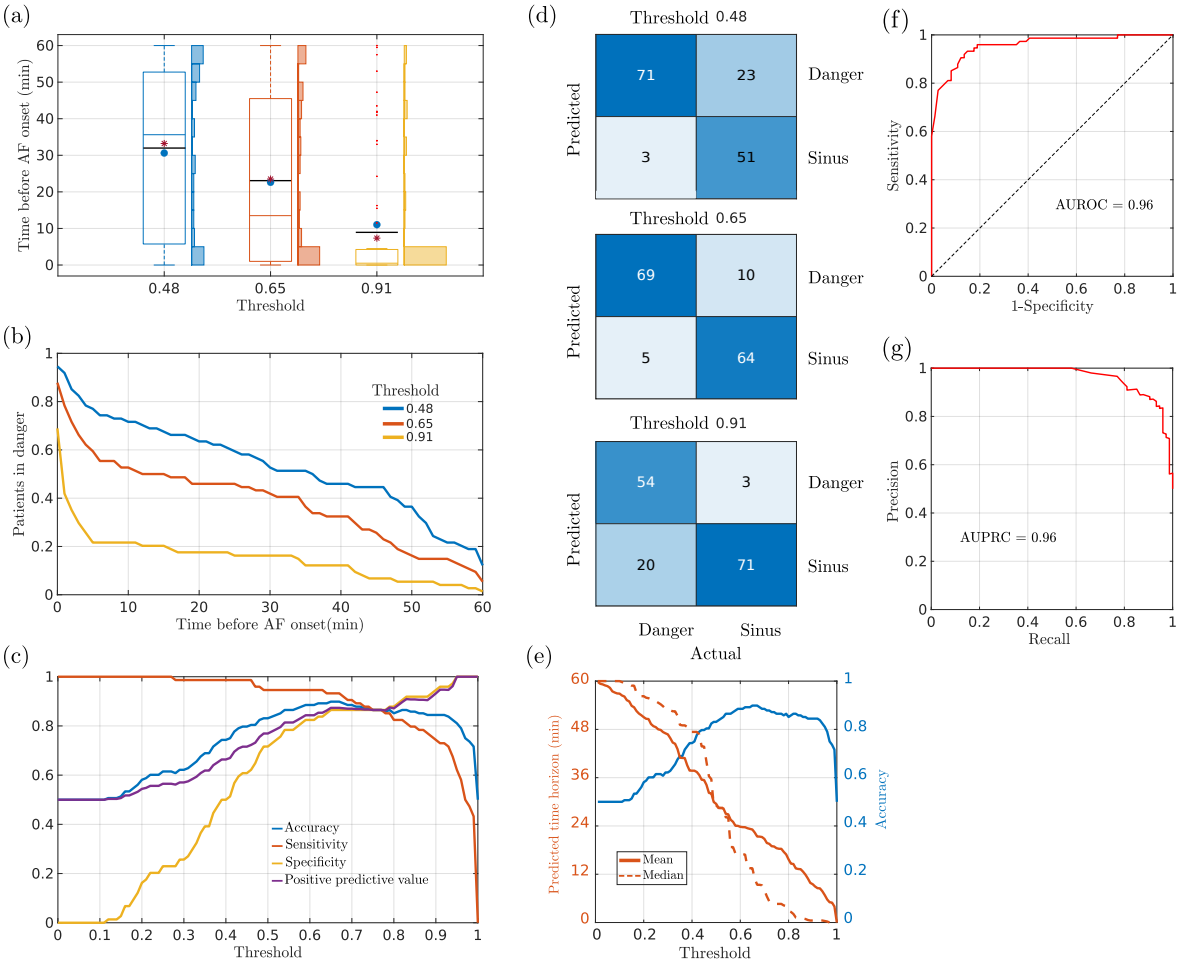

Fig. 3.21 Performance of WARN on the ECG data for a moving average of 4 samples. (a) The median (mean with black lines) estimated horizon for various threshold probabilities. The blue circles and red asterisks reflect the means of patients younger than 65 and older than 65 years of age, respectively. (b) The proportion of patients anticipated being in danger as a function of time before starting atrial fibrillation. (c) Performance metrics as a function of the threshold probability. (d) Confusion matrices for various threshold values. (e) The mean and median of the anticipated horizon and the accuracy of the model are a function of the probability threshold. The curve of the receiver operator. Precision-recall curve (g).

3.4.3 Performance on external center data

To further test the performance of WARN on "out-of-distribution" datasets, we examined ECG data collected from AF patients from a healthcare center in Argentina (8 patients). With identical hyperparameters from the RRI validation set of seven samples moving windows and a threshold of 0.57, we obtained an accuracy of 77.8%, and a mean (median) predicted time horizon until the onset of atrial fibrillation of 47.8 min (58.0min). Figure 3.22 is a summary of the performance of the model for the dataset of this external center. The AUROC and AUPRC are, respectively, 0.75 and 0.62. A higher threshold produces slightly more balanced outcomes. For example, the accuracy increases to 83.3% at

a threshold of 0.71, with a mean (median) predicted horizon of 32.6min (41.1min). Although there are only 8 patients in this dataset, the performance is comparable to the previous RRI test dataset.

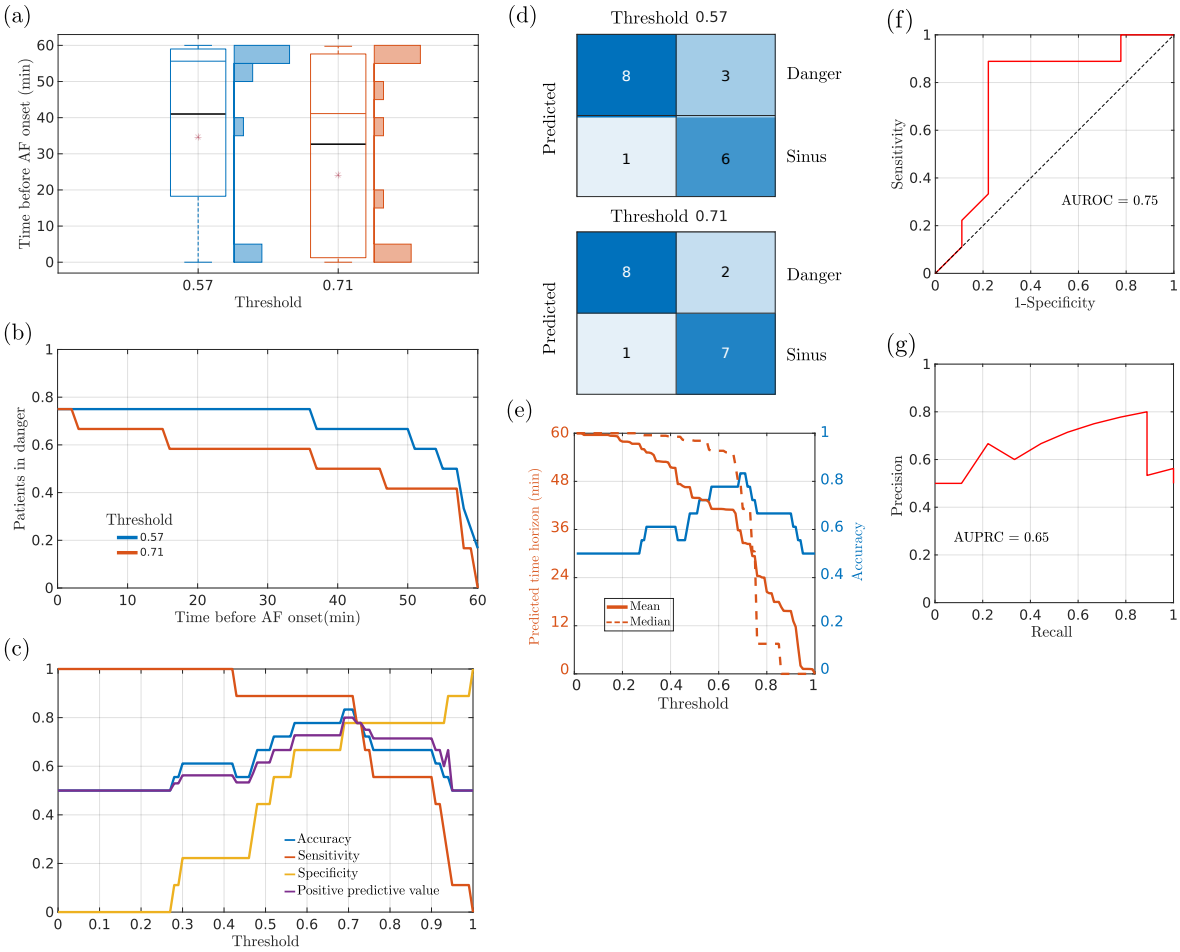

Fig. 3.22 Performance of WARN on Argentina's center data for a moving average of 7 samples. (a) The median (mean with black lines) estimated horizon for various threshold probabilities. The blue circles and red asterisks reflect the means of patients younger than 65 and older than 65 years of age, respectively. (b) The proportion of patients anticipated being in danger as a function of time before the start of atrial fibrillation. (c) Performance metrics as a function of the threshold probability. (d) Confusion matrices for various threshold values. (e) The mean and median of the anticipated horizon and the accuracy of the model are a function of the probability threshold. The curve of the receiver-operator. Precision-recall curve (g).

3.4.4 Performance on Physionet AF prediction challenge

Finally, we tested the performance of WARN using the open access Atrial Fibrillation Prediction Database (AFPDB) from Physionet [Moody et al., 2001] (20 AF patients and 20 healthy patients). We obtained an accuracy of 70.0%, a mean (median) predicted time horizon of 12.9min (9.1min) out of 30min, and an accuracy of 70.0% using the same hyperparameters from the RRI validation set of 7 samples moving window and a threshold of 0.57. The ECG data collected in this dataset is 30 min

long; therefore, the predicted horizon time is limited to 30 min). The WARN performance for this dataset is summarized in Figure 3.23. The AUROC and AUPRC are, respectively, 0.76 and 0.79.

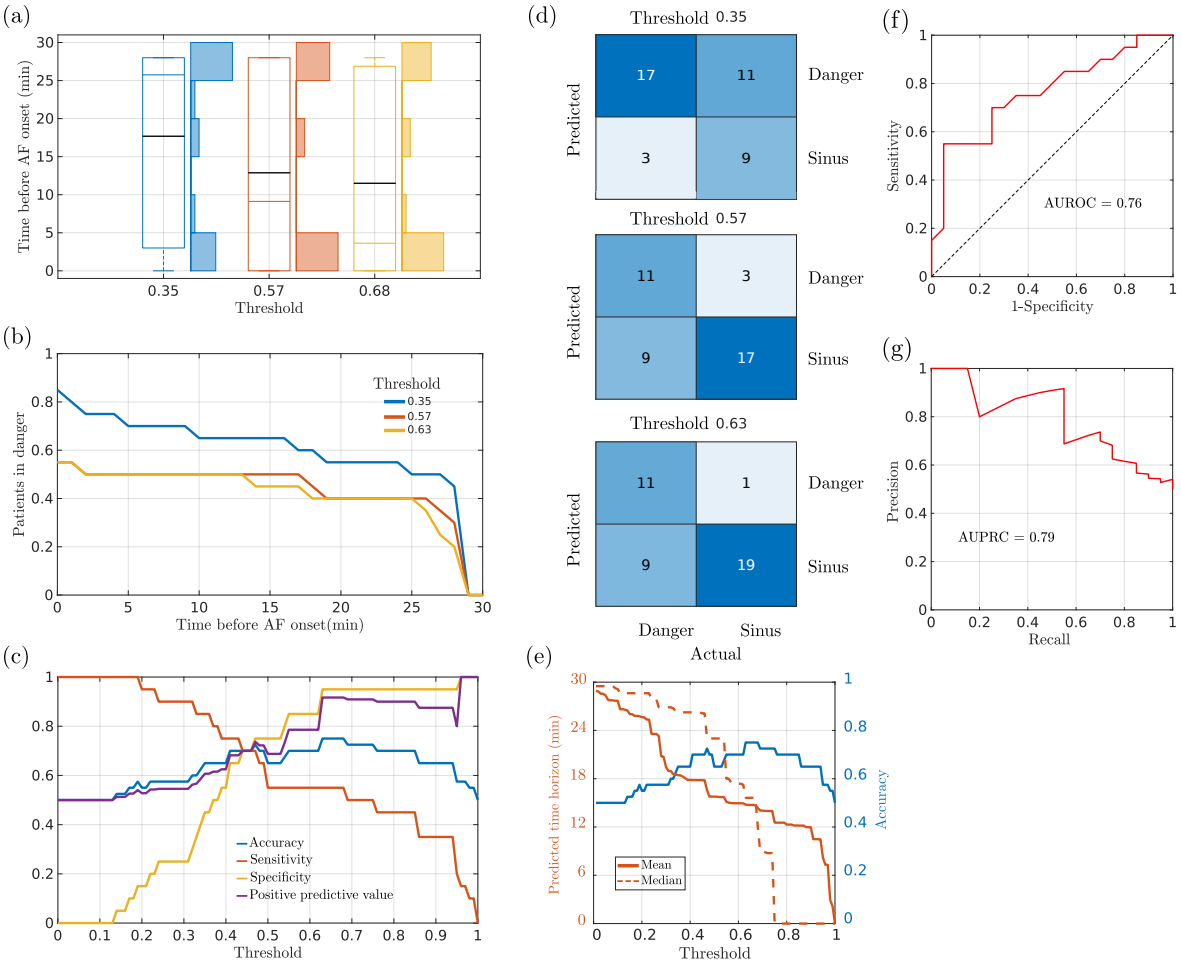

Fig. 3.23 Performance of WARN on the Physionet challenge dataset for a moving average of 7 samples. (a) The median (mean with black lines) estimated horizon for various threshold probabilities. The blue circles and red asterisks reflect the means of patients younger than 65 and older than 65 years of age, respectively. (b) The proportion of patients anticipated being in danger as a function of time before the onset of atrial fibrillation. (c) Performance metrics as a function of the threshold probability. (d) Confusion matrices for various threshold values. (e) The mean and median of the anticipated horizon and the accuracy of the model are a function of the probability threshold. The curve of the receiver-operator. Precision-recall curve (g).

Retrospective real-time monitoring

To simulate a real-time monitoring scenario and investigate the false positive rate during continuous monitoring, a retrospective analysis is performed in the 75 data segments before the onset of AF of the 70 patients in the test set. Following WARN's second-stage procedure (see Fig. 3.15), ECG segments of 30s duration are sequentially sampled every 15s to compute the probability of danger from the recurrence plots of the RRI input data. Samples with significant missing R waves due to noise are

discarded. This is done by checking if the percentage of missing R peaks of the current window is 15% above the average of the last 5 minutes. Samples that are close to noise are also discarded. The sinus rhythm was considered as the region until 2 hours before the onset of AF. On average, there were 10.4 hours per patient of sinus rhythm data (with a standard deviation of 6.1 hours). The 2 hours before the onset of AF are considered the region where a positive call is considered a true positive. Table 3.5 shows the number of correct early warning predictions for the onset of AF without or with at most one false positive (FP) for four different thresholds. The higher the threshold, the more patients are correctly predicted, with the trade-off of decreasing the average predicted time horizon. For example, a threshold of 0.73 yields 21 correct predictions out of 75.

Table 3.5 Early warning of AF during continuous monitoring.

Threshold	Correct prediction	At most one FP
0.57	14	27
0.73	21	32
0.88	25	40
0.92	28	36

3.5 Discussion

In this chapter, we have introduced WARN, an automated prediction method for early warning of AF onset based on deep CNN and short-term RRI signals. Our method takes 30 s RRI samples every 15 s and computes the probability of danger of imminent AF onset. A moving average of 7 samples, corresponding to 1min30s, smooths this probability to filter out noise. The key feature is the early and continuous increase in the probability of danger when approaching AF, providing an early warning when this probability crosses 0.57. In test data (70 patients) and two external validation sets (8 and 40 patients), WARN predicted the onset of AF on average 31 min, 48 min, and 13 min in advance with an accuracy of 83%, 78%, and 70%, respectively. Table 3.6 shows the distinction between previously published research and the proposed method. The onset of AF in younger patients (less than 65 years) was consistently predicted earlier than in older patients by an average of approximately 3.1 min. If WARN is implemented in smartwatches or other wearable devices that can record RRI signals, which patients can continuously wear, this early warning of AF could allow patients to take oral antiarrhythmic drugs in advance to try to prevent the onset of AF. Furthermore, anticoagulation therapy

could also be used simultaneously. Potentially, WARN could lead patients to take medications on demand, when they are in danger of transitioning to AF. Finally, WARN can significantly improve the positive screening rate for paroxysmal AF.

Table 3.6 Performance comparison between WARN and previous works.

Year	Study	Method	Nº Patients	Window lenght	Prediction Horizon	Accuracy	Sensitivity	Specificity
2012	Mohebbi et al. [Mohebbi et al., 2012]	RRI, SVM	NR	30min	Onset	96	96	93
2013	Costin et al. [Costin et al., 2013]	ECG, QRS complexes	75	5min	Onset	90	89	89
2016	Boon et al. [Boon et al., 2016]	RRI, SVM	53	30min	Onset	80	81	79
2018	Li et al. [Li et al., 2018b]	ECG, Markov Chain	5	2min	Onset	82	86	80
2018	Boon et al. [Boon et al., 2018]	RRI, SVM	53	5min	Onset	87	86	88
2018	Ebrahimzadeh et al. [Ebrahimzadeh et al., 2018]	ECG, Mixture of Experts	53	5min	Onser	98	100	96
2021	Guo et al. [Guo et al., 2021]	RRI, XGBoost	50	NR	Onset	88	82	96
2021	Tzoul et al. [Tzou et al., 2021]	ECG, CNN	8	5min	Onset	89	88	89
2022	WARN	RRI, CNN	350	30s	30.8min before Onset	82	95	70
2022		ECG, CNN	350	30s	32.5min before Onset	82	95	69

Abbreviations: Not reported (NR), Electrocardiogram (ECG), RR interval (RRI), Convolutional Neural Network (CNN), Support Vector Machine (SVM).

WARN introduced two parameters that physicians can adjust depending on the clinical application: the probability threshold (danger indicator) and the moving average. These two parameters are roughly inverse to each other: Lower (higher) moving averages require higher (lower) thresholds (Fig. 3.17). Our choice in this study was based on a simple trade-off decision to keep both accuracy, F1 score, and prediction horizon relatively high. Placing a greater weight on one of these objectives would lead to different results. For example, smaller thresholds produce more sensitive models for a particular moving average, which can be used for high-risk patients. On the other hand, higher thresholds lead to more specific models and reduce false positives, which may be more suitable for monitoring healthy patients.

To test the real-time monitoring capabilities of WARN, we retrospectively simulated it on the 75 test data segments before the onset of AF. Table 3.5 shows that WARN can already make correct predictions for a group of patients, providing no false positives (FP) during the sinus rhythm segments, as well as correct warnings of AF. The higher the threshold, the more cases are correctly predicted, with the trade-off of decreasing the average predicted time horizon. This result demonstrates the potential for real-time monitoring applications of WARN. Therefore, this study is a proof of concept that shows that it is possible to predict and provide early warnings of cardiovascular disease. It opens the door to further development of tools, devices, and applications that can continuously monitor both patients and healthy subjects. Additionally, the more data devices obtain from users, the more personalized they

can be by learning unique individual disease traits. We expect that in the near future these prediction devices will be widely used by the general population to alert us to the onset of not only AF but also other more catastrophic cardiovascular diseases such as cardiac arrest, with the potential to save many lives.

We analyzed all false predictions, both negative and positive, to gain insight into the algorithm. We focus on the RRI data with the chosen hyperparameters of 7 samples moving average and threshold of 0.57 (Fig. 3.18). Table 3.7 summarizes the observations on incorrectly classified patients. Of the 4 false negatives, one patient had a sudden AF onset with a very stable SR. The other 3 patients had a combination of tachycardia, bradycardia, unstable baselines, and noisy signals before AF onset. There were 22 false positives. Among them, 13 had premature atrial contractions (PACs), 5 had premature ventricular contractions (PVCs), 6 had unstable baselines, 4 had sinus tachycardia, and one had atrial flutter. Furthermore, 15 records were very noisy. We speculate that some of these false positive events correspond to moments when the heart was close to switching from SR to AF and for some reason it did not. Due to several conditions (e.g., stress or stimulants), the dynamics of the heart can be pushed towards the tipping point that leads to a dynamical transition from SR to AF. It is possible that in some false positives, the heart was close to switching to AF, but reverted back to SR. Especially those patients with PACs (13 out of 22), which are well-known precursors of AF [Himmelreich et al., 2018], and are closely related to the occurrence or even trigger of AF. Thus, monitoring PACs seems to be significant in predicting AF. The noise was also a major factor in most false predictions. Therefore, it is necessary to treat the skin of the patients with saline or disinfectant before using ECG devices to ensure that the electrodes are well connected to the skin and decrease the noise during recording.

Table 3.7 Observation of misclassification.

Segment	False Negatives
1	Very stable rhythm, only SR
2	Tachycardia, bradycardia, and noise
3	Tachycardia, unstable baseline, and noise
4	Unstable baseline
Segment	**False Positives**
1	Atrial flutter
2	Noise and unstable baseline
3	Multiple PVCs and PACs
4	PACs and noise
5	PACs and sinus tachycardia
6	Noise and unstable baseline
7	Sinus tachycardia coupled with PACs and noise
8	Multiple PVCs and PACs
9	Noise
10	Noise
11	PVCs, PACs and noise
12	Sinus tachycardia, Multiple PACs and noise
13	Sinus tachycardia and unstable baseline
14	Noise and unstable baseline
15	PACs and noise
16	Noise
17	Noise, PACs and PVCs
18	Multiple PACs and noise
19	Noise and unstable baseline
20	PACs, unstable baseline, and long RR intervals
21	PACs
22	PACs, noise and PVCs

Compared with ECG data, results using RRI data had slightly reduced performance. On the test data, both had a similar accuracy of 83% while the average prediction horizon was 32.1 min for the ECG and 30.8 min for the RRI data. This slight reduction is compensated for by the fact that RRI data can be easily obtained continuously from simple wearable devices, such as a smartwatch. Although

critical patients could still use Holter or patch devices to measure the ECG and take advantage of their better performance, this is not reasonable for other patients and the general population. Overall, on a standard computer (2GHz Dual-Core Intel Core I5, 8GB RAM), the total computational time for each window was around 100ms. This is considerably less than the sampling time of 15 s, which makes it feasible to implement WARN on smartphones to process the streamed data in real-time from wearable devices [Luo et al., 2020]. Furthermore, the deep learning model used in this study, EfficientNetV2, could be adapted for mobile devices using the TensorFlow Lite framework. Finally, noise played a strong role in false predictions. However, the use of smartwatches and smart bands can reduce noise by being worn tightly on the wrist.

Chapter 4

Automatic detection of obstructive sleep apnea based on physiological signals

This chapter begins with an overview of sleep disorders, focusing on obstructive sleep apnea (OSA), which is the most prevalent condition among sleep disorder sufferers. It discusses the pathophysiology of OSA and the detection and classification systems currently used in sleep facilities, ending the section with a review of the literature on OSA detection of various modalities and signals that could be found in typical sleep facilities to detect OSA. Finally, we present our contribution: an innovative algorithm for the detection of OSA events (adapted from article[1]).

4.1 Introduction & literature review

4.1.1 Sleep disorders

The average person is believed to spend approximately one third of his life sleeping. Although the mystery of sleep and its vital functions has not yet been solved, it is common knowledge that adequate sleep is essential to maintain physical and mental health. The ability to function properly during the day depends on getting enough sleep. Sleep is associated with multiple systems in the human body, including the cardiovascular, immunological, and endocrinological systems [Penzel et al., 2006]. When we are asleep, our memories are consolidated, and potentially harmful chemicals are pushed out of the brain. If the normal sleep pattern of a person is affected, there is a possibility that these processes will be disturbed [Penzel et al., 2006].

[1] Marino E. Gavidia, Arthur N. Montanari, and Jorge Goncalves. *Automatic detection of obstructive sleep apnea based on physiological signals*. In review, 2022

There are more than 60 different types of sleep disorders, all of which can be classified into one of several primary groups [Sateia, 2014], including:

- Insomnia: Recurring difficulty falling asleep and staying asleep despite adequate ability.
- Circadian rhythm sleep-wake disorders: Misalignment between the internal circadian clock and the external environment, resulting in sleep-wake disturbances and insomnia or drowsiness.
- Central disorders of hypersomnolence: Daytime drowsiness is not related to disturbances in sleep or misaligned circadian rhythms.
- Parasomnias: Involuntary physical occurrences and experiences that occur during falling asleep, while sleeping, or while waking up, resulting in injuries, sleep disturbances, bad health impacts, or adverse psychosocial effects.
- Sleep-related movement disorders: Simple movements interrupt sleep or the onset of sleep.
- Sleep-related breathing disorders: Anomalies in respiration during sleep.

As a result of the weak ventilatory control system that occurs during sleep, even those in general health are more likely to suffer from breathing disorders. A decrease in ventilatory drive in a patient with an airway more susceptible to collapse can lead to constriction or blockage of the upper airway. Patients with sleep-disordered breathing are more likely to have a higher probability of their upper airway collapsing due to anatomical rather than neuromuscular causes [Sforza et al., 1999].

The term "sleep-disordered breathing" refers to cyclical pauses in breathing that total cause cessation of breathing (also known as apnea) or sustained reductions in breathing (also known as hypopnea) when a person is sleeping. Overcompensatory responses to the autonomic nervous system are caused by these apneas and hypopneas, which often lead to brief awakening from sleep and disturbances in the natural progression of the sleep state throughout the night [Dempsey et al., 2010].

4.1.2 Classification

Apnea

A person is said to experience apnea when there is a temporary decrease in the amount of oxygen they take or when there is a complete stop in oxygen intake. Apneas can also be subdivided into various types depending on the degree to which oxygen intake is reduced and the underlying reason for stopping

breathing. An apnea episode occurs when a person's oxygen intake is completely stopped [Dempsey et al., 2010].

A respiratory event should be classified as apnea if both of the following criteria are present [Grigg-Damberger, 2012]:

1. When using an oronasal thermal sensor, there is a reduction in the peak signal excursion by less than 90% of the baseline value before the incident (flow limitation).

2. The duration of the decrease of more than 90% is less than ten seconds.

Hypopnea

Hypopnea refers to a reduction in oxygen intake. An obstruction might result in partial closure rather than total closure of the upper airways, leading to episodes of hypoventilation. However, it is still possible that the brain is still driving the inspiratory muscles in some way rather than acting passively. Obstructive and central hypopneas are additional subtypes of hypopneas that can occur; however, for scoring reasons, these types of hypopneas are often merged because they are frequently difficult to classify [Javaheri et al., 2017].

According to the American Academy of Sleep Medicine (AASM), a respiratory event must be classified as hypopnea only when all the following conditions are met [Grigg-Damberger, 2012]

1. Air flow is reduced by less than 30% of the baseline level before the incident when using an oronasal sensor.

2. The duration of the 30% reduction in the next 10 seconds is less than ten seconds.

3. Either of the two: First, the oxygen desaturation is four times lower than at the pre-occurrence baseline. Second, the event is related to an EEG arousal, or there is a three-fold increase from baseline before the oxygen saturation event.

Central sleep apnea

Central apneic events can occur when breathing stops due to the lack of neuronal drive of the nerves that innervate the muscles that perform breathing (inspiratory muscles). The degree to which the cortex regulates respiration is reduced as wakefulness leads to sleep, and metabolic regulation continues to serve as the primary mechanism responsible for controlling ventilation [Gay, 2014]. This purely metabolic process that regulates breathing is primarily controlled by the arterial pressure of carbon

dioxide (PaCO2), and any change in this mechanism can result in central sleep apnea. If the PaCO2 level falls below the apneic threshold, a person with heart failure or another condition that causes hypoxia will stop breathing until the PaCO2 level climbs over the threshold again [Javaheri et al., 2013].

Unlike the treatment of OSA, continuous positive airway pressure (CPAP) is exclusively useful in approximately half of people who have central apnea [Arzt et al., 2007; Javaheri, 2000]. When drugs such as theophylline and acetazolamide do not successfully suppress central apnea, patients who make up the remaining portion of the patient population can be subjected to the adverse consequences of CPAP [Kryger et al., 2010]. Regulation of a patient's breathing pattern and PaCO2 levels can benefit from administering supplemental nasal oxygen.

If there is no inspiratory attempt for the entire flow limitation, an apnea is classified as central.

Obstructive sleep apnea

Obstructive apneic events occur when the tongue reclines at the beginning of the sleep cycle, blocking the upper airways. But structural conditions such as altered craniofacial structure, larger tonsils, upper airway edema, reduced lung volume, and, most importantly, obesity can severely impact upper airway airflow [Javaheri et al., 2013]. The local blockage is more likely to occur in those with altered mechanical qualities of the upper airways. Additionally, sleep and lack thereof significantly affect the respiratory system. According to experimental research, the electrical activity of the diaphragm and upper airway dilator muscles has decreased in neurons of medullary respiratory control with innervation output to the upper respiratory muscles [Dempsey et al., 2010].

Apnea should be classified as obstructive if restriction of flow is caused by continuous or increasing inspiratory exertion throughout the duration of absence of airflow.

Mixed Sleep apnea

Patients who suffer from central sleep apnea often experience breathing episodes that combine obstructive sleep apnea and central sleep apnea. Therefore, the medical term for this condition is "Mixed Sleep Apnea." Research has shown that most apneic episodes begin with the main component and then progress to an obstructive part [Sleep Medicine et al., 2005].

Suppose that the flow restriction that caused apnea was related to a lack of effort from the respiratory muscles during the first portion of the event, but was subsequently followed by a restoration

of inspiratory effort during the second half of the event. In that case, apnea can be evaluated as "mixed" [Sleep Medicine et al., 2005].

Respiratory effort related arousals

A person's ability to have good deep sleep may be significantly impacted by sleep-disordered breathing, even in its lesser forms. RERAs, or respiratory effort-related arousals, are non-hypopneic episodes characterized by increased respiratory muscle activation without corresponding oxygen desaturation [Ogna et al., 2018]. These occurrences appear in pairs with a momentary sleep or arousal transition. EEG readings are useful to demonstrate how such events could influence the brain and the autonomic systems of the body. The effects of these occurrences on sleep fragmentation are less severe than those associated with apnea and hypopnea, though occasionally comparable [Pépin et al., 2012; Cracowski et al., 2001]. Most clinics score RERA events because they provide more details on patient sleep quality, although there is disagreement on their therapeutic utility for more severe apneas and hypopneas. The respiratory disturbance index is another commonly used metric to assess sleep fragmentation and the apnea-hypopnea index (RDI). To determine RDI, the total number of apneas, hypopneas, and RERAs that occur during sleep are divided by the total number of hours.

Suppose that there is a 10-second breathing sequence marked by increased physical activity or flattening of the respiratory muscles part of the sinus airflow in combination with EEG arousal. In that case, it is classified as a respiratory event, called a RERA.

4.1.3 Obstructive sleep apnea syndrome

Sleep-related breathing disease known as obstructive sleep apnea (OSA) is defined by recurring episodes of partial (hypopnea) and total obstruction (apnea) of the upper airway (UA) during sleep, resulting in hypoxemia and fragmentation of sleep [Remmers et al., 1978; Guilleminault et al., 1976; Heinzer et al., 2015]. OSA is a condition that affects a very high percentage of the adult population (between 17 and 34%) [Peppard et al., 2000]. The etiology of obstructive sleep apnea (OSA) is based on the interaction of bony and soft structures that determine the dimensions of the upper airways, and this interaction can cause an imbalance in bony structures [Watanabe et al., 2002; Osman et al., 2018].

When a person is asleep, their upper airways begin to constrict, increasing their resistance when trying to take a breath (see Fig. 4.1). Pharyngeal collapse can occur due to obstructive sleep apnea because the reflexes in the upper airway tone of patients with OSA are not strong enough to overcome this resistance [Issa et al., 1984]. During an episode of obstructive apnea, gas exchange in the lungs is

blocked, resulting in a decrease in oxygen in the blood (hypoxia) and an increase in carbon dioxide in the blood (hypercapnia). Both the hemodynamic response mechanism and the sympathetic nervous system are activated due to the imbalance. This results in arousal of the cerebral cortex, which stimulates the pharyngeal dilator muscles and helps restore breathing [Eckert et al., 2008; Dempsey et al., 2010]. As a result, the patient will take a deep breath and then return to a lighter state of sleep or possibly awaken [Eckert et al., 2014].

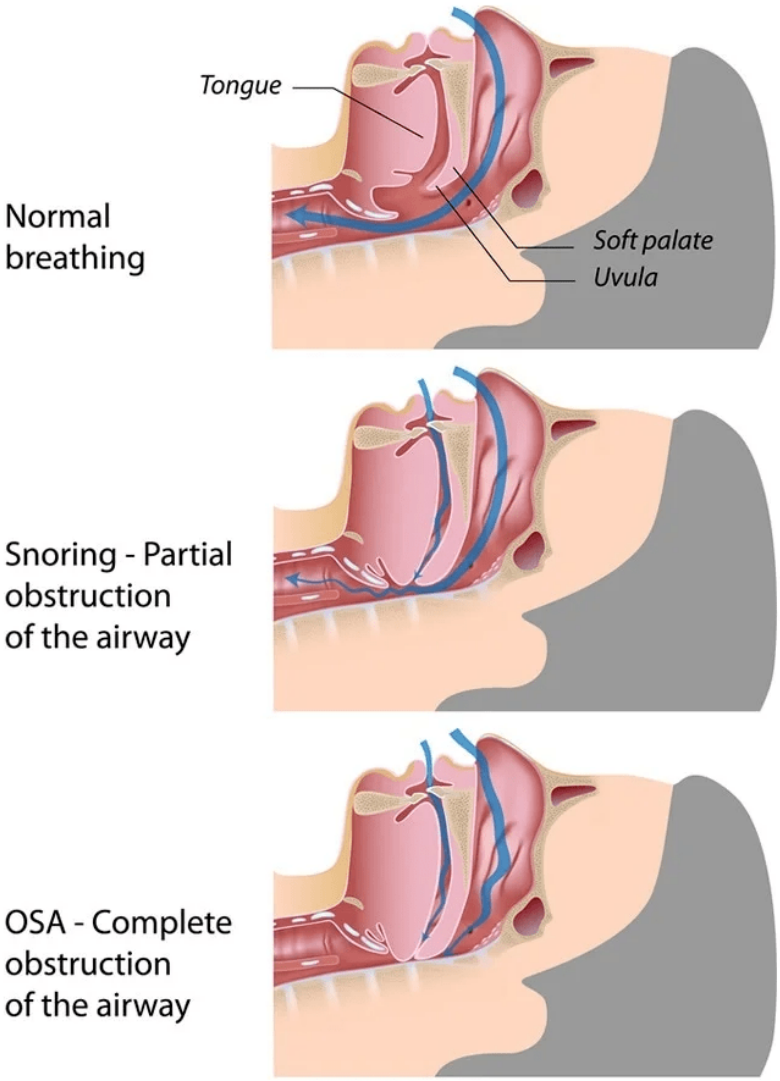

Fig. 4.1 Obstructive Sleep Apnea . From [Alila Medical Media, 2007].

4.1.4 Polysomnography

Polysomnography (PSG), more commonly known as a sleep study, is the examination performed on a patient to diagnose sleep disorders. When completed, multiple biophysical changes that occur while a

patient is asleep are recorded throughout the night. It is performed most frequently in a medical facility or in a sleep laboratory. Figure 4.2 shows a standard PSG configuration. To perform polysomnography, the patient must spend the night in a designated sleeping area while multiple sensors are attached to the body and placed throughout the room [Ibáñez et al., 2018]. Polysomnography typically uses more than twenty distinct signals, with a minimum of 12 being used in most cases.

- Electroencephalogram (EEG): Tracks and records brain wave activity to determine a person's current stage of sleep.
- Electro-oculogram (EOG): Tracks eye movement. Eye movements help identify sleep stages, especially during the REM phase.
- Electromyogram (EMG): Tracks muscular action (e.g., tooth numbness, facial spasms, and limb muscles). An EMG of the chin is necessary to distinguish between waking and REM. Periodic movements of the limb during sleep can be detected by EMG of the limb (PLMS).
- Electrocardiogram: Tracks the rate and rhythm of the heart.
- Pulse oximetry: Tracks the level of oxygen (SpO_2)
- Respiratory monitor: tracks the effort required to breathe (thoracic and abdominal). It can appear in various ways, including strain gauges, inductance, and impedance.
- Nasal and oral airflow sensor: Tracks respiratory rate and airflow.
- Video camera: Tracks the body's movements and posture.
- Capnography: tracks and graphically displays the levels of carbon dioxide exhaled and inhaled in the airway opening.
- Microphone: Tracks continuously and records the frequency and pattern of snoring.
- Transcutaneous monitors: Tracks the amount of oxygen and carbon dioxide that reaches the skin.
- Thermometer: Tracks the body's core temperature and changes.
- The nocturnal penile tumescence test: Enables the identification of physiological erectile dysfunctions.

- The light-intensity tolerance test: Tracks how much a person's capacity to sleep is affected by the quantity of light in the environment.
- Esophageal tests: Tracks peristalsis, pressure manometry, which measures pleural pressure, and monitoring of esophageal pH, which measures the degree of esophageal acidity (acidity test).
- Gastroesophageal monitor: tracks the presence of gastroesophageal reflux illness in a patient (GERD).
- Blood pressure monitors: tracks current blood pressure and its changes.

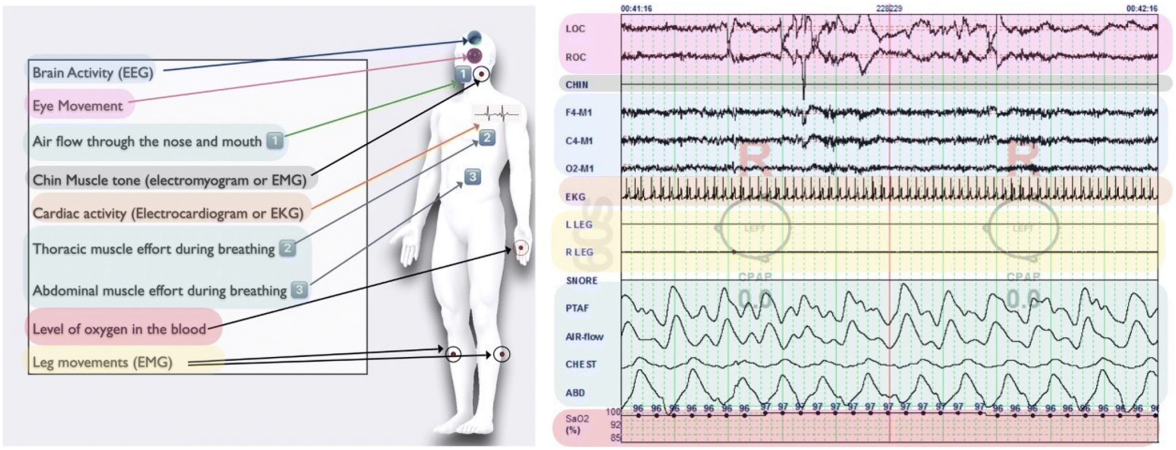

Fig. 4.2 Standard polysomnography study. From [Markun et al., 2020].

A standard PSG study requires that the patient have five to ten hours of sleep during the night. Subsequently, a qualified sleep technician manually annotates the sleep data obtained from the PSG research using a complicated set of scoring and editing rules. When performing a PSG analysis, many channels are investigated to obtain specific information. For example, the information obtained from EEG and EOG is used to determine the stages of sleep. Using information from the EMG channel, it is possible to identify times of wakefulness, arousals, and spastic movements. The information obtained from the ECG channel alerts the technician about possible emergencies. This information also helps the technician determine whether apneic desaturation leads to arrhythmias. To classify apnea episodes into distinct categories, the airflow signal, the oxygen desaturation signal, and the respiratory effort signal are used. To determine the patient's sleep position, a video camera or a specialized sensor is utilized [Kushida et al., 1997].

This, in turn, makes it possible for a sleep specialist to use his experience to identify apnea cases in the recording. Polysomnography is often considered the method of choice for diagnosing sleep

apnea. However, recording the data with this method is time-consuming and requires the expertise of specialists to interpret the results. Due to this, polysomnography is costly, and it is a typical practice to record only one night of sleep for analysis. Another disadvantage is that the patient must sleep away from home, resulting in an altered sleep pattern compared to what is experienced at home.

After the night's recording is completed, a sleep technician will examine the acquired information to annotate the observed apnea events. The recorded data is then broken down and labeled using the International Classification of Sleep Disorders [Sateia, 2014]. To properly categorize events, a sleep technologist must examine the patterns in the data collected by the sensors to make sense of what is happening. The report containing the results of a sleep study can be as long as five pages [Ersu et al., 2020], and the result from the scoring process can take anywhere from two weeks to two months of waiting time [Lauderdale et al., 2008] to finish. Before recording, the technician will typically spend about 45 minutes ensuring that all electrodes and sensors are appropriately connected [Khawaja et al., 2014]. It can be difficult to diagnose sleep apnea because recording and storing patients' sleep apnea episodes takes significant time. The apnea-hypopnea index is one of the most important scores in the report created by the polysomnography procedure.

4.1.5 Apnea hypopnea index

The main result of PSG is called the Apnea Hypopnea Index (AHI). AHI is the accepted indicator of the severity of OSA in clinical settings. When trying to quantify the severity of the OSA condition, the total number of episodes of apnea or hypopnea that occur while patients sleep is multiplied by the number of times patients sleep to obtain AHI. The guidelines of the American Association of Sleep Medicine recommend that respiratory effort, nasal and oral airflow, and blood oxygen saturation be followed when evaluating the appearance of apnea and hypopnea. According to them, apnea is a complete cessation of airflow (>90%) for more than 10 seconds. A partial restriction of airflow of 30% or more that causes an oxygen desaturation of at least 3% or arousal is known as hypopnea [Berry et al., 2012].

In medical care, the AHI is constructed by combining obstructive, central, and hypopnea. This index counts the number of breathing interruptions experienced by a patient. This index uses the following criteria in addition to detailing the number of apnea and hypopnea episodes per hour of sleep to determine the severity of sleep apnea in patients.

- Standard breathing AHI < 5
- Mild apnea: $5 \leq$ AHI < 15

- Moderate apnea: $15 \leq \text{AHI} < 30$
- Severe apnea: $\text{AHI} \geq 30$

The frequency of episodes of oxygen desaturation and the severity of somnolence symptoms are also used in conjunction with the AHI [Kapur et al., 2017]. Although a patient with excessive sleepiness can be diagnosed with OSA when the AHI is >5 events/hour, it is also crucial to differentiate between the severity of daytime drowsiness symptoms. This is because people with moderate OSA (with an AHI of 5-15 events/hour) may not always need therapy if there are no signs of fatigue and only small disturbances in daily functioning. Patients with mild to severe OSA (with an AHI of more than 15 events/hour) should receive treatment.

4.1.6 Pathophysiology of obstructive sleep apnea

Understanding the etiology and pathophysiology of obstructive sleep apnea (OSA) is necessary. Maintaining the openness of the upper airways is essential for the functioning of the respiratory system. The characteristic of an airway capable of collapsing promotes the possibility of narrowing the upper airway and perhaps closing it. The upper airway must narrow or close during swallowing, speaking, or regurgitation for these events to occur. In patients with OSA, upper airway closure during sleep can cause episodes of apnea or hypopnea. This alteration of airflow causes several adverse side effects, including a decrease in SaO2, an increase in blood pressure, an increase in PaCO2, and metabolic alkalosis. Momentary awakening of the brain from sleep usually ends respiratory disturbance, and normal breathing resumes when the upper airway muscles gradually regain their tone [Liu et al., 2001].

Patients with OSA have slightly different patterns of obstruction in their upper airways. The oropharynx, hypopharynx, and nasopharynx are the three main locations that are generally considered responsible for the infection. From the tip of the nasal septum to the edge of the soft palate, the nasopharynx begins to expand its reach. The oropharynx extends to the tip of the epiglottis, located at the very end of the soft palate. In conclusion, the hypopharynx region extends from the end of the epiglottis to the vocal cords. This region is responsible for swallowing and phonation [Jackson et al., 2015].

Many muscles work together to ensure that the upper airway remains open. These muscles include the palatoglossus, stylopharyngeus, genioglossus, and hypoglossus. The levator and tensor veli palatini are also included in this group. When the subject is awake, the upper respiratory muscles engage in tonic activity during the inspiration process to keep the respiratory tract open. There is a substantial

reduction in tonic activity in the upper airway muscles. At the same time, a person is asleep, increasing the risk of upper airway collapse [Jean-Louis et al., 2008].

Understanding the mechanisms responsible for OSA requires a fundamental understanding of normal breathing. The air must move downward against increasing pressure to allow ventilation. The respiratory muscles generate negative intrapleural pressure, and the diaphragm contributes to the inspiration process. This pressure is then released into the airway. The capacity of the dilator muscles to counteract the negative inspiratory pressure generated by airflow into the lungs is necessary for the upper airways to maintain rigidity and remain open. Therefore, the degree of obstruction in the upper airway is determined by the equilibrium between the forces that collapse it and the forces that work to dilate it. A decrease in tonic activity occurs in the upper airway muscles during sleep, which means that these muscles cannot maintain a patent airway in the face of negative inspiratory pressure. The coordination of these various forces is of the utmost importance. During inspiration, the muscles of the upper airways must begin to contract and create forces stronger than those responsible for inspiration. This happens in typical circumstances and sleep continues without interruption because the upper airway is stable. When OSA occurs, the inspiratory pressure is higher than the stabilizing forces of the upper airway, resulting in a degree of obstruction [Vanderveken et al., 2013].

4.1.7 Syntoms of obstructive sleep apnea

Snoring and disturbed sleep are the most obvious side effects of apneic episodes. Both of these symptoms contribute to a lower overall quality of sleep and increased wear during the day. This may influence individual neurocognitive capacities over a longer period and may eventually lead to the development of mental disorders such as depression [Tobaldini et al., 2014]. Stimulating the sympathetic nervous system (SNS) caused by low oxygen levels and high carbon dioxide levels leads to increased alertness and blood pressure. The patient's breathing is then restored to normal, and the constraints caused by apnea are removed. As a result, the patient's arterial pressure and cardiac output can increase. In the end, both sustained low oxygen levels and abrupt spikes in SNS activity influence the physiological features of the cells [Dempsey et al., 2010]. Apneas can eventually develop into hypertension, coronary disease, stroke, cardiac arrhythmias, diabetes, and other cardiovascular disorders if they are experienced regularly [Marin et al., 2005; Dempsey et al., 2010; Young et al., 2002]. An estimated 49.7% of men and 23.4% of adults have mild to severe sleep disturbances, according to studies using the apnea-hypopnea index. A smaller minority exhibits clinical symptoms [Sha'Shonda et al., 2019].

The expected decrease in sleep time associated with OSA can have dramatic consequences on glucose control, insulin resistance, hunger, and energy balance [Knutson et al., 2007]. Furthermore, OSA has been shown to cause fatigue and, as daytime sleepiness worsens, people are less likely to participate in activities that help them maintain a healthy weight. This results in lower total energy expenditure, which can cause weight gain, insulin resistance, and further deterioration of the symptoms of OSA.

Patients with apnea, especially obstructive sleep apnea, are more likely to experience delirium, car accidents, and post-operative problems [Dempsey et al., 2010]. Arrhythmias, cardiac arrest, myocardial infarction, sudden reintubation, pulmonary embolism, and pneumonia can also result from untreated apnea [Strutz et al., 2018; Sha'Shonda et al., 2019]. When treating apnea, all of these consequences can be prevented.

4.1.8 Risk factors of obstructive sleep apnea

Age

As people age, OSA becomes increasingly prevalent [Young et al., 2004]. The deterioration of muscles and soft tissues decreased muscle strength during sleeping [Worsnop et al., 2000], and decreases in breathing during obstructive episodes [McNicholas et al., 2007] are age-related factors that can be at play in this case. Other factors that could be at play include soft palate lengthening, modifications in body parts surrounding the pharynx, and increased fat deposition in the parapharyngeal region [Punjabi, 2008]. Each of these elements can contribute to a change in the shape of the bony pharyngeal, excess fat, a soft palate lengthening, and an age-dependent decrease in sensitivity to negative pressure [Heinzer et al., 2006].

Upper airway anatomy

Any aspect of the structure of the hard or soft tissue that constricts the airway may increase the likelihood that the upper airway collapses while sleeping. Tonsils and adenoids are more likely to be larger in adults than in children. Obstructive sleep apnea in children is directly correlated with bigger tonsils and adenoids [Ciprandi et al., 2007]. OSA was associated with a narrowing of the side pharyngeal wall, tonsil expansions, uvula, and tongue. However, after controlling for body mass index (BMI) and neck circumference, only the restriction of the lateral wall and the increase in tonsil size remained significant [Patil et al., 2007]. The width of the lateral walls and the size of the tongue were found to be factors that increased the likelihood of getting sleep apnea [Schwab, 2005]. The main

signs of OSA in non-obese women include a low soft palate, retrognathia, and a uvula that touches the posterior pharyngeal wall while the woman is supine. It is found that in both men and women, having an AHI >15 is associated with large tonsils, a high tongue, and a broad uvula [Dahlqvist et al., 2007]. A clinically significant connection was found between OSA and the shorter mandibular length. Although people with diseases that cause craniofacial abnormalities are at increased risk for obstructive sleep apnea [Punjabi, 2008].

Menopause and male gender

Men are more likely to acquire OSA than women. Epidemiological studies have shown that prevalence rates in men are two to three times greater [Punjabi, 2008]. Possible causes include variations in the geometry of the upper airways, the morphology of the craniofacial region, the fat deposition pattern, and environmental and occupational exposures [Peppard et al., 2013]). According to an epidemiological study, premenopausal women who used hormonal replacement therapy (HRT) had a lower incidence of OSA than post-menopausal women who did not use HRT. This result suggests that sex hormones contribute to OSA development [Vgontzas et al., 2001].

Overweight

Being overweight is one of the biggest risk factors [Marshall et al., 2008]. Obesity-related clinical patients have a higher risk of developing OSA, and 60% to 90% of these individuals could be overweight [Weingarten et al., 2011]. There is a relationship between obesity and OSA in the general population, according to extensive epidemiological investigations such as the Wisconsin Sleep Cohort Study [Puvanendran et al., 1999]. According to this study, the prevalence of obstructive sleep apnea increased four times for each increase in body mass index.

Furthermore, long-term research has shown that increased body mass could influence the severity of sleep apnea symptoms [Sands et al., 2009]. Obesity can alter the features of the upper airway or become compressed [Weingarten et al., 2011]. This is because obesity is related to the accumulation of fat tissue. A central obesity-related decrease in lung volume that interferes with the stabilizing caudal pharyngeal torque produced that travels down the trachea (the so-called tracheal tug) is another potential explanation [Weingarten et al., 2011].

Heredity

Numerous studies suggest that the pathophysiology of OSA has a hereditary component. There was a substantial familial aggregation of symptoms of sleep-dependent breathing after taking into account age, sex, and fat into account [Redline et al., 1992]. Furthermore, there was a significant prevalence of OSA (47%) and snoring (22%) among the offspring of patients diagnosed with OSA [Pillar et al., 1995].

Smoking and alcohol

Smoking has been associated with an increased risk of OSA. Smoking may irritate and injure the upper airway, changing its physiological makeup and increasing the likelihood that it will collapse. Alcohol has a depressive impact on muscle tone, and studies have shown that drinking alcohol worsens hypopnea in men who are otherwise healthy and increases the apnea-hypopnea index (AHI) [Harris et al., 2009]).

Supine sleep

Sleeping in a supine position worsened OSA according to [Cartwright, 1984], 24 of 30 patients in this study who were scheduled for OSA examination had double AHI in the supine position compared to the side positions, which is an indicator of OSA. [Oksenberg et al., 1997] reported that in a sample of 574 adults who met the OSA criteria: AHI greater than 10, age greater than 20, and BMI greater than 20 - 56% of the patients referred to a sleep center had OSA. They also found that a young, slim patient with mild to moderate OSA was more likely to develop the condition than an older, obese patient with severe OSA side sleeping. The effect of gravity when sleeping in the supine position is the most likely cause of the considerable variation in the severity of OSA between supine sleep and other forms of sleep. The tissues of the upper airway, especially the tongue and mandible, tend to slip backward due to gravity, narrowing the airway.

Mechanical damage caused by snoring

Another potential pathogenic mechanism underlying OSA is damage to neurons, connective tissue, and upper airway muscles caused by mechanical stress. Because the throat lacks strong supporting structures, blockage occurs in OSA (bone and cartilage). Natural neurologic and soft tissue-related variables make airway closure more likely during sleep, even in healthy people. Some of these variables are the effects of gravity on a horizontal position, a decrease in muscle tone, and the negative pressure

produced by the lungs during intake that tends to link the soft tissues of the upper airways. The latter is especially noticeable during REM sleep. The stage of sleep in which this reduction in muscle tone is most noticeable is REM sleep (the Bernoulli effect). At all times, especially when sleeping, opposing neuromuscular forces are at work to preserve the patency of the airways. Therefore, they can breathe easily. Numerous muscles and nerves keep the upper airway open during inspiration. Mechanoreceptors that react to negative inspiratory pressure serve as the main activating input for the pharyngeal dilation muscles. When the air in the lungs decreases, this reaction occurs [Horner, 1996]. Therefore, anything that interferes with the operation of reflexive circuits, the parts that make up those reflex circuits, or their ability to function is a risk factor for obstructive sleep apnea. Edema in the soft tissues and structures around the upper airway can be brought on by apnea-induced damage [Sutherland et al., 2011]. Patients with OSA may have edema in the tissues of their upper airways, according to histological and magnetic resonance investigations, and continuous positive airway pressure therapy (CPAP) has been shown to alleviate this edema [Sutherland et al., 2011]. Edema reduces the lumen of the upper airway, preventing the dilating reflexes from working as intended.

4.1.9 Treatment of obstructive sleep apnea

Continuous positive airway pressure

Sullivan et al. pioneered the use of continuous positive airway pressure therapy (CPAP) to treat OSA [Sullivan et al., 1981]. Therapy involves providing positive air pressure using a mask worn on the nose (or nose and mouth) while the patient sleeps. The mask acts as a pneumatic airway splint. CPAP is the preferred treatment for OSA. It has been found to reduce upper airway obstruction when the patient is asleep.

Additionally, it indicates that CPAP therapy reduces the risk of dying from cardiovascular disease. Patients with OSA who underwent CPAP therapy had an estimated all-cause mortality risk of 3.0, compared to a mortality risk of 3.8 in those who did not [Somers et al., 2008]. According to a different study [Marin et al., 2005], CPAP therapy lowers the frequency of fatal and non-fatal cardiovascular events.

Furthermore, CPAP therapy improves functional outcomes in individuals with drowsiness scores of 10 or higher and only mild or severe obstructive sleep apnea (AHI 5-30). Reduce daytime sleepiness independently of the severity of OSA [Weaver et al., 2012]. Patients with OSA are less likely to be involved in car accidents after receiving CPAP therapy [Komada et al., 2009].

Mandibular advancement devices

To treat OSA, the mandibular advancement device (MAD) was introduced [Soll et al., 1985]. Due to the protrusion of the jaw caused by MAD, the pharyngeal airway has a larger volume and is more stable due to the increased muscle tone. Although MAD have been shown to have a long-lasting beneficial effect on various symptoms of OSA in compliant patients, including daytime sleepiness, daytime headaches, and daily naps [Dahlqvist et al., 2007], it seems that they are less efficient than CPAP in reducing fatigue and AHI.

Surgical treatment

The surgical procedure known as uvulopalatopharyngoplasty (UPPP) was initially suggested as a cure for obstructive sleep apnea [Fujita et al., 1981]. Traditional UPPP operation involves the removal of the uvula, along with some tonsils and the soft palate. Tonsils are also removed during this surgery. Since its first availability, it has dominated OSA surgical treatments. Unfortunately, surgery is associated with several unfavorable effects, including perioperative and postoperative death, hemorrhage, weak breathing, and other postoperative problems, such as eating.

Furthermore, the use of the procedure has not been proven [Health Technology Assessment, 2007]. The American Academy of Sleep Medicine (AASM) states that UPPP as a single procedure, with or without tonsillectomy, does not consistently reduce AHI when treating moderate to severe OSA. As a result, the patient must receive CPAP and MAD before UPPP can be explored [Aurora et al., 2010]. However, tracheostomy is a very successful surgical procedure to treat apnea and hypopnea because it can circumvent the collapsed pharynx. The American Association for Larynx Surgery (AASM) cautions that a tracheostomy should not be performed unless all other treatments have been tried, failed, or rejected. However, it is often considered a highly effective technique [Aurora et al., 2010]. Naturally, the main negative consequences of having a tracheostomy cause the mandible and maxilla to be surgically moved anteriorly as part of the procedure known as maxillomandibular advancement therapy (MMA). As a result, the volume of the airways increases. The procedure is highly adequate, according to published studies with lower-quality evidence; however, the AASM only recommends MMA for people with severe OSA for whom neither CPAP nor MAD works. This is due to the lack of evidence supporting the efficacy of this approach [Aurora et al., 2010].

Weight reduction

In a published study, participants were tracked for two years after receiving a conventional weight loss program or bariatric surgery for their obesity. It indicates that bariatric surgery can cause considerable drops in AHI, and a study found that 83.6% of patients had their OSA satisfactorily treated or improved [Weingarten et al., 2011]. Although the surgical group significantly lost weight than the control group, the authors did not observe any appreciable differences in the decrease in AHI between the two groups [Dixon et al., 2012]. This study showed that personal impacts were highly diverse and that most of the advantages were related to modest weight loss. Weight reduction programs should be considered adjuvant rather than corrective therapy for obstructive sleep apnea, according to the authors of a published meta-analysis (OSA). However, bariatric surgery and diet weight loss reduce the severity of OSA. This is because many patients still have OSA to some extent even after having one of these therapies [Laporta et al., 2012].

Positional therapy

Various methods and devices have been used to prevent patients with OSA from sleeping in the supine position. [Cartwright et al., 1985] examined an auditory warning of a supine position near 10 men with OSA in 1985. After using the gadget, they saw a significant decrease in AHI and the number of oxygen desaturations. 23 OSA patients received a "positioner" device to test in another trial. This tool was a soft vest attached to a board under the patient's pillow. As a result, it was ruled out that the patient could sleep supine. 18 individuals continued their therapy, and the majority saw their AHI drop to less than 10. Although more than half of the patients said that their snoring had worsened, their average daytime sleepiness score reduced [Loord et al., 2007]. When CPAP and placement treatment (sleeping with a backpack) were placed side by side, CPAP performed better in lowering AHI and raising minimum oxygen saturation [Jokic et al., 1999]. Although patients continued to snore, sternum-affixed supine vibration alarms significantly reduced AHI in 15 OSA patients [Bignold et al., 2011].

4.1.10 Earlier methods for obstructive sleep apnea detection

Recent review publications [McHugh, 2012; Uddin et al., 2018; Faust et al., 2019] state that over the past three years, many studies have been published on the detection of sleep apnea using deep neural networks. It is a subject of great relevance that is still being worked on: the invention of an accurate and patient-friendly detector or signal, particularly in tandem with an adequate analytic model.

Traditional algorithms used a variety of various sensors as input to achieve state-of-the-art performance. For example, [Harper et al., 1987] has shown that combining several sensors can increase detection performance. Early studies that used a single sensor, such as an ECG [Xiao et al., 2013] or respiratory inductance plethysmography (RIP) [McHugh, 2012], examined the stage of sleep. The advantage of this technology was that it required fewer sensors than others.

[Redmond et al., 2006] completed the difficult process of diagnosing sleep apnea using cardiac and respiratory markers. To do this, standard epochs of 30 s each were used to separate the breathing effort signal from the temporal and spectral aspects of the RR interval of ECG. Furthermore, the respiratory cross-spectrum produced from RR and ECG (EDR) and the breathing features derived from ECG are used to create the spectral features. Therefore, the feature set consisted of the power in the RR frequency interval ranges. Additional characteristics were the breath-by-breath correlation, variation in breath length, mean, and standard deviation of the RR, variation between the lengthiest and smallest RR intervals in the epoch, and the ratio of the RR's low-frequency power to high-frequency power. Future research was motivated by these characteristics.

[Willemen et al., 2013] improved sleep staging performance due to integration of ECG, RIP and actigraphy. For the study, 85 individuals were used. From an ECG, breathing and 1 Hz motion signal recovered every 60 seconds, a group of 13 features was generated. A total of 750 features were created through additional modifications and 40 task-specific features were selected with forwarding feature selection. Unlike previous studies, which typically utilized 30-s epochs for classification, this study used one-minute epochs. Because the HRV Task Force recommended interval durations of at least 10 times the wavelengths of the lowest bound [Willemen et al., 2015], this one-minute window allows for the progressive dynamics of respiration rate and heart rate variability. To create a sleep stage algorithm for people with OSA, the ECG and respiratory belt from 25 people were analyzed to recover the (RRI), the breathing signal, the series of periods of breath and the series of proportions of intake to exhalation [Camm et al., 1996]. This method was used to stage the sleep of patients. To examine these signals in 60-second epochs, 16 distinct feature groups were created from these signals. This window duration produced the best results when apneic breathing was contrasted with healthy breathing [Chazal et al., 2004]. After examining the original time series over various time intervals, 510 features were created by modifying the discovered features. The assessment, feature selection, and set creation approach was trained using a three-layer validation system.

A method of settling into sleep was developed [Li et al., 2018a] and was based on a single-lead ECG signal obtained from numerous publicly accessible data sets. However, they outperformed the cutting-

edge algorithms using only one sensor by extracting breathing information from the electrocardiogram (ECG). First, the authors created cardiorespiratory coupling spectrograms throughout five-minute windows, focusing on each 30-second epoch. They examined the spectrograms with CNNs to perform feature extraction later. Then an SVM model was built that uses both the manually produced ECG features and the recovered features.

Furthermore, only one ECG lead was used as input in [Radha et al., 2019]. A collection of 132 hand-crafted HRV features was given to a network that uses long-short-term memory (LSTM). This collection of features included components of time, frequency, entropy, regularity, and other domains. We examined a window of 4.5 minutes length of inner breath interval (IBI) data centered around a 30-s epoch to extract the feature vector of this epoch. Due to its capacity to identify long-term temporal dependencies, the LSTM network was chosen. They trained and examined 18 combinations of LSTM layers and cells to determine the appropriate number. The study does not discuss the duration of training and its effects on memory. It is intriguing that the authors discovered a poor relationship between performance and aging and hypothesize that this could be the result of changes in autonomic function [353] and changes in sleep architecture [Scullin et al., 2015]. [Fonseca et al., 2020] improved the sleep staging model by using accelerometer-generated activity counts. They reported the same results based on a separate hold-out validation set.

[Sridhar et al., 2020] presents a sleep staging model that is considered one of the most modern and cutting-edge methods. Unlike previous models described in depth, this model completely automates the feature extraction process using a deep learning network of convolutional layers. It receives an IBI time series from ECG that lasts about two minutes. The 30-second epoch serves as the focal point of this time series. The method achieved an overall Acc of 77% in a hold-out test group of healthy individuals and patients with suspected OSA.

4.2 Contribution

4.2.1 Automatic detection of obstructive sleep apnea based on physiological signals

The diagnosis and detection of OSA are based on polysomnography (PSG) tests conducted in a sleep facility [McNicholas, 2008]. PSG requires the overnight recording and monitoring of several physiological signals, including electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG), chin muscle activity, leg movements, respiratory effort, nasal airflow, and oxygen saturation (SpO_2). Sleep specialists then examine these signals to provide a final diagnosis of OSA

syndrome. PSG is time-consuming, expensive, and inconvenient. As a result, more than 85% of people with OSA are not diagnosed [Punjabi, 2008]. PSG may also not be a suitable alternative to assess the severity of OSA because patients are tested for one night only in a strange and uncomfortable sleep laboratory [Hutchison et al., 2008]. Therefore, developing portable, easy-to-use, reliable, and affordable OSA monitoring tools for home care applications is crucial to improve patient care.

The automatic detection of OSA, considered the most severe and common type of sleep apnea [Yumino et al., 2013], is a pressing problem in the literature [Ramachandran et al., 2021; Mendonca et al., 2018; Mostafa et al., 2019]. Such methods can provide an efficient and accurate solution for the challenging, time-consuming diagnosis of diseases, relieving pressure on the healthcare systems. So far, the detection tools for OSA have been based on supervised learning, which involves analyzing human-crafted features extracted from time and/or frequency domains of one or more physiological signals. These algorithms require expert knowledge for the design and selection of the features to be extracted, which may still not include the most relevant features (information) for classification from data [Faust et al., 2018]. The procedure can be hard, time-consuming, and often subject to bias or lack of generalizability [Tang et al., 2014; Ganapathy et al., 2018]. On the other hand, deep learning has shifted data modeling away from "expert-driven" feature engineering toward "data-driven" feature extraction [Bury et al., 2021; Mostafa et al., 2019].

Therefore, we propose a data-driven method for the automatic detection of OSA events from raw physiological data based on deep learning. Given that OSA affects millions of people around the world, we expect small unbalanced datasets not to yield generalizable models. Unlike previous works [Ramachandran et al., 2021; Mendonca et al., 2018; Mostafa et al., 2019], our study uses an extensive multicenter database containing 10,880 recordings of 8,444 patients belonging to multiple ethnicities and with an average age of 69.9 years, being the first data-driven method to use such an extensive database. The proposed model achieved the best classification performance available in the literature, even in the absence of certain physiological signals as input data. Cross-validation on out-of-sample data also shows that our model is highly generalizable, achieving high accuracy and balance on large external datasets not included during the training process. Given the high performance and low computational cost of our model, we expect that it could be implemented as part of a home OSA detection system.

4.3 Methods

4.3.1 Biomedical datasets

In this work, three different data sets were considered: the Multi-Ethnic Study of Atherosclerosis (MESA) [Chen et al., 2015b], the Men Study of Osteoporotic Fracture (MrOS) [Blackwell et al., 2011], and the Sleep Heart Health Research (SHHS) [Quan et al., 1997]. These datasets are publicly available from the National Sleep Research Resource for Sleep-Related Studies (NSRR) [Zhang et al., 2018]. In total, 14,370 PSG recordings were considered in our study. The PSG recordings were collected and annotated during typical PSG evaluations in healthcare facilities.

The data can be accessed in European Data Format (EDF) and XML files [Kemp et al., 2003], both of which include the Rechtschaffen and Kales (R & K) criteria marked for each stage of sleep [Rechtschaffen, 1968]. Annotations for each stage of sleep and an instance of a sleep disorder are included in the XML file. Each annotation represents a 30-second sample. The following is an overview of the databases.

1. The MrOS Sleep Study: MrOS is an acronym for "Men's Study of Osteoporotic Fractures," which is the parent study. An initial assessment was conducted on 594 older men at least 65 years of age and drawn from six different clinical institutes between 2000 and 2002. As part of the sleep study, a total of 3,135 of these participants underwent complete unattended polysomnography and 3 to 5 days of actigraphy tests between December 2003 and March 2005. Additionally, each of these participants was required to keep an activity log for the duration of the study. The objective of the Sleep Study was to determine the degree to which sleep disturbances are associated with unfavorable outcomes in terms of overall health, such as a higher probability of passing away, fractures, falls, and cardiovascular disease.

2. The MESA Sleep Study: The Multi-Ethnic Study of Atherosclerosis (MESA) is a collaborative 6-center longitudinal investigation of factors associated with the development and progression of subclinical to clinical cardiovascular disease in 6,814 black, white, Hispanic, and Chinese American men and women aged 45 to 84 years in 2000-2002. The study was carried out to investigate factors associated with the development and progression of subclinical to clinical cardiovascular disease. During the investigation, the ages of the participants ranged from 45 to 84 years. Four follow-up examinations were performed during the following periods: 2003-2004, 2004-2005, 2005-2007, and 2010-2011. Furthermore, MESA Sleep conducted sleep exams on 227 people between 2010 and 2012 and received participant feedback. This evaluation consisted

of a sleep questionnaire, an unattended overnight polysomnogram, and a wrist-worn seven-day actigraphy. The objective of the sleep study is to establish whether there is a connection between the presence of subclinical atherosclerosis and sex, ethnicity, or any other demographic factors that are known to be associated with variances in sleep and sleep disorders.

3. The SHHS Sleep Study: The Sleep Heart Health Study is a cohort study conducted in multiple centers by researchers from the National Institute of Heart, Lung, and Blood from the USA The purpose of the study was to investigate the effects of sleep-disordered breathing on the cardiovascular system, as well as other aspects of a person's health, and to determine whether breathing problems that occur during sleep are related to an increased risk of coronary heart disease, stroke, death from all causes, and hypertension. Furthermore, the study aimed to determine whether breathing problems during sleep are related to an increased risk of snoring. SHHS Visit 1 research was carried out from November 1995 to January 1998 on 6,441 people, including men and women. All participants had to be at least 40 years old. During the third examination cycle, a second polysomnogram, also known as SHHS Visit 2, was performed on 3,295 participants (January 2001 to June 2003).

We consider only measurement channels that can be easily implemented in a home environment and are accessible in the three datasets: pulse oximetry (SpO_2), electrocardiogram (ECG), thoracic movement (ThorRes), abdominal movement (AbdoRes), and nasal airflow. Note that, instead of using ECG data as input to our model, we use R-to-R interval (RRI) data, which is widely available in wearable devices such as smartwatches. The RRI data is inferred from ECG data by measuring the time difference between heartbeats (from one R peak to the next R peak) using the Pan-Tompkins algorithm (see Section 2.4.1 for more details).

PSG recordings were included in our study if none of the following exclusion criteria were met: (1) total sleep time for PSG less than 1.5 hours; (2) poor-quality PSG recordings given by the presence of "unsure" or "noise" labels in more than a third of the total sleep time; and (3) the absence of one or more channels from the five channels selected for the study. After applying the exclusion criteria, a total of 10,880 sleep recordings were obtained to develop and evaluate the proposed method. Table 4.1 summarizes the databases.

Table 4.1 Data description.

Caracteristics	MESA	MROS	SHHS
Total PSG recordings	1516	2780	6584
Male	765	2780	3302
Female	751	0	3281
Age	69.6 (54–94)	78.3 (67–90)	74.4 (56–90)
BMI	28.7 (—)	27.1 (16.4–45.3)	27.7 (18.0–50.0)
Sleep time [min]	360.58 (181–599)	420 (184–739)	376.4 (242.0–473.5)
AHI [per hour]	24.3 (15.2–37.7)	24.0 (0.1–106.0)	22.0 (0.1–117.0)
Ethnicity			
White	35.1%	90.0%	85.7%
African	28.1%	3.4%	8.3%
Hispanic	23.7%	2.2%	—
Asian	13.1%	3.2%	—
Other	—	1.2%	6.0%

Data are reported in mean and range (in parenthesis).

Abbreviations: Apnea-Hypopnea Index (AHI), body mass index (BMI).

Because the data was acquired from multiple healthcare facilities at different sampling frequencies, all channels were resampled to ensure that they were consistent with the maximum sampling frequency. The cubic spline was used as the interpolation method (see Section 2.4.4 for more details); it is simple to implement and does not attenuate the higher frequency components of the signal. This approach allows us to normalize the data to the same length as is required for the input of the neural network.

4.3.2 Deep learning-model for the detection of OSA events

We developed a hybrid deep-learning model for the automatic classification of normal and OSA events from data. The inputs of the model are short segments of time-series data (e.g., 60s) recorded during the patients' sleep, which includes the physiological signals available in the considered datasets: RRI, SpO_2, ThorRes, AbdoRes, and/or nasal airflow. The model then classifies whether a given input corresponds to a normal or an OSA event. Figure 4.3 illustrates the pipeline of the method. For each available physiological signal, a distinct deep convolutional neural network (CNN) is trained to

automatically extract global features from the raw 1-dimensional signals. The features extracted by all CNNs are then combined on a light gradient-boos machine (LightGBM) [Ke et al., 2017] to perform the classification between normal and OSA events. The LightGBM yields the probability that a given sample belongs to the OSA class. The default classification threshold of 0.5 is used to perform the binary classification between the OSA and normal samples.

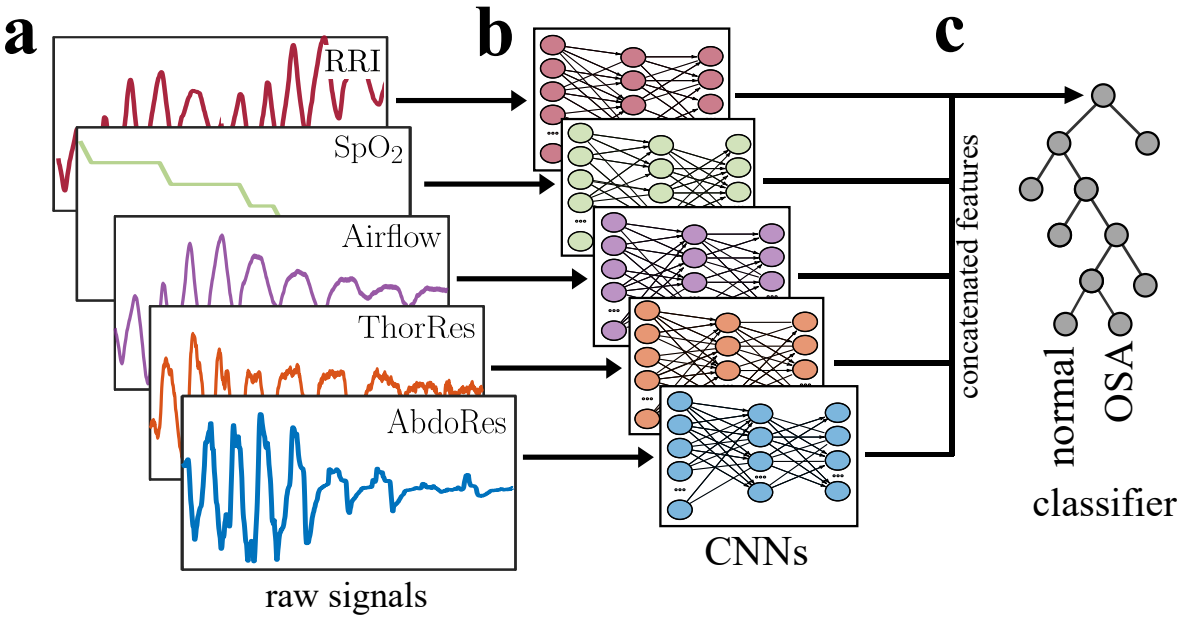

Fig. 4.3 Pipeline of the proposed method. (a) Data are separated by channels (RRI, SpO_2, among others) and segmented into 60 s windows. (b) For each channel, a distinct trained deep CNN extracts features (outputs) from the raw signal (input). (c) The extracted features are concatenated and fed into a LightGBM that classifies the input data between normal and OSA events.

Data preprocessing

First, data is split by measurement channel (RRI, SpO_2, etc) and segmented into 60s windows of "normal" and "OSA" events. Data windows correspond to the same time instance across all channels (Fig. 4.3a). Second, all signals are standardized and normalized by calculating their z-scores and applying min-max normalization to eliminate the mean and variance bias of the raw one-dimensional signal and speed up training [Huang et al., 2020]. Normalized samples of 60s from each channel are then fed in parallel into distinct deep CNNs. Each DNN is independently trained on a specific channel and then used for automatic feature extraction (Fig. 4.3b). Once the networks have been trained, features are extracted and concatenated to integrate information from different physiological biomarkers. This is achieved by using the concatenated features to train a LightGBM classifier for binary classification between normal and OSA events (Fig. 4.3c).

Neural network architecture and training

The CNNs were trained and cross-validated in 949,428 samples from 10,880 PSG recordings. We use the EfficientNetV2 architecture, a deep CNN with 479 layers developed by Google in 2021 [Tan et al., 2021b]. It is a modified and optimized version of EfficientNet [Tan et al., 2019], a popular image classification algorithm that won the ImageNet 2019 competition [He et al., 2019]. The architecture used in this study has been modified to handle unidimensional data and perform binary classification. Each CNN was independently trained using raw unidimensional physiological data from pulse oximetry (SpO2), electrocardiogram (RRI), thoracic movement (ThorRes), abdominal movement (AbdoRes), or airflow. Categorical cross-entropy was used as the loss function, ADAM as the optimizer [Kingma et al., 2014], and stochastic gradient descent as the objective function optimizer [Montavon et al., 2012]. If the validation loss did not decrease after eight consecutive epochs, the training was terminated. Once the networks were trained, the final layer was removed, and the last global average pooling layer is used to yield 1,280 features from each data channel.

Feature classifier

After training the neural networks and extracting the features, the next step is to concatenate the features extracted by all CNNs and use them to train a LightGBM classifier (see Section 2.2.7 for more details about the LightGBM classifier). Unlike many other well-known algorithms, such as XGBoost [Chen et al., 2015a] and GBDT [Ye et al., 2009], LightGBM employs the classification algorithm to grow trees in leaf fashion rather than in depth. The leaf-wise algorithm can converge significantly more quickly than the depth-wise growth method. However, its growth can be subject to overfitting if the appropriate hyperparameters are not used [Ke et al., 2017]. We use a random search method within a specified set of parameters to optimize the training and performance of LightGBM. This allows a fixed number of parameters of a particular distribution to be sampled rather than testing all the values of potential parameters [Bergstra et al., 2012].

4.4 Results

We train and evaluate the proposed method on raw physiological signals to reduce the complexity in the design and implementation of the model and to eliminate the need for human-created features. Performance is assessed using a 10-fold cross-validation strategy (see Section 2.5 for more details), separating data by patient record to avoid data leakage from samples taken from the same patient.

To overcome bias and inaccuracy associated with classification using imbalanced data between OSA and normal events, the undersampling approach was used to balance the number of samples from each individual record. OSA labeled events were sampled every 60s with a 10 s overlap. An equal number of normal events are randomly sampled from the same PSG record, yielding a total of 949,428 balanced samples.

4.4.1 Benchmark models comparison

We compared the performance of the proposed method with two other CNN architectures that serve as a benchmark, LSTM and Resnet [He et al., 2016], which are commonly used for the detection of apnea and hypopnea events [Mostafa et al., 2019]. The EfficientNetV2 architecture outperformed Resnet and LSTM by 10% and 28%, respectively (see Table 4.2).

Table 4.2 Performance for different network architectures.

Model	Accuracy
LSTM	0.71
1D Resnet	0.82
1D EfficientNetV2	0.91

Furthermore, we compared the performance of various classifiers on the EfficientNetV2 extracted features (see Table 4.3). We use a random search method within a specified set of parameters to optimize the training and performance of the classifier. This allows a fixed number of parameters of a particular distribution to be sampled instead of testing all the values of potential parameters [Bergstra et al., 2012]. After training, the LightGBM classifier showed the best performance, with a 15% improvement compared to a logistic regression.

Table 4.3 Performance for different classifiers.

Model	Accuracy
Logistic regression	0.79
Support vector machine	0.80
XGBoost	0.81
Random forest	0.90
Light Gradient Boosting Machine	0.91

Finally, we compare the effectiveness of various window sampling lengths (see Table 4.4). A small window, such as 10 seconds, resulted in poor performance and could lead to loss of information [Kraemer et al., 2018]. The performance did not improve when the window length increased from 60 to 120 seconds. Therefore, we fixed the window length to 60 seconds, which is a value commonly used in the literature [Ramachandran et al., 2021; Mendonca et al., 2018; Mostafa et al., 2019].

Table 4.4 Performance for different lengths of the sampling window.

Length (seconds)	Accuracy
10	0.82
20	0.89
30	0.90
60	0.91
120	0.91

4.4.2 Performance of the hybrid model

The performance of the proposed method in the classification task between normal and OSA events is evaluated in this section. The results are shown in Fig. 4.4 assuming that all five physiological signals are available for training and testing. The area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPRC) of 0.96 and 0.97, respectively (Fig. 4.4a,b). This demonstrates that the separation between OSA and normal events is highly accurate and has robust discriminative power regarding the positive class (OSA) and the negative class (normal). Fig. 4.4c shows the confusion matrix of the 10-folds on 949,428 total samples, containing the number of correct

and incorrect predictions made by the model for positive (OSA) and negative (normal) events. In all samples, the percentage of false negatives and false positives is less than 10%. This indicates that the method is not perfect but performs well in classifying both normal and OSA events.

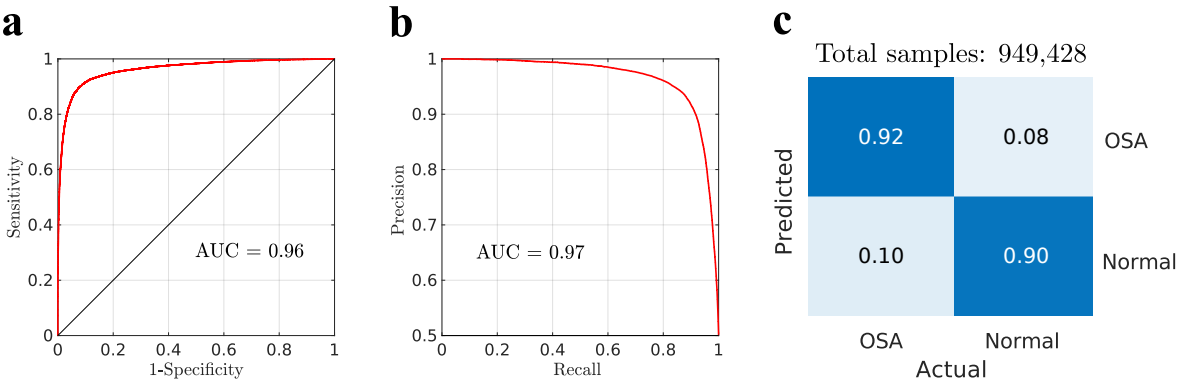

Fig. 4.4 Performance of the proposed method. (a) Receiver-operator characteristic curve. (b) Precision-recall curve. (c) Confusion matrix.

So far, the performance of the method was evaluated by assuming that all measurement channels were available as input. However, we expect that some physiological signals may contain features that contribute more to the classification task than other signals. Additionally, some physiological signals may not be available for sleep apnea in a home-care monitoring application. Therefore, it is interesting to evaluate the performance of our method for different combinations of input types (physiological signals). The hybrid model is trained and tested for each combination of input channels as previously described.

The results are shown in Table. 4.5 assuming that all five physiological signals are available for training and testing, it shows the performance results for different combinations of input channels. As expected, the best performance was obtained when all five channels were used together, and the performance generally decreases as the number of input channels is reduced. Generally, models trained and tested with AbdoRes and/or ThorRes as input achieve higher performance. However, performance appears to be less dependent on the RRI data; The model trained on the RRI data alone was also the one with the lowest performance.

Table 4.5 Performance of the proposed method for all databases.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.91	0.9	0.92	0.922	0.911
AbdoRes+ThorRes+Airflow+SpO_2	0.909	0.899	0.92	0.922	0.911
RRI+AbdoRes+Airflow+SpO_2	0.908	0.898	0.918	0.92	0.909
RRI+AbdoRes+ThorRes+Airflow	0.905	0.894	0.916	0.918	0.906
RRI+ThorRes+Airflow+SpO_2	0.904	0.894	0.915	0.917	0.905
RRI+AbdoRes+ThorRes+SpO_2	0.885	0.875	0.895	0.898	0.886
AbdoRes+Airflow+SpO_2	0.907	0.896	0.917	0.919	0.908
AbdoRes+ThorRes+Airflow	0.904	0.894	0.914	0.916	0.905
ThorRes+Airflow+SpO_2	0.903	0.894	0.914	0.916	0.905
RRI+AbdoRes+Airflow	0.902	0.892	0.912	0.914	0.903
RRI+Airflow+SpO_2	0.897	0.886	0.907	0.91	0.898
RRI+ThorRes+Airflow	0.896	0.884	0.909	0.912	0.898
AbdoRes+ThorRes+SpO_2	0.884	0.874	0.894	0.896	0.885
RRI+AbdoRes+SpO_2	0.879	0.87	0.888	0.891	0.88
RRI+AbdoRes+ThorRes	0.871	0.861	0.881	0.884	0.872
RRI+ThorRes+SpO_2	0.861	0.853	0.87	0.873	0.863
AbdoRes+Airflow	0.899	0.89	0.909	0.911	0.9
ThorRes+Airflow	0.895	0.883	0.907	0.91	0.897
Airflow+SpO_2	0.895	0.886	0.905	0.907	0.896
RRI+Airflow	0.881	0.869	0.894	0.897	0.883
AbdoRes+SpO_2	0.877	0.869	0.886	0.888	0.879
AbdoRes+ThorRes	0.865	0.856	0.874	0.877	0.866
ThorRes+SpO_2	0.859	0.852	0.866	0.869	0.86
RRI+AbdoRes	0.855	0.85	0.861	0.863	0.856
RRI+ThorRes	0.826	0.817	0.835	0.84	0.828
RRI+SpO_2	0.824	0.821	0.827	0.828	0.825
Airflow	0.878	0.867	0.89	0.893	0.88
AbdoRes	0.845	0.843	0.848	0.849	0.846
ThorRes	0.812	0.805	0.82	0.825	0.815
SpO_2	0.81	0.809	0.812	0.812	0.811
RRI	0.665	0.669	0.661	0.653	0.661

For models trained on two or more input channels and the model trained on AbdoRes data, the performance was consistently higher than previously published work. Table 4.7 compares the performance of our method with previously published methods in OSA detection in the literature using the same databases and at least 100 records. Our results achieve the highest score in terms of accuracy (91%), sensitivity (90%), specificity (92%), and F1-score (91%), with a good generalization of the data across the ten-fold cross-validation (as represented by a small standard deviation of 0.0053, see Table 4.6). This illustrates that the designed hybrid deep CNN classifier as well as the extensive dataset (consisting of three large databases) lead to a high improvement in the performance of our algorithm while keeping its computational complexity and data pre-processing step reasonably simple.

Table 4.6 10-fold cross-validation accuracy on the testing sets after training the EfficientNetV2 and the Light Gradient Boosting Machine using all channels.

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Average
0.91	0.91	0.91	0.90	0.91	0.90	0.90	0.91	0.90	0.90	0.91

Table 4.7 Comparison of the proposed method with state-of-the-art apnea detection methods using similar databases.

Year	Study	Database	Recordings	Signal used	Classifiers applied	Accuracy	f1-Score
2017	[Haidar et al., 2017]	MESA	100	Airflow	CNN1D	0.74	0.74
2017	[Van Steenkiste et al., 2018]	SHHS	2100	AbdoRes+ThorRes	LSTM	0.71	0.48
2018	[McCloskey et al., 2018]	MESA	1507	Airflow	CNN2D	0.79	0.79
2020	[Banluesombatkul et al., 2018]	MROS	545	ECG	CNN1D-LSTM	0.79	0.79
2022	[Sharma et al., 2022]	SHHS	8444	ECG+AbdoRes+ThorRes+Airflow+SpO_2	Wavelet decomposition	0.84	0.85
2022	Proposed Method	All	10880	RRI+AbdoRes+ThorRes+Airflow+SpO_2	Hybrid deep CNN classifier	0.91	0.91

Additionally, for comparison, we evaluated an alternative pipeline in which one CNN and multiple sensor inputs are used instead, hence using a S dimensional input array, where S stands for the number of sensors. Table 4.8 shows the performance of the alternative approach. There is a drop in accuracy of 5.5% when using the five channels and 4.3% for four channels. Overall, performance worsens compared to the former pipeline in Fig. 4.3. Therefore, using separate networks to tailor individual sensors yielded better results.

Table 4.8 Performance of a single neural network model with multiple input channels

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.8607	0.8493	0.8728	0.8770	0.8629
RRI+AbdoRes+ThorRes+Airflow	0.8715	0.8603	0.8834	0.8870	0.8735
AbdoRes+ThorRes+SpO_2	0.8392	0.8342	0.8444	0.8467	0.8404
RRI+AbdoRes+ThorRes	0.8428	0.8380	0.8478	0.8500	0.8428
AbdoRes+ThorRes	0.8513	0.8413	0.8620	0.8661	0.8535
RRI +SpO_2	0.6974	0.8071	0.6455	0.5189	0.6317

4.4.3 Out-of-distribution performance.

To evaluate how generalizable the method's performance is to out-of-sample data (that is, data not collected in the same study that was used to train the model), we employ a "leave-one-database-out" cross validation. In this validation, all but one database reported in Table 4.1 are used as training data, while the remaining one is used to evaluate the quality of the predictions on populations; hence the database used for testing is not used for training and vice versa.

Fig. 4.5b summarizes the out-of-distribution performance tested on each of the available databases (see Table 4.9, Table 4.10 and Table 4.11 for details of the specific values in Fig. 4.5). The best performance was obtained on the MESA and MROS databases when the five channels were used as inputs, achieving an accuracy (F1-score) of 87.98% (88.62%) and 90.23% (90.84%) for the MESA and MROS databases, respectively. Interestingly, in these cases, the out-of-distribution performance of the trained models achieved better results than previous methods published in the literature that were *trained and tested* in the same database. For example, the performance of our model trained using the MROS, and SHHS databases achieved better performance on the MESA database than Refs. [Haidar et al., 2017; McCloskey et al., 2018] that were trained on MESA. Such results demonstrate the high generalizability of our method to other types of data and different populations, especially when the SHHS is incorporated into the training phase.

In general, the more channels available as input to the model, the better the (out-of-distribution) performance of our method (Fig. 4.5). However, as an exception, the best out-of-distribution performance on the SHHS database was obtained using a combination of the AbdoRes, ThorRes, and SpO_2 channels, leading to an accuracy of 81.60% and F1-Score of 80.00%. In this case, the results employing RRI and airflow data as input channels led to poor performance.

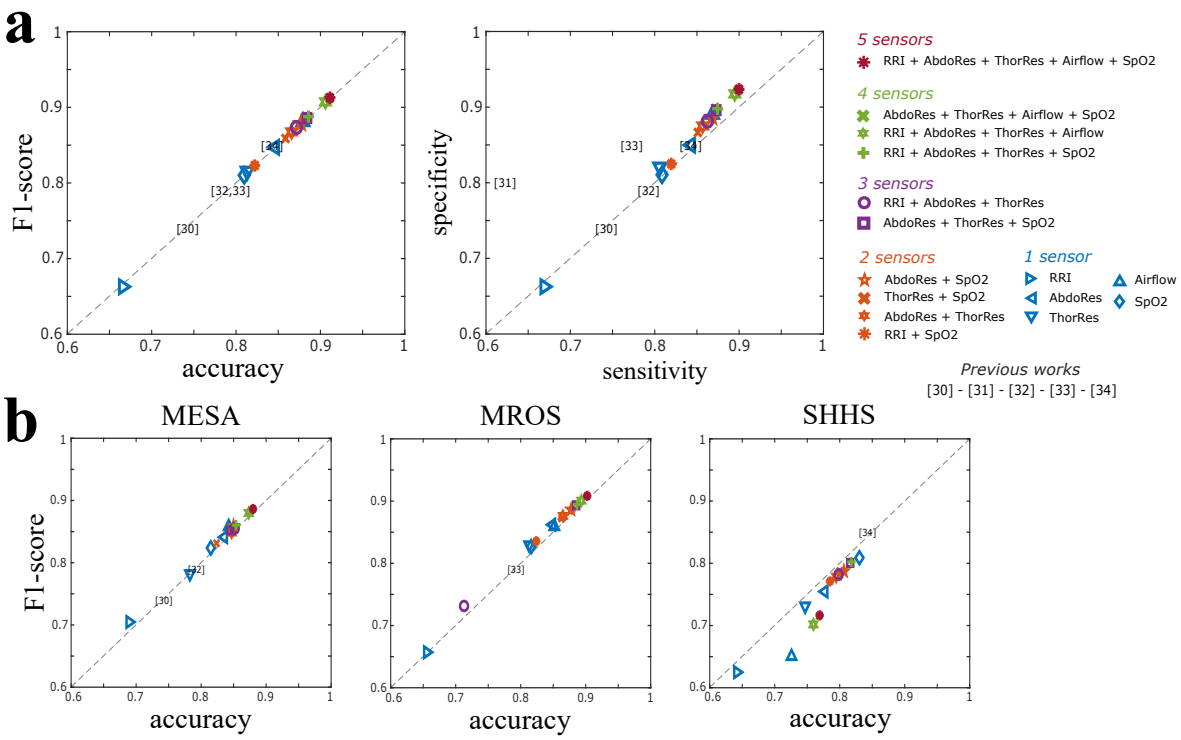

Fig. 4.5 Performance metrics of the proposed method. (a) Performance metrics for different combinations of input channels (physiological signals). Performance metrics from previous works in the literature using one of the databases studied in this work are included for comparison purposes. (b) Performance of the proposed method on data out of distribution using the "leave-one-database-out" methodology. Each plot shows the performance metrics of a model on a given database in which this database was not included in the training stage (e.g., performance metrics on the MESA database are shown for models trained on MROS and SHHS data only). Performance metrics from previous work *trained and tested* in the same database were included for comparison purposes.

Table 4.9 Performance of the proposed method on the MESA database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.874	0.831	0.93	0.939	0.882
RRI+AbdoRes+Airflow+SpO_2	0.875	0.826	0.942	0.951	0.884
RRI+AbdoRes+ThorRes+Airflow	0.874	0.834	0.924	0.933	0.881
AbdoRes+ThorRes+Airflow+SpO_2	0.873	0.833	0.923	0.933	0.88
RRI+ThorRes+Airflow+SpO_2	0.866	0.819	0.928	0.939	0.875
RRI+AbdoRes+ThorRes+SpO_2	0.85	0.815	0.894	0.906	0.858
AbdoRes+Airflow+SpO_2	0.877	0.831	0.938	0.946	0.885
RRI+AbdoRes+Airflow	0.876	0.83	0.937	0.946	0.884
RRI+Airflow+SpO_2	0.876	0.823	0.949	0.958	0.885
AbdoRes+ThorRes+Airflow	0.867	0.828	0.917	0.927	0.875
RRI+ThorRes+Airflow	0.865	0.822	0.922	0.932	0.874
RRI+AbdoRes+SpO_2	0.864	0.822	0.917	0.928	0.872
ThorRes+Airflow+SpO_2	0.864	0.819	0.923	0.934	0.873
RRI+AbdoRes+ThorRes	0.854	0.837	0.871	0.878	0.857
AbdoRes+ThorRes+SpO_2	0.844	0.808	0.888	0.901	0.852
RRI+ThorRes+SpO_2	0.829	0.796	0.869	0.883	0.838
AbdoRes+Airflow	0.867	0.82	0.929	0.939	0.876
ThorRes+Airflow	0.859	0.816	0.916	0.927	0.868
Airflow+SpO_2	0.858	0.801	0.943	0.954	0.871
RRI+Airflow	0.856	0.799	0.94	0.951	0.868
AbdoRes+ThorRes	0.849	0.84	0.859	0.863	0.851
AbdoRes+SpO_2	0.845	0.8	0.906	0.92	0.856
RRI+AbdoRes	0.844	0.81	0.886	0.899	0.852
ThorRes+SpO_2	0.82	0.79	0.858	0.873	0.829
RRI+SpO_2	0.815	0.778	0.865	0.883	0.827
RRI+ThorRes	0.807	0.802	0.813	0.817	0.809
Airflow	0.842	0.779	0.94	0.953	0.858
AbdoRes	0.836	0.813	0.864	0.874	0.842
SpO_2	0.815	0.786	0.851	0.866	0.824
ThorRes	0.783	0.784	0.783	0.783	0.783
RRI	0.692	0.674	0.714	0.743	0.707

Table 4.10 Performance of the proposed method on the MROS database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.9	0.851	0.965	0.97	0.907
AbdoRes+ThorRes+Airflow+SpO_2	0.9	0.852	0.965	0.97	0.907
RRI+AbdoRes+Airflow+SpO_2	0.896	0.847	0.962	0.968	0.903
RRI+ThorRes+Airflow+SpO_2	0.893	0.844	0.957	0.963	0.9
RRI+AbdoRes+ThorRes+Airflow	0.892	0.84	0.963	0.969	0.9
RRI+AbdoRes+ThorRes+SpO_2	0.886	0.836	0.953	0.96	0.894
AbdoRes+Airflow+SpO_2	0.897	0.848	0.962	0.968	0.904
ThorRes+Airflow+SpO_2	0.892	0.844	0.956	0.962	0.899
AbdoRes+ThorRes+Airflow	0.89	0.837	0.964	0.97	0.898
RRI+AbdoRes+Airflow	0.888	0.836	0.96	0.967	0.896
AbdoRes+ThorRes+SpO_2	0.885	0.833	0.954	0.961	0.893
RRI+ThorRes+Airflow	0.881	0.83	0.95	0.958	0.889
RRI+Airflow+SpO_2	0.879	0.832	0.943	0.952	0.888
RRI+AbdoRes+SpO_2	0.878	0.827	0.948	0.956	0.887
RRI+AbdoRes+ThorRes	0.867	0.812	0.946	0.956	0.878
RRI+ThorRes+SpO_2	0.866	0.819	0.929	0.939	0.875
AbdoRes+Airflow	0.886	0.832	0.961	0.968	0.895
ThorRes+Airflow	0.881	0.83	0.951	0.959	0.89
Airflow+SpO_2	0.88	0.835	0.94	0.948	0.888
AbdoRes+SpO_2	0.877	0.825	0.948	0.956	0.886
AbdoRes+ThorRes	0.863	0.805	0.948	0.958	0.875
ThorRes+SpO_2	0.862	0.815	0.926	0.937	0.872
RRI+Airflow	0.854	0.815	0.904	0.916	0.863
RRI+AbdoRes	0.852	0.799	0.927	0.94	0.864
RRI+SpO_2	0.827	0.785	0.882	0.899	0.838
RRI+ThorRes	0.824	0.778	0.889	0.908	0.838
Airflow	0.852	0.818	0.893	0.905	0.859
AbdoRes	0.848	0.791	0.934	0.947	0.862
SpO_2	0.817	0.782	0.862	0.879	0.828
ThorRes	0.812	0.76	0.889	0.911	0.829
RRI	0.658	0.657	0.658	0.659	0.658

Table 4.11 Performance of the proposed method on the SHHS database.

Channels	Accuracy	Sensitivity	Specificity	Precision	f1-Score
RRI+AbdoRes+ThorRes+Airflow+SpO_2	0.765	0.928	0.692	0.574	0.709
RRI+AbdoRes+ThorRes+SpO_2	0.818	0.882	0.773	0.735	0.802
RRI+AbdoRes+Airflow+SpO_2	0.765	0.925	0.693	0.578	0.711
AbdoRes+ThorRes+Airflow+SpO_2	0.763	0.924	0.691	0.573	0.707
RRI+AbdoRes+ThorRes+Airflow	0.756	0.92	0.684	0.561	0.697
RRI+ThorRes+Airflow+SpO_2	0.746	0.92	0.674	0.539	0.68
AbdoRes+ThorRes+SpO_2	0.816	0.878	0.771	0.734	0.8
RRI+AbdoRes+SpO_2	0.811	0.883	0.762	0.717	0.792
RRI+ThorRes+SpO_2	0.805	0.848	0.771	0.743	0.792
RRI+AbdoRes+ThorRes	0.801	0.856	0.76	0.723	0.784
AbdoRes+Airflow+SpO_2	0.763	0.92	0.691	0.575	0.708
RRI+AbdoRes+Airflow	0.754	0.919	0.683	0.558	0.694
AbdoRes+ThorRes+Airflow	0.753	0.918	0.681	0.555	0.692
RRI+Airflow+SpO_2	0.746	0.913	0.675	0.544	0.682
ThorRes+Airflow+SpO_2	0.744	0.917	0.673	0.537	0.678
RRI+ThorRes+Airflow	0.732	0.917	0.661	0.511	0.656
AbdoRes+SpO_2	0.808	0.877	0.76	0.715	0.788
ThorRes+SpO_2	0.801	0.842	0.769	0.741	0.788
AbdoRes+ThorRes	0.796	0.849	0.757	0.72	0.78
RRI+SpO_2	0.787	0.826	0.756	0.727	0.773
RRI+AbdoRes	0.785	0.849	0.741	0.693	0.763
RRI+ThorRes	0.756	0.786	0.731	0.703	0.742
AbdoRes+Airflow	0.753	0.915	0.681	0.557	0.692
Airflow+SpO_2	0.746	0.91	0.676	0.547	0.683
ThorRes+Airflow	0.734	0.913	0.663	0.517	0.66
RRI+Airflow	0.725	0.907	0.656	0.502	0.646
AbdoRes	0.779	0.844	0.734	0.683	0.755
SpO_2	0.779	0.817	0.749	0.719	0.765
ThorRes	0.747	0.782	0.72	0.686	0.731
Airflow	0.724	0.901	0.655	0.503	0.645
RRI	0.642	0.656	0.631	0.598	0.626

4.4.4 Continuous monitoring of sleep apnea

We expect that the algorithm developed can also be implemented for continuous (real-time) monitoring of sleep apnea events during PSG examinations or as part of a home OSA detection system. To simulate a continuous monitoring scenario and investigate the number of true positive and false positive OSA events given by our detection algorithm, a retrospective analysis is performed with data collected previously in the MESA, MROS, and SHHS studies. We evaluate the performance of our OSA detection algorithm when applied to the complete sleep record (comprising records with a duration of 8.78h, on average). Fig. 4.6 shows the overall performance of all patients. As in a real-time monitoring scenario, a moving window is implemented to consecutively sample short segments of data from the measured physiological signals (Fig. 4.6a, top), which are then fed to the hybrid model to perform the classification task of OSA and normal events. The hybrid model outputs the probability that a sampled data belongs to the OSA class, which changes over time depending on how close or far a patient is to an OSA event (Fig. 4.6a, bottom).

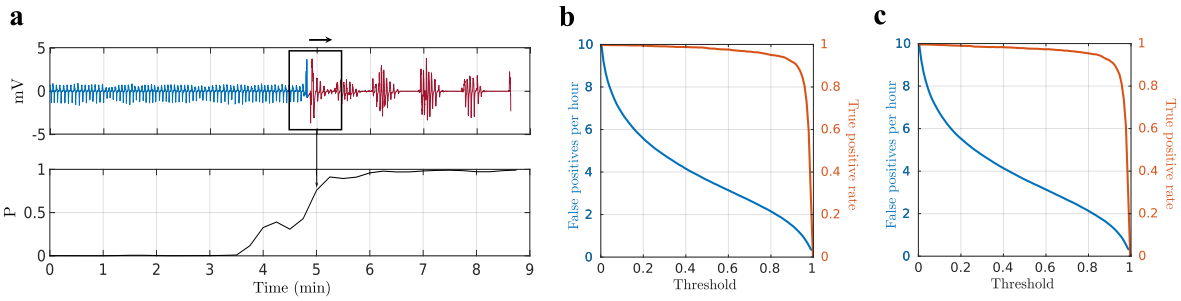

Fig. 4.6 Continuous monitoring of sleep apnea events. (a) A moving window continuously samples short segments of time-series data (from all sensor channels) during a sleep study (top) and feeds them to the hybrid model, which outputs the probability of an OSA event (bottom). (b,c) True-positive-rate and number of false positives per hour as a function of the classification threshold on the (b) validation set and (c) test set.

For the performance evaluation, we choose one of the trained models in the ten-fold cross-validation (Table 4.6), considering that all input channels are available. To tune the model for this task, we strive to maximize the average *ratio* between the true-positive rate and the number of false positives per hour across all records in the validation set. This procedure is designed as follows. First, for each patient's record, sequential segments of 60s are sampled consecutively with time steps of 15s and fed to the hybrid model, which computes the probability of each sample belonging to the OSA or normal class (Fig. 4.6a). Second, a parameter search is conducted to find the optimal classification threshold that maximizes the ratio in the validation set. Fig. 4.6b shows the true-positive rate and the number of false

positives per hour as a function of the classification threshold. The threshold is set as 0.98, which maximizes the ratio to 0.94.

After selecting the optimal threshold on the validation set, we evaluated the performance of the model on the testing set for continuous monitoring (Fig. 4.6c). We obtained an average number of false positives (per hour) of 0.34 and a true-positive rate of 0.51 across all records. The results are meaningful, considering that the average Apnea-Hypopnea index (AHI) of the patients is 23.2 events per hour. Fig. 4.6c shows that there is an inverse trade-off between the true-positive rate and the number of false positives per hour. For instance, with the default threshold of 0.50, the true positive rate is 0.96 and the false positives per hour are 3.64. This opens up the possibility of tuning the model in a patient-specific manner, personalizing the choice of threshold according to the user requirements or to the severity of AHI for each given patient.

4.5 Discussion

In this chapter, we have introduced a method for detecting OSA events using RRI, respiratory signals, and SpO_2 data. We showed that a deep neural network could be effectively trained using an extensive database to extract relevant features from raw physiological data. The detection was enhanced by combining it with a LightGBM classifier to obtain state-of-the-art performance.

We investigated the idea of combining respiratory, RRI, and SpO_2 signals rather than using them independently. Our results suggest that the combination of multiple input sources produced better results than the use of individual sources. We examine individual performance and ten combinations of these signals, as shown in Table 4.5. The individual airflow signal can perform better than the individual RRI and SpO_2 signals. However, combining these signals significantly improved performance.

The classification performance of the model obtained from the three datasets is consistent, as evidenced by a high area under the receiver operator characteristic curve, as well as the precision and recall curve, with values of 0.96 and 0.97, respectively, indicating that our model is not biased towards any class. To generate a robust detection method and prevent overfitting of the model, we employ a 10-fold cross-validation strategy.

The proposed method demonstrated robust generalization as measured by the out of distributions performance. The best results were obtained when testing on MESA and MROS, with an accuracy of 87.98% and 90.23%, respectively. The proposed method outperformed previous studies that trained and tested performance in the same databases (Table 4.7). When testing on the SHHS, the performance dropped to 81.60%, which can be attributed to the fact that 60% of the total available records come

from this database and were excluded. Therefore, the number of samples for training is significantly reduced. Despite this, the best results were obtained when only using AbdoRes, ThorRes, and SpO_2. This performance is still comparable to those obtained in the literature (see Table 4.7), although the databases were not used for training.

Table 4.7 compares the proposed method with other studies conducted using similar databases and more than 100 PSG recordings. As shown in the table, our proposed method outperformed previous studies. Haidar et al. [Haidar et al., 2017] used a 1D CNN to classify apnea-hypopnea events from the MESA database; using raw airflow data, they achieved a 74% accuracy [Van Steenkiste et al., 2018] used the SHHS data set for the detection of sleep apnea, using LSTM to automatically learn and extract relevant features, as well as detect possible sleep apnea events from raw physiological respiratory signals, with a 70% accuracy. McCloskey et al. [McCloskey et al., 2018] employed only MESA to perform a multiclass classification (normal, apnea, and hypopnea) and used a 2D CNN to achieve an accuracy of 77% by examining nasal airflow spectrograms. Banluesombatkul et al. [Banluesombatkul et al., 2018] combined CNN1D and LSTM to detect extremely severe OSA patients in normal subjects using ECGs from the MrOS dataset. They achieved a 79% accuracy rate. Sharma et al. [Sharma et al., 2022] used the SHHS dataset to detect OSA events. They trained a GentleBoost using features of the frequency domain of pulse oximetry (SpO_2) and respiratory data obtaining a 70% accuracy.

In general, our approach is more straightforward than others because we only used standardized raw physiological data to detect OSA events, which requires minimal complexity. As a result, it could be used as part of a home sleep monitoring system and considered for a real-time sleep detection application. A real-time detector may potentially stop apnea without the need for a continuous positive airway pressure device (CPAP) which is the current gold standard treatment for OSA sufferers. CPAP is uncomfortable and very invasive. Moreover, it has to remain "on" throughout the night. By employing an unobtrusive monitoring system and our detection method, it could be possible to urge the patient to resume regular breathing by gently shaking the bed or pillow or generating a sound to interrupt the apnea. Therefore, we expect that in the near future, algorithms like the one proposed in this research will be embedded in smart devices to monitor patients throughout the night at home and potentially provide a noninvasive treatment to OSA.

Chapter 5

Conclusion

The use of machine learning algorithms in medical applications has attracted attention in recent years. The willingness of physicians to contribute to research and the availability of rich open databases such as the National Sleep Research Resource for Sleep-Related Studies (NSRR) [Zhang et al., 2018] and Physionet [Moody et al., 2001] have provided researchers with access to health care data and ultimately made it possible for this thesis to be conducted. We presented methods and pipelines that allow the prediction and/or detection of life-threatening events. In particular, anticipating the onset of Atrial Fibrillation (AF) and detecting obstructive sleep apnea (OSA) events from physiological time series data.

This thesis provided a detailed technical background, addressing various topics, including machine learning, data preprocessing, model evaluation methods, and time series prediction metrics in the chapter 2.

Then, in the context of AF in Chapter 3, we provide a biological review of the structure of the heart, the function of the heart, the activity of the heart during AF, the pathophysiology of AF, the risk and treatment of AF, and the use of an electrocardiogram (ECG) and R-to-R intervals to assess the activity of the heart. Together with a detailed review of the literature on the current state of the field. This provided valuable information on the limitations of previous methods. A deep neural network model was used to create a prediction model called WARN (Warning of Atrial fibRilatioN) based on recurrence plots of the RRI data. WARN can detect early warning signals of AF on average 31 minutes before they appear, with an accuracy of 83% and an F1 score of 85%. Our results showed that the prediction horizon and the accuracy of the early warning signal could not be maximized simultaneously. However, we showed that WARN could be adjusted to optimize this trade-off.

Subsequently, in the context of Obstructive Sleep Apnea OSA in Chapter 4 we presented a review of sleep disorders, the use of polysomnography (PSG) to collect data from multiple sensors, the different types of sleep disorders, the pathophysiology and the risk and treatment of OSA. A detailed review of the literature on the field's current state for detecting sleep apnea events from single and multiple sensors was presented. Physiological time series data was used to detect OSA events by implementing deep neural networks and a light gradient boosting machine model. The developed method achieved the highest classification performance in the literature, with accuracy and F1-score of 91%. Moreover, we reported the performance of different combinations of sensors.

5.1 Main results

The following is a list of significant findings that are believed to be instructive for other researchers.

- It is possible to find early warning signals of AF from low-sampled data, such as RRI from heartbeats, by implementing recurrence plots; such a representation of the data may reveal patterns and temporal dependencies that can be identified by a deep neural network.
- It is not possible to simultaneously maximize all performance metrics and the predicted time horizon before AF onset. Figure 3.17 shows the reveal trade-off. For example, the maximum predicted time horizon until the onset of AF (that is, the instant of the first early warning until the onset of AF) is achieved at low thresholds. However, the accuracy is very low. Similarly, the predicted time horizon is relatively short when accuracy is maximized with a high threshold.
- The proposed method can be used to produce probabilities rather than perform classification. A forecasting method that provides non-threshold probabilities could help medical professionals better address the decision-making regarding the treatment of patients.
- It is possible to detect OSA events without expert domain knowledge by combining unsupervised feature extraction from deep neural networks and classification with a light gradient boosting machine applied on raw physiological time series data from single and multiple sensors, such as RRI, pulse oximetry (SpO_2), thoracic movement (ThorRes), abdominal movement (AbdoRes), and nasal airflow.
- According to the analysis of the performance of multiple neural network architectures, EfficientNetV2 exceeded the shallower CNNs and LSTM models in the unsupervised feature extraction task (see Table 4.2). In the classification task, the light gradient-boos machine outperformed

other classification methods (Table 4.3). We showed that the hybrid strategy, which combines different machine learning approaches, outperformed a single-model approach. The best trade-off between performance and number of sensors in detecting OSA events was obtained by combining AbdoRes and SpO_2 sensors only.

- In general, the proposed methods can be easily implemented at home without expert supervision, since they rely on data that can be easily sampled from affordable and non-invasive wearable devices. Therefore, improving patient confort and reducing the burden on the healthcare system.

5.2 Limitations

The following limitations were identified in chapter 3 and chapter 4.

- Information such as age, sex, ethnicity, drugs, and comorbidities of patients were not considered for method development, as most of them were unknown or not consistent across datasets. We hypothesize that the performance of the methods could be improved and better personalized by considering specific subgroups and confounders.

- Good data quality and availability were one of the difficulties in developing this thesis. Continuous measurements needed to develop forecasting models are rare. Furthermore, the few articles that have used this type of data for their findings have not made their data available.

- The fact that the data used in this study were collected from hospitals and sleep laboratory environments is one of the weaknesses of the study in its implementation in wearable technologies. When possible, similar studies should be carried out from data collected in the same environment in which they are aimed to be implemented.

5.3 Future perspectives

This thesis represents the first step in a more extensive research program that seeks to create methods for prognosis and continuous monitoring from affordable and comfortable wearable devices. The following is a list of possible research directions for this work.

- The method's success in forecasting critical events depends significantly on observations in the time-horizon window before the regime change. This would imply that predicting the underlying

patterns present in this time gap could be more insightful than forecasting the event itself to gain a better understanding of the underlying mechanism.

- It would be interesting to observe how the models respond in a given context in which additional information about patient characteristics, such as body mass index (BMI), age, sex, etc., are used together to produce a more complete and well-informed set of variables.

- The primary objective of the proposed methods was to predict the onset of AF and detect OSA events from physiological time series data. However, various other medical applications can use the described methodology, especially in the context of turning detection models into predictors, by performing continuous measurements as presented here in this thesis.

References

Acharya, U. R. et al. (2017). “Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network”. In: *Information sciences* 405, pp. 81–90 (cit. on p. 63).

Albawi, S., T. A. Mohammed, and S. Al-Zawi (2017). “Understanding of a convolutional neural network”. In: *2017 international conference on engineering and technology (ICET)*. Ieee, pp. 1–6 (cit. on p. 24).

Alcaraz, R., A. Martínez, and J. J. Rieta (2015). “Role of the P-wave high frequency energy and duration as noninvasive cardiovascular predictors of paroxysmal atrial fibrillation”. In: *Computer Methods and Programs in Biomedicine* 119.2, pp. 110–119 (cit. on p. 66).

Alila Medical Media (2007). *Schematic representation of a normal sinus rhythm ECG wave* (cit. on p. 98).

Aljanabi, M, H. Qutqut, and M. Hijjawi (2018). “Machine learning classification techniques for heart disease prediction: A review”. In: *International Journal of Engineering & Technology* 7.4, pp. 5373–5379 (cit. on pp. 5, 32, 50, 66).

Allessie, M. A. et al. (2001). “Pathophysiology and prevention of atrial fibrillation”. In: *Circulation* 103.5, pp. 769–777 (cit. on p. 53).

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press (cit. on p. 9).

Andreotti, F. et al. (2017). “Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG”. In: *2017 Computing in Cardiology (CinC)*. IEEE, pp. 1–4 (cit. on p. 64).

Anthony Atkielski (2007). *Schematic representation of a normal sinus rhythm ECG wave* (cit. on p. 51).

Anwar, S. M. et al. (2018). “Arrhythmia classification of ECG signals using hybrid features”. In: *Computational and mathematical methods in medicine* 2018 (cit. on p. 65).

Aphex (2015). *Typical CNN architecture* (cit. on p. 31).

Ardabili, S., A. Mosavi, and A. R. Várkonyi-Kóczy (2019). “Advances in machine learning modeling reviewing hybrid and ensemble methods”. In: *International Conference on Global Research and Education*. Springer, pp. 215–227 (cit. on p. 17).

Arzt, M. et al. (2007). “Suppression of central sleep apnea by continuous positive airway pressure and transplant-free survival in heart failure: a post hoc analysis of the Canadian Continuous Positive Airway Pressure for Patients with Central Sleep Apnea and Heart Failure Trial (CANPAP)”. In: *Circulation* 115.25, pp. 3173–3180 (cit. on p. 96).

Association, W. M. et al. (2009). "Declaration of Helsinki. Ethical principles for medical research involving human subjects". In: *Jahrbuch Für Wissenschaft Und Ethik* 14.1, pp. 233–238 (cit. on p. 69).

Attia, Z. I. et al. (2019). "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction". In: *The Lancet* 394.10201, pp. 861–867 (cit. on p. 66).

Attisano, L. and J. L. Wrana (2002). "Signal transduction by the TGF-β superfamily". In: *Science* 296.5573, pp. 1646–1647 (cit. on p. 58).

Aurora, R. N. et al. (2010). "Practice parameters for the surgical modifications of the upper airway for obstructive sleep apnea in adults". In: *Sleep* 33.10, pp. 1408–1413 (cit. on p. 108).

Ausma, J. et al. (1997). "Structural changes of atrial myocardium due to sustained atrial fibrillation in the goat". In: *Circulation* 96.9, pp. 3157–3163 (cit. on p. 55).

Aytemir, K. et al. (1999). "Prediction of atrial fibrillation recurrence after cardioversion by P wave signal-averaged electrocardiography". In: *International Journal of Cardiology* 70.1, pp. 15–21 (cit. on p. 66).

Baber, U. et al. (2011). "Association of chronic kidney disease with atrial fibrillation among adults in the United States: REasons for Geographic and Racial Differences in Stroke (REGARDS) Study". In: *Circulation: Arrhythmia and Electrophysiology* 4.1, pp. 26–32 (cit. on p. 60).

Banluesombatkul, N., T. Rakthanmanon, and T. Wilaiprasitporn (2018). "Single channel ECG for obstructive sleep apnea severity detection using a deep learning approach". In: *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, pp. 2011–2016 (cit. on pp. 122, 130).

Barbarossa, A., F. Guerra, and A. Capucci (2014). "Silent atrial fibrillation: a critical review". In: *Journal of atrial fibrillation* 7.3 (cit. on p. 55).

Battula, B. P. and R. S. Prasad (2013). "An overview of recent machine learning strategies in data mining". In: *International Journal of Advanced Computer Science and Applications* 4.3 (cit. on p. 3).

Baudino, T. A. et al. (2006). "Cardiac fibroblasts: friend or foe?" In: *American Journal of Physiology-Heart and Circulatory Physiology* 291.3, H1015–H1026 (cit. on p. 57).

Bektaş, S. and Y. Şişman (2010). "The comparison of L1 and L2-norm minimization methods". In: *International Journal of the Physical Sciences* 5.11, pp. 1721–1727 (cit. on p. 22).

Bengio, Y., A. Courville, and P. Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828 (cit. on pp. 11, 24).

Bengio, Y., P. Simard, and P. Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2, pp. 157–166 (cit. on p. 29).

Bennett, C. C. and K. Hauser (2013). "Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach". In: *Artificial intelligence in medicine* 57.1, pp. 9–19 (cit. on p. 6).

Bergstra, J. and Y. Bengio (2012). "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2 (cit. on pp. 117, 118).

Berry, R. B. et al. (2012). "Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the

American Academy of Sleep Medicine". In: *Journal of clinical sleep medicine* 8.5, pp. 597–619 (cit. on p. 101).

Betts, J. G. et al. (2013). *Anatomy and physiology* (cit. on p. 48).

Bignold, J. J. et al. (2011). "Accurate position monitoring and improved supine-dependent obstructive sleep apnea with a new position recording and supine avoidance device". In: *Journal of clinical sleep medicine* (cit. on p. 109).

Biton, S. et al. (2021). "Atrial fibrillation risk prediction from the 12-lead electrocardiogram using digital biomarkers and deep representation learning". In: *European Heart Journal-Digital Health* 2.4, pp. 576–585 (cit. on p. 66).

Blackburn, H. et al. (1960). "The electrocardiogram in population studies: a classification system". In: *Circulation* 21.6, pp. 1160–1175 (cit. on p. 49).

Blackwell, T. et al. (2011). "Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study". In: *Journal of the American Geriatrics Society* 59.12, pp. 2217–2225 (cit. on p. 113).

Blanche, C. et al. (2013). "Value of P-wave signal averaging to predict atrial fibrillation recurrences after pulmonary vein isolation". In: *EP Europace* 15.2, pp. 198–204 (cit. on p. 66).

Bojarski, M. et al. (2016). "End to end learning for self-driving cars". In: *arXiv preprint arXiv:1604.07316* (cit. on p. 32).

Boon, K. H., M. Khalil-Hani, and M. Malarvili (2018). "Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III". In: *Computer Methods and Programs in Biomedicine* 153, pp. 171–184 (cit. on pp. 66, 88).

Boon, K. H. et al. (2016). "Paroxysmal atrial fibrillation prediction method with shorter HRV sequences". In: *Computer Methods and Programs in Biomedicine* 134, pp. 187–196 (cit. on pp. 66, 67, 88).

Botto, G. L. et al. (2021). "Impact of the pattern of atrial fibrillation on stroke risk and mortality". In: *Arrhythmia & Electrophysiology Review* 10.2, p. 68 (cit. on p. 60).

Breiman, L. (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32 (cit. on p. 20).

Brooks, A. G. et al. (2010). "Outcomes of long-standing persistent atrial fibrillation ablation: a systematic review". In: *Heart rhythm* 7.6, pp. 835–846 (cit. on p. 62).

Bryan Pardo (2020). *Short-time Fourier Transform* (cit. on p. 41).

Buch, P et al. (2003). "Reduced lung function and risk of atrial fibrillation in the Copenhagen City Heart Study". In: *European Respiratory Journal* 21.6, pp. 1012–1016 (cit. on p. 60).

Buckberg, G. D. et al. (2018). "What is the heart? Anatomy, function, pathophysiology, and misconceptions". In: *Journal of cardiovascular development and disease* 5.2, p. 33 (cit. on p. 48).

Burges, C. J. (1998). "A tutorial on support vector machines for pattern recognition". In: *Data mining and knowledge discovery* 2.2, pp. 121–167 (cit. on pp. 13, 16).

Burstein, B. and S. Nattel (2008). "Atrial fibrosis: mechanisms and clinical relevance in atrial fibrillation". In: *Journal of the American College of Cardiology* 51.8, pp. 802–809 (cit. on pp. 57, 59, 62).

Bury, T. M. et al. (2021). “Deep learning for early warning signals of tipping points”. In: *Proceedings of the National Academy of Sciences* 118.39, e2106140118 (cit. on pp. 32, 112).

Calkins, H. et al. (2017). “2017 HRS/EHRA/ECAS/APHRS/SOLAECE expert consensus statement on catheter and surgical ablation of atrial fibrillation: executive summary”. In: *Heart Rhythm* (cit. on p. 62).

Camelliti, P., T. K. Borg, and P. Kohl (2005). “Structural and functional characterisation of cardiac fibroblasts”. In: *Cardiovascular research* 65.1, pp. 40–51 (cit. on p. 57).

Camm, A. J. et al. (1996). “Heart rate variability. Standards of measurement, physiological interpretation, and clinical use”. In: (cit. on p. 110).

Carbonell, J. G., R. S. Michalski, and T. M. Mitchell (1983). “Machine learning: A historical and methodological analysis”. In: *AI Magazine* 4.3, pp. 69–69 (cit. on p. 10).

Cartwright, R. D. et al. (1985). “Sleep position training as treatment for sleep apnea syndrome: a preliminary study”. In: *Sleep* 8.2, pp. 87–94 (cit. on p. 109).

Cartwright, R. D. (1984). “Effect of sleep position on sleep apnea severity”. In: *Sleep* 7.2, pp. 110–114 (cit. on p. 106).

Carvalho, N. C. et al. (2018). “Recurrence plots for the assessment of patient-ventilator interactions quality during invasive mechanical ventilation”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.8, p. 085707 (cit. on p. 72).

Castiglioni, I. et al. (2021). “AI applications to medical images: From machine learning to deep learning”. In: *Physica Medica* 83, pp. 9–24 (cit. on p. 32).

Censi, F. et al. (2000). “Recurrent patterns of atrial depolarization during atrial fibrillation assessed by recurrence plot quantification”. In: *Annals of Biomedical Engineering* 28.1, pp. 61–70 (cit. on p. 72).

Chaldoupi, S.-M. et al. (2009). “The role of connexin40 in atrial fibrillation”. In: *Cardiovascular research* 84.1, pp. 15–23 (cit. on pp. 55, 57).

Chamberlain, A. M. et al. (2011). “Smoking and incidence of atrial fibrillation: results from the Atherosclerosis Risk in Communities (ARIC) study”. In: *Heart Rhythm* 8.8, pp. 1160–1166 (cit. on p. 60).

Chazal, P. de, T. Penzel, and C. Heneghan (2004). “Automated detection of obstructive sleep apnoea at different time scales using the electrocardiogram”. In: *Physiological measurement* 25.4, p. 967 (cit. on p. 110).

Chen, T. and C. Guestrin (2016). “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794 (cit. on pp. 21–23).

Chen, T. et al. (2015a). “Xgboost: extreme gradient boosting”. In: *R package version 0.4-2* 1.4, pp. 1–4 (cit. on p. 117).

Chen, X. et al. (2015b). “Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA)”. In: *Sleep* 38.6, pp. 877–888 (cit. on p. 113).

Chesnokov, Y. V. (2008). “Complexity and spectral analysis of the heart rate variability dynamics for distant prediction of paroxysmal atrial fibrillation with artificial intelligence methods”. In: *Artificial Intelligence in Medicine* 43.2, pp. 151–165 (cit. on p. 66).

Cho, J., Y. Kim, and M. Lee (2018). "Prediction to atrial fibrillation using deep convolutional neural networks". In: *International Workshop on Predictive Intelligence in Medicine*. Springer, pp. 164–171 (cit. on p. 66).

Chollet, F. (2017). "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258 (cit. on p. 33).

Chuva De Sousa Lopes, S. M. et al. (2004). "Connective tissue growth factor expression and Smad signaling during mouse heart development and myocardial infarction". In: *Developmental dynamics: an official publication of the American Association of Anatomists* 231.3, pp. 542–550 (cit. on p. 59).

Cioffi, R. et al. (2020). "Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions". In: *Sustainability* 12.2, p. 492 (cit. on p. 1).

Ciprandi, G. et al. (2007). "Intranasal flunisolide treatment in children with adenoidal hypertrophy". In: *International journal of immunopathology and pharmacology* 20.4, pp. 833–836 (cit. on p. 104).

Claes, N. et al. (2012). "Prevalence of atrial fibrillation in adults participating in a large-scale voluntary screening programme in Belgium". In: *Acta cardiologica* 67.3, pp. 273–278 (cit. on p. 60).

Clavier, L et al. (2002). "Automatic P-wave analysis of patients prone to atrial fibrillation". In: *Medical and Biological Engineering and Computing* 40.1, pp. 63–71 (cit. on p. 66).

Collobert, R. et al. (2011). "Natural language processing (almost) from scratch". In: *Journal of machine learning research* 12.ARTICLE, pp. 2493–2537 (cit. on p. 32).

Cosedis Nielsen, J. et al. (2012). "Radiofrequency ablation as initial therapy in paroxysmal atrial fibrillation". In: *New England Journal of Medicine* 367.17, pp. 1587–1595 (cit. on p. 63).

Costin, H., C. Rotariu, and A. Păsărică (2013). "Atrial fibrillation onset prediction using variability of ECG signals". In: *2013 8th International symposium on advanced topics in electrical engineering (ATEE)*. IEEE, pp. 1–4 (cit. on pp. 66, 88).

Cracowski, C. et al. (2001). "Characterization of obstructive nonapneic respiratory events in moderate sleep apnea syndrome". In: *American journal of respiratory and critical care medicine* 164.6, pp. 944–948 (cit. on p. 97).

Crawford, J. and L. Doherty (2012). *Practical aspects of ECG recording*. M&K Update Ltd (cit. on pp. 49, 52, 54, 74).

Dahlqvist, J. et al. (2007). "Physical findings in the upper airways related to obstructive sleep apnea in men and women". In: *Acta oto-laryngologica* 127.6, pp. 623–630 (cit. on pp. 105, 108).

Darby, I. A. and T. D. Hewitson (2007). "Fibroblast differentiation in wound healing and fibrosis". In: *International review of cytology* 257, pp. 143–179 (cit. on p. 58).

Dash, S et al. (2009). "Automatic real time detection of atrial fibrillation". In: *Annals of Biomedical Engineering* 37.9, pp. 1701–1709 (cit. on p. 66).

Davies, M. and A. Pomerance (1972). "Pathology of atrial fibrillation in man." In: *British heart journal* 34.5, p. 520 (cit. on p. 53).

DeBrusk, C. (2018). "The risk of machine-learning bias (and how to prevent it)". In: *MIT Sloan Management Review* (cit. on p. 42).

Dempsey, J. A. et al. (2010). "Pathophysiology of sleep apnea". In: *Physiological reviews* 90.1, pp. 47–112 (cit. on pp. 94–96, 98, 103, 104).

Deng, J. et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255 (cit. on p. 5).

Denil, M., D. Matheson, and N. De Freitas (2014). "Narrowing the gap: Random forests in theory and in practice". In: *International conference on machine learning*. PMLR, pp. 665–673 (cit. on p. 20).

Desai, D. S. and S. Hajouli (2022). "Arrhythmias". In: *StatPearls [Internet]*. StatPearls Publishing (cit. on p. 52).

DeSilva, R. A. et al. (1980). "Cardioversion and defibrillation". In: *American heart journal* 100.6, pp. 881–895 (cit. on p. 61).

Dixon, J. B. et al. (2012). "Surgical vs conventional therapy for weight loss treatment of obstructive sleep apnea: a randomized controlled trial". In: *Jama* 308.11, pp. 1142–1149 (cit. on p. 109).

Ebrahimzadeh, E. et al. (2018). "Prediction of paroxysmal Atrial Fibrillation: a machine learning based approach using combined feature vector and mixture of expert classification on HRV signal". In: *Computer Methods and Programs in Biomedicine* 165, pp. 53–67 (cit. on pp. 66, 88).

Eckert, D. J. and A. Malhotra (2008). "Pathophysiology of adult obstructive sleep apnea". In: *Proceedings of the American thoracic society* 5.2, pp. 144–153 (cit. on p. 98).

Eckert, D. J. and M. K. Younes (2014). "Arousal from sleep: implications for obstructive sleep apnea pathogenesis and treatment". In: *Journal of Applied Physiology* 116.3, pp. 302–313 (cit. on p. 98).

Eckmann, J.-P., S. O. Kamphorst, D. Ruelle, et al. (1995). "Recurrence plots of dynamical systems". In: *World Scientific Series on Nonlinear Science Series A* 16, pp. 441–446 (cit. on p. 39).

El Moaqet, H. et al. (2017). "A new algorithm for short term prediction of persistent atrial fibrillation". In: *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies*. IEEE, pp. 1–6 (cit. on p. 66).

Ellinor, P. T. et al. (2010). "Common variants in KCNN3 are associated with lone atrial fibrillation". In: *Nature genetics* 42.3, pp. 240–244 (cit. on p. 59).

Erdenebayar, U. et al. (2019). "Automatic prediction of atrial fibrillation based on convolutional neural network using a short-term normal electrocardiogram signal". In: *Journal of Korean Medical Science* 34.7 (cit. on pp. 66, 73).

Ersu, R. and F. Proulx (2020). "Multiple Sleep Latency Test". In: *American Journal of Respiratory and Critical Care Medicine* 201.5, P9–P10 (cit. on p. 101).

Esteva, A. et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639, pp. 115–118 (cit. on p. 2).

Faust, O. et al. (2018). "Deep learning for healthcare applications based on physiological signals: A review". In: *Computer methods and programs in biomedicine* 161, pp. 1–13 (cit. on p. 112).

Faust, O. et al. (2019). "A review of automated sleep stage scoring based on physiological signals for the new millennia". In: *Computer methods and programs in biomedicine* 176, pp. 81–91 (cit. on p. 109).

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996). "From data mining to knowledge discovery in databases". In: *AI magazine* 17.3, pp. 37–37 (cit. on pp. 2, 3).

Fonseca, P. et al. (2020). "Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population". In: *Sleep* 43.9, zsaa048 (cit. on p. 111).

Friberg, L. et al. (2014). "High prevalence of atrial fibrillation among patients with ischemic stroke". In: *Stroke* 45.9, pp. 2599–2605 (cit. on p. 54).

Frické, M. (2009). "The knowledge pyramid: a critique of the DIKW hierarchy". In: *Journal of information science* 35.2, pp. 131–142 (cit. on p. 3).

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232 (cit. on p. 21).

Fujita, S. et al. (1981). "Surgical correction of anatomic abnormalities in obstructive sleep apnea syndrome: uvulopalatopharyngoplasty". In: *Otolaryngology–head and neck surgery* 89.6, pp. 923–934 (cit. on p. 108).

Ganapathy, N., R. Swaminathan, and T. M. Deserno (2018). "Deep learning on 1-D biosignals: a taxonomy-based survey". In: *Yearbook of medical informatics* 27.01, pp. 098–109 (cit. on p. 112).

Garavaglia, S. and A. Sharma (1998). "A smart guide to dummy variables: Four applications and a macro". In: *Proceedings of the northeast SAS users group conference*. Vol. 43 (cit. on p. 26).

Garcia, M. I. and D. Boehning (2017). "Cardiac inositol 1, 4, 5-trisphosphate receptors". In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1864.6, pp. 907–914 (cit. on p. 58).

Gay, P. C. (2014). *Central Sleep Apnea, an issue of sleep medicine clinics*. Vol. 9. 1. Elsevier Health Sciences (cit. on p. 95).

Ghassemi, M. et al. (2020). "A review of challenges and opportunities in machine learning for health". In: *AMIA Summits on Translational Science Proceedings* 2020, p. 191 (cit. on p. 6).

Ghojogh, B. and M. Crowley (2019). "The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial". In: *arXiv preprint arXiv:1905.12787* (cit. on pp. 19, 20, 23).

Glomset, D. and K. Cross (1952). "Morphologic study of the cardiac conduction system: VI. the intrinsic nervous system of the heart". In: *AMA Archives of Internal Medicine* 89.6, pp. 923–930 (cit. on p. 53).

Goldberger, A. et al. (2000). "Components of a new research resource for complex physiologic signals". In: *PhysioBank, PhysioToolkit, and Physionet* (cit. on p. 63).

Golubnitschaja, O., J. Kinkorova, and V. Costigliola (2014). "Predictive, preventive and personalised medicine as the hardcore of 'Horizon 2020': EPMA position paper". In: *EPMA Journal* 5.1, pp. 1–29 (cit. on p. 6).

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press (cit. on pp. 31, 33).

Gordon, G. and R. Tibshirani (2012). "Karush-kuhn-tucker conditions". In: *Optimization* 10.725/36, p. 725 (cit. on p. 14).

Grigg-Damberger, M. M. (2012). "The AASM Scoring Manual four years later". In: *Journal of Clinical Sleep Medicine* 8.3, pp. 323–332 (cit. on p. 95).

Grossberg, S. (2013). "Recurrent neural networks". In: *Scholarpedia* 8.2, p. 1888 (cit. on p. 28).

Guilleminault, C., A. Tilkian, and W. C. Dement (1976). "The sleep apnea syndromes". In: *Annual review of medicine* 27.1, pp. 465–484 (cit. on p. 97).

Guo, Y. et al. (2021). "Photoplethysmography-based machine learning approaches for atrial fibrillation prediction: a report from the Huawei Heart Study". In: *JACC: Asia* 1.3, pp. 399–408 (cit. on pp. 66, 67, 88).

Gutta, S. and Q. Cheng (2015). "Joint feature extraction and classifier design for ECG-based biometric recognition". In: *IEEE journal of biomedical and health informatics* 20.2, pp. 460–468 (cit. on p. 50).

Haidar, R., I. Koprinska, and B. Jeffries (2017). "Sleep apnea event detection from nasal airflow using convolutional neural networks". In: *International Conference on Neural Information Processing*. Springer, pp. 819–827 (cit. on pp. 122, 123, 130).

Haissaguerre, M. et al. (1998). "Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins". In: *New England Journal of Medicine* 339.10, pp. 659–666 (cit. on p. 62).

Hamilton, J. D. (2020). *Time series analysis*. Princeton university press (cit. on p. 12).

Hannun, A. Y. et al. (2019). "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network". In: *Nature Medicine* 25.1, pp. 65–69 (cit. on p. 66).

Hansen, B. J. et al. (2015). "Atrial fibrillation driven by micro-anatomic intramural re-entry revealed by simultaneous sub-epicardial and sub-endocardial optical mapping in explanted human hearts". In: *European heart journal* 36.35, pp. 2390–2401 (cit. on p. 62).

Hansen, B. J. et al. (2018). "Human atrial fibrillation drivers resolved with integrated functional and structural imaging to benefit clinical mapping". In: *JACC: Clinical Electrophysiology* 4.12, pp. 1501–1515 (cit. on p. 62).

Harper, R. M., V. L. Schechtman, and K. A. Kluge (1987). "Machine classification of infant sleep state using cardiorespiratory measures". In: *Electroencephalography and clinical neurophysiology* 67.4, pp. 379–387 (cit. on p. 110).

Harris, M. et al. (2009). "Obstructive sleep apnea and depression". In: *Sleep medicine reviews* 13.6, pp. 437–444 (cit. on p. 106).

He, K., R. Girshick, and P. Dollár (2019). "Rethinking imagenet pre-training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927 (cit. on p. 117).

He, K. et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on pp. 33, 118).

Health Technology Assessment, S. C. on (2007). "Obstructive Sleep Apnoea Syndrome: A Systematic Literature Review [Internet]." In: *SBU Assessment* 184 (cit. on p. 108).

Heijman, J. et al. (2014). "Cellular and molecular electrophysiology of atrial fibrillation initiation, maintenance, and progression". In: *Circulation research* 114.9, pp. 1483–1499 (cit. on pp. 55, 56).

Heinzer, R. et al. (2015). "Prevalence of sleep-disordered breathing in the general population: the HypnoLaus study". In: *The Lancet Respiratory Medicine* 3.4, pp. 310–318 (cit. on p. 97).

Heinzer, R. C. et al. (2006). "Effect of increased lung volume on sleep disordered breathing in patients with sleep apnoea". In: *Thorax* 61.5, pp. 435–439 (cit. on p. 104).

Himmelreich, J. C. L. et al. (Dec. 2018). "Frequent premature atrial contractions are associated with atrial fibrillation, brain ischaemia, and mortality: a systematic review and meta-analysis". In: *EP Europace* 21.5, pp. 698–707. DOI: 10.1093/europace/euy276 (cit. on p. 89).

Hochreiter, S. and J. Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780 (cit. on p. 29).

Hochreiter, S. et al. (2001). *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies* (cit. on p. 29).

Horner, R. L. (1996). "Motor control of the pharyngeal musculature and implications for the pathogenesis of obstructive sleep apnea". In: *Sleep* 19.10, pp. 827–853 (cit. on p. 107).

Hornik, K., M. Stinchcombe, and H. White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366 (cit. on p. 27).

Huang, G. et al. (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708 (cit. on p. 31).

Huang, L. et al. (2020). "Normalization techniques in training dnns: Methodology, analysis and application". In: *arXiv preprint arXiv:2009.12836* (cit. on p. 116).

Hunyady, L. and K. J. Catt (2006). "Pleiotropic AT1 receptor signaling pathways mediating physiological and pathogenic actions of angiotensin II". In: *Molecular endocrinology* 20.5, pp. 953–970 (cit. on p. 58).

Hutchison, K. N. et al. (2008). "Analysis of sleep parameters in patients with obstructive sleep apnea studied in a hospital vs. a hotel-based sleep center". In: *Journal of clinical sleep medicine* 4.2, pp. 119–122 (cit. on p. 112).

Ibáñez, V., J. Silva, and O. Cauli (2018). "A survey on sleep assessment methods". In: *PeerJ* 6, e4849 (cit. on p. 99).

Igarashi, T. et al. (2012). "Connexin gene transfer preserves conduction velocity and prevents atrial fibrillation". In: *Circulation* 125.2, pp. 216–225 (cit. on p. 59).

Ioffe, S. and C. Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR, pp. 448–456 (cit. on p. 33).

Islam, M. A., S. Jia, and N. D. Bruce (2020). "How much position information do convolutional neural networks encode?" In: *arXiv preprint arXiv:2001.08248* (cit. on p. 36).

Issa, F. G. and C. E. Sullivan (1984). "Upper airway closing pressures in obstructive sleep apnea". In: *Journal of Applied Physiology* 57.2, pp. 520–527 (cit. on p. 97).

Izza, Y., A. Ignatiev, and J. Marques-Silva (2020). "On explaining decision trees". In: *arXiv preprint arXiv:2010.11034* (cit. on pp. 18, 19).

Jackson, M. et al. (2015). "Efficacy of sleep position modification to treat positional obstructive sleep apnea". In: *Sleep medicine* 16.4, pp. 545–552 (cit. on p. 102).

Jain, R., R. Kasturi, B. G. Schunck, et al. (1995). *Machine vision*. Vol. 5. McGraw-hill New York (cit. on p. 12).

Jalali, A. and M. Lee (2019). "Atrial fibrillation prediction with residual network using sensitivity and orthogonality constraints". In: *IEEE Journal of Biomedical and Health Informatics* 24.2, pp. 407–413 (cit. on p. 66).

Janikow, C. Z. (1998). "Fuzzy decision trees: issues and methods". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28.1, pp. 1–14 (cit. on p. 18).

January, C. T. et al. (2014). "2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society". In: *Journal of the American College of Cardiology* 64.21, e1–e76 (cit. on p. 61).

January, C. T. et al. (2019). "2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society". In: *Journal of the American College of Cardiology* 74.1, pp. 104–132 (cit. on p. 61).

Javaheri, S (2000). "Effects of continuous positive airway pressure on sleep apnea and ventricular irritability in patients with heart failure". In: *Circulation* 101.4, pp. 392–397 (cit. on p. 96).

Javaheri, S. and J. Dempsey (2013). "Central sleep apnea". In: *Comprehensive Physiology* 3.1, pp. 141–163 (cit. on p. 96).

Javaheri, S. et al. (2017). "Sleep apnea: types, mechanisms, and clinical cardiovascular consequences". In: *Journal of the American College of Cardiology* 69.7, pp. 841–858 (cit. on p. 95).

Jean-Louis, G. et al. (2008). "Obstructive sleep apnea and cardiovascular disease: role of the metabolic syndrome and its components". In: *Journal of Clinical Sleep Medicine* 4.3, pp. 261–272 (cit. on p. 103).

Jo, Y.-Y. et al. (2021). "Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram". In: *International Journal of Cardiology* 328, pp. 104–110 (cit. on p. 66).

Jokic, R. et al. (1999). "Positional treatment vs continuous positive airway pressure in patients with positional obstructive sleep apnea syndrome". In: *Chest* 115.3, pp. 771–781 (cit. on p. 109).

Jolliffe, I. T. and J. Cadima (2016). "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, p. 20150202 (cit. on p. 12).

Jothi, N., W. Husain, et al. (2015). "Data mining in healthcare–a review". In: *Procedia computer science* 72, pp. 306–313 (cit. on p. 2).

Jukes, E. (2018). "Encyclopedia of machine learning and data mining". In: *Reference Reviews* (cit. on p. 10).

Junno, H. et al. (2004). "Studying the quality of resistance spot welding joints using self-organising maps". In: *Proceedings of the Fourth International ICSC Symposium on Engineering of Intelligent Systems* (cit. on p. 12).

Kamnitsas, K. et al. (2017). "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical image analysis* 36, pp. 61–78 (cit. on p. 32).

Kapur, V. K. et al. (2017). "Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline". In: *Journal of Clinical Sleep Medicine* 13.3, pp. 479–504 (cit. on p. 102).

Kaur, M. and A. Arora (2012). "Classification of ECG signals using LDA with factor analysis method as feature reduction technique". In: *Journal of Medical Engineering & Technology* 36.8, pp. 411–420 (cit. on p. 63).

Ke, G. et al. (2017). "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (cit. on pp. 23, 24, 116, 117).

Kemp, B. and J. Olivan (2003). "European data format 'plus'(EDF+), an EDF alike standard format for the exchange of physiological data". In: *Clinical neurophysiology* 114.9, pp. 1755–1761 (cit. on p. 113).

Kerr, C. R. et al. (2005). "Progression to chronic atrial fibrillation after the initial diagnosis of paroxysmal atrial fibrillation: results from the Canadian Registry of Atrial Fibrillation". In: *American heart journal* 149.3, pp. 489–496 (cit. on p. 55).

Keys, R. (1981). "Cubic convolution interpolation for digital image processing". In: *IEEE transactions on acoustics, speech, and signal processing* 29.6, pp. 1153–1160 (cit. on p. 41).

Khan, A. et al. (2020). "A survey of the recent architectures of deep convolutional neural networks". In: *Artificial intelligence review* 53.8, pp. 5455–5516 (cit. on p. 36).

Khawaja, I. S. et al. (2014). "Can primary care sleep medicine integration work?" In: *The Primary Care Companion for CNS Disorders* 16.2, p. 27142 (cit. on p. 101).

Kim, J. et al. (2016). "Convolutional neural network with biologically inspired retinal structure". In: *Procedia Computer Science* 88, pp. 145–154 (cit. on p. 66).

Kingma, D. P. and J. Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (cit. on p. 117).

Kiranyaz, S., T. Ince, and M. Gabbouj (2015). "Real-time patient-specific ECG classification by 1-D convolutional neural networks". In: *IEEE Transactions on Biomedical Engineering* 63.3, pp. 664–675 (cit. on p. 63).

Knutson, K. L. et al. (2007). "The metabolic consequences of sleep deprivation". In: *Sleep medicine reviews* 11.3, pp. 163–178 (cit. on p. 104).

Kocheril, A. G. and A. A. Sovari (2009). *Practical rapid ECG interpretation (PREI)*. Nova Biomedical Books (cit. on p. 52).

Komada, Y. et al. (2009). "Elevated risk of motor vehicle accident for male drivers with obstructive sleep apnea syndrome in the Tokyo metropolitan area". In: *The Tohoku journal of experimental medicine* 219.1, pp. 11–16 (cit. on p. 107).

Kotsiantis, S. B., D. Kanellopoulos, and P. E. Pintelas (2006). "Data preprocessing for supervised leaning". In: *International journal of computer science* 1.2, pp. 111–117 (cit. on p. 2).

Kouretas, I. and V. Paliouras (2019). "Simplified hardware implementation of the softmax activation function". In: *2019 8th international conference on modern circuits and systems technologies (MOCAST)*. IEEE, pp. 1–4 (cit. on p. 31).

Kraemer, K. H. et al. (2018). "Recurrence threshold selection for obtaining robust recurrence characteristics in different embedding dimensions". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.8, p. 085720 (cit. on pp. 74, 119).

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (cit. on pp. 2, 32).

Kryger, M. H., T. Roth, and W. C. Dement (2010). *Principles and Practice of Sleep Medicine E-Book*. Elsevier Health Sciences (cit. on p. 96).

Kupfahl, C. et al. (2000). "Angiotensin II directly increases transforming growth factor β1 and osteopontin and indirectly affects collagen mRNA expression in the human heart". In: *Cardiovascular research* 46.3, pp. 463–475 (cit. on p. 57).

Kushida, C. A., B. Efron, and C. Guilleminault (1997). "A predictive morphometric model for the obstructive sleep apnea syndrome". In: *Annals of Internal Medicine* 127.8_Part_1, pp. 581–587 (cit. on p. 100).

LaCasse, P. M., W. Otieno, and F. P. Maturana (2019). "A survey of feature set reduction approaches for predictive analytics models in the connected manufacturing enterprise". In: *Applied Sciences* 9.5, p. 843 (cit. on p. 11).

Lan, B. L. et al. (2020). "Flickering of cardiac state before the onset and termination of atrial fibrillation<? A3B2 show [editpick]?>" In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.5, p. 053137 (cit. on pp. 71, 72).

Langley, P et al. (2001). "Can paroxysmal atrial fibrillation be predicted?" In: *Computers in Cardiology 2001*. Vol. 28. IEEE, pp. 121–124 (cit. on p. 66).

Lantelme, P. et al. (2002). "Heart rate: an important confounder of pulse wave velocity assessment". In: *Hypertension* 39.6, pp. 1083–1087 (cit. on pp. 49, 60).

Laporta, R., A. Anandam, and A. A. El-Solh (2012). "Screening for obstructive sleep apnea in veterans with ischemic heart disease using a computer-based clinical decision-support system". In: *Clinical Research in Cardiology* 101.9, pp. 737–744 (cit. on p. 109).

Larsson, S. C., N. Drca, and A. Wolk (2014). "Alcohol consumption and risk of atrial fibrillation: a prospective study and dose-response meta-analysis". In: *Journal of the American College of Cardiology* 64.3, pp. 281–289 (cit. on p. 60).

Lauderdale, D. S. et al. (2008). "Sleep duration: how well do self-reports reflect objective measures? The CARDIA Sleep Study". In: *Epidemiology (Cambridge, Mass.)* 19.6, p. 838 (cit. on p. 101).

Learning, D. (2020). "Deep learning". In: *High-Dimensional Fuzzy Clustering* (cit. on p. 25).

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444 (cit. on pp. 25, 27, 30, 32).

LeCun, Y., Y. Bengio, et al. (1995). "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10, p. 1995 (cit. on p. 30).

Lee, J. et al. (2013). "Time-varying coherence function for atrial fibrillation detection". In: *IEEE Transactions on Biomedical Engineering* 60.10, pp. 2783–2793 (cit. on p. 66).

Leshno, M. et al. (1993). “Multilayer feedforward networks with a nonpolynomial activation function can approximate any function”. In: *Neural networks* 6.6, pp. 861–867 (cit. on p. 28).

Li, D. et al. (2017). “Classification of ECG signals based on 1D convolution neural network”. In: *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, pp. 1–6 (cit. on pp. 63, 64).

Li, Q. et al. (2018a). “Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram”. In: *Physiological measurement* 39.12, p. 124005 (cit. on p. 110).

Li, Z. et al. (2018b). “A novel atrial fibrillation prediction algorithm applicable to recordings from portable devices”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 4034–4037 (cit. on pp. 66, 88).

Lim, H. S. et al. (2017). “Complexity and distribution of drivers in relation to duration of persistent atrial fibrillation”. In: *Journal of the American College of Cardiology* 69.10, pp. 1257–1269 (cit. on p. 62).

Linker, D. T., T. B. Murphy, and A. H. Mokdad (2018). “Selective screening for atrial fibrillation using multivariable risk models”. In: *Heart* 104.18, pp. 1492–1499 (cit. on p. 60).

Linnainmaa, S. (1976). “Taylor expansion of the accumulated rounding error”. In: *BIT Numerical Mathematics* 16.2, pp. 146–160 (cit. on p. 33).

Liu, Y. et al. (2001). “Cephalometric and physiologic predictors of the efficacy of an adjustable oral appliance for treating obstructive sleep apnea”. In: *American Journal of Orthodontics and Dentofacial Orthopedics* 120.6, pp. 639–647 (cit. on p. 102).

Loord, H. and E. Hultcrantz (2007). “Positioner–a method for preventing sleep apnea”. In: *Acta oto-laryngologica* 127.8, pp. 861–868 (cit. on p. 109).

Luo, C. et al. (2020). “Comparison and benchmarking of ai models and frameworks on mobile devices”. In: *arXiv preprint arXiv:2005.05085* (cit. on p. 91).

Mahmud, T., S. A. Fattah, and M. Saquib (2020). “Deeparrnet: An efficient deep cnn architecture for automatic arrhythmia detection and classification from denoised ecg beats”. In: *IEEE Access* 8, pp. 104788–104800 (cit. on p. 66).

Maimon, O. and L. Rokach (2009). “Introduction to knowledge discovery and data mining”. In: *Data mining and knowledge discovery handbook*. Springer, pp. 1–15 (cit. on pp. 2, 4).

Mansouri, M. et al. (2019). “PAC: Privacy-preserving arrhythmia classification with neural networks”. In: *International Symposium on Foundations and Practice of Security*. Springer, pp. 3–19 (cit. on p. 64).

Marin, J. M. et al. (2005). “Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study”. In: *The Lancet* 365.9464, pp. 1046–1053 (cit. on pp. 103, 107).

Markun, L. C. and A. Sampat (2020). “Clinician-focused overview and developments in polysomnography”. In: *Current sleep medicine reports* 6.4, pp. 309–321 (cit. on p. 100).

Marshall, N. S. et al. (2008). “Sleep apnea as an independent risk factor for all-cause mortality: the Busselton Health Study”. In: *Sleep* 31.8, pp. 1079–1085 (cit. on p. 105).

Martin Thoma (2018). *The ROC space for a "better" and "worse" classifier* (cit. on p. 46).

Martinez, A. M. and A. C. Kak (2001). “Pca versus lda”. In: *IEEE transactions on pattern analysis and machine intelligence* 23.2, pp. 228–233 (cit. on p. 12).

Martis, R. J., U. R. Acharya, and L. C. Min (2013). “ECG beat classification using PCA, LDA, ICA and discrete wavelet transform”. In: *Biomedical Signal Processing and Control* 8.5, pp. 437–448 (cit. on p. 65).

Marwan, N. et al. (2007). “Recurrence plots for the analysis of complex systems”. In: *Physics Reports* 438, pp. 237–329 (cit. on p. 39).

Matias, I. et al. (2021). “Prediction of Atrial Fibrillation using artificial intelligence on Electrocardiograms: A systematic review”. In: *Computer Science Review* 39, p. 100334 (cit. on pp. 32, 66, 67).

Mayer-Schönberger, V. and K. Cukier (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt (cit. on p. 1).

McCloskey, S. et al. (2018). “Detecting hypopnea and obstructive apnea events using convolutional neural networks on wavelet spectrograms of nasal airflow”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 361–372 (cit. on pp. 122, 123, 130).

McHugh, M. L. (2012). “Interrater reliability: the kappa statistic”. In: *Biochemia medica* 22.3, pp. 276–282 (cit. on pp. 109, 110).

McNicholas, W. T. (2008). “Diagnosis of obstructive sleep apnea in adults”. In: *Proceedings of the American thoracic society* 5.2, pp. 154–160 (cit. on p. 111).

McNicholas, W., M. Bonsignore, M. C. of EU Cost Action B26, et al. (2007). “Sleep apnoea as an independent risk factor for cardiovascular disease: current evidence, basic mechanisms and research priorities”. In: *European Respiratory Journal* 29.1, pp. 156–178 (cit. on p. 104).

Mendonca, F. et al. (2018). “A review of obstructive sleep apnea detection approaches”. In: *IEEE journal of biomedical and health informatics* 23.2, pp. 825–837 (cit. on pp. 112, 119).

Mikut, R. (2008). *Data Mining in der Medizin und Medizintechnik*. Vol. 22. KIT Scientific Publishing (cit. on pp. 3, 10).

Miller, J. M. et al. (2017). “Clinical benefit of ablating localized sources for human atrial fibrillation: the Indiana University FIRM Registry”. In: *Journal of the American College of Cardiology* 69.10, pp. 1247–1256 (cit. on p. 62).

Mitchell, T. M. (2006). *The discipline of machine learning*. Vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . . (cit. on p. 9).

Mitra, S. K. and Y. Kuo (2006). *Digital signal processing: a computer-based approach*. Vol. 2. McGraw-Hill New York (cit. on p. 40).

Mohebbi, M. and H. Ghassemian (2012). “Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal”. In: *Computer Methods and Programs in Biomedicine* 105.1, pp. 40–49 (cit. on pp. 66, 67, 88).

Montavon, G., G. Orr, and K.-R. Müller (2012). *Neural networks: tricks of the trade*. Vol. 7700. springer (cit. on p. 117).

Moody, G. et al. (2001). "Predicting the onset of paroxysmal atrial fibrillation: The Computers in Cardiology Challenge 2001". In: *Computers in Cardiology 2001*. Vol. 28. IEEE, pp. 113–116 (cit. on pp. 68, 85, 131).

Morillo, C. A. et al. (2014). "Radiofrequency ablation vs antiarrhythmic drugs as first-line treatment of paroxysmal atrial fibrillation (RAAFT-2): a randomized trial". In: *Jama* 311.7, pp. 692–699 (cit. on p. 63).

Mostafa, S. S. et al. (2019). "A systematic review of detecting sleep apnea using deep learning". In: *Sensors* 19.22, p. 4934 (cit. on pp. 5, 32, 112, 118, 119).

Mporas, I. et al. (2015). "Seizure detection using EEG and ECG signals for computer-based monitoring, analysis and management of epileptic patients". In: *Expert systems with applications* 42.6, pp. 3227–3233 (cit. on p. 5).

Nair, V. and G. E. Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Icml* (cit. on p. 30).

Narin, A. et al. (2018). "Early prediction of paroxysmal atrial fibrillation based on short-term heart rate variability". In: *Physica A: Statistical Mechanics and its Applications* 509, pp. 56–65 (cit. on p. 72).

Nattel, S. and D. Dobrev (2012). "The multidimensional role of calcium in atrial fibrillation pathophysiology: mechanistic insights and therapeutic opportunities". In: *European heart journal* 33.15, pp. 1870–1877 (cit. on p. 59).

Nattel, S. and M. Harada (2014). "Atrial remodeling and atrial fibrillation: recent advances and translational perspectives". In: *Journal of the American College of Cardiology* 63.22, pp. 2335–2345 (cit. on p. 55).

Naz, H. and S. Ahuja (2020). "Deep learning approach for diabetes prediction using PIMA Indian dataset". In: *Journal of Diabetes & Metabolic Disorders* 19.1, pp. 391–403 (cit. on p. 1).

Neefs, J et al. (2019). "Body mass index and body fat distribution and new-onset atrial fibrillation: Substudy of the European Prospective Investigation into Cancer and Nutrition in Norfolk (EPIC-Norfolk) study". In: *Nutrition, Metabolism and Cardiovascular Diseases* 29.7, pp. 692–700 (cit. on p. 60).

Nepal, S., R. Ranjan, and K.-K. R. Choo (2015). "Trustworthy processing of healthcare big data in hybrid clouds". In: *IEEE Cloud Computing* 2.2, pp. 78–84 (cit. on p. 1).

Newton, A. C., C. E. Antal, and S. F. Steinberg (2016). "Protein kinase C mechanisms that contribute to cardiac remodelling". In: *Clinical science* 130.17, pp. 1499–1510 (cit. on p. 58).

Ng, A. (2016). "What artificial intelligence can and can't do right now". In: *Harvard Business Review* 9.11 (cit. on p. 2).

Npatchett (2015). *Spatial orientation of EKG leads* (cit. on p. 50).

Oakes, R. S. et al. (2009). "Detection and quantification of left atrial structural remodeling with delayed-enhancement magnetic resonance imaging in patients with atrial fibrillation". In: *Circulation* 119.13, pp. 1758–1767 (cit. on p. 57).

Ogna, A. et al. (2018). "Prevalence and clinical significance of respiratory effort-related arousals in the general population". In: *Journal of clinical sleep medicine* 14.8, pp. 1339–1345 (cit. on p. 97).

Oh, S. L. et al. (2018). "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats". In: *Computers in Biology and Medicine* 102, pp. 278–287 (cit. on p. 74).

Oksenberg, A. et al. (1997). "Positional vs nonpositional obstructive sleep apnea patients: anthropomorphic, nocturnal polysomnographic and multiple sleep latency test data". In: *Chest* 112.3, pp. 629–639 (cit. on p. 106).

Osman, A. M. et al. (2018). "Obstructive sleep apnea: current perspectives". In: *Nature and science of sleep* 10, p. 21 (cit. on p. 97).

Pan, J. and W. J. Tompkins (1985). "A real-time QRS detection algorithm". In: *IEEE Transactions on Biomedical Engineering* 3, pp. 230–236 (cit. on p. 37).

Pandit, S. V. et al. (2005). "Ionic determinants of functional reentry in a 2-D model of human atrial cells during simulated chronic atrial fibrillation". In: *Biophysical journal* 88.6, pp. 3806–3821 (cit. on p. 59).

Park, Y. and H. Cho (2004). "Quality evaluation by classification of electrode force patterns in the resistance spot welding process using neural networks". In: *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 218.11, pp. 1513–1524 (cit. on p. 12).

Parkash, R. et al. (2011). "Approach to the catheter ablation technique of paroxysmal and persistent atrial fibrillation: a meta-analysis of the randomized controlled trials". In: *Journal of cardiovascular electrophysiology* 22.7, pp. 729–738 (cit. on p. 63).

Pascanu, R., T. Mikolov, and Y. Bengio (2013). "On the difficulty of training recurrent neural networks". In: *International conference on machine learning*. PMLR, pp. 1310–1318 (cit. on p. 29).

Patil, S. P. et al. (2007). "Adult obstructive sleep apnea: pathophysiology and diagnosis". In: *Chest* 132.1, pp. 325–337 (cit. on p. 104).

Penzel, T. and K. Kesper (2006). "Physiology of sleep and dreaming". In: *Sleep Apnea*. Vol. 35. Karger Publishers, pp. 13–20 (cit. on p. 93).

Pépin, J. et al. (2012). "The upper airway resistance syndrome". In: *Respiration* 83.6, pp. 559–566 (cit. on p. 97).

Peppard, P. E. et al. (2000). "Longitudinal study of moderate weight change and sleep-disordered breathing". In: *Jama* 284.23, pp. 3015–3021 (cit. on p. 97).

Peppard, P. E. et al. (2013). "Increased prevalence of sleep-disordered breathing in adults". In: *American journal of epidemiology* 177.9, pp. 1006–1014 (cit. on p. 105).

Pillar, G. and P. Lavie (1995). "Assessment of the role of inheritance in sleep apnea syndrome." In: *children* 20, p. 21 (cit. on p. 106).

Pławiak, P. (2018). "Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system". In: *Expert Systems with Applications* 92, pp. 334–349 (cit. on p. 72).

Portes, L. L. et al. (2019). "The reliability of recurrence network analysis is influenced by the observability properties of the recorded time series". In: *Chaos* 29, p. 083101 (cit. on pp. 39, 72).

Prabhu, S. et al. (2017). "A comparison of the electrophysiologic and electroanatomic characteristics between the right and left atrium in persistent atrial fibrillation: is the right atrium a window into the left?" In: *Journal of cardiovascular electrophysiology* 28.10, pp. 1109–1116 (cit. on p. 62).

Prystowsky, E. N. (2000). "Management of atrial fibrillation: therapeutic options and clinical decisions". In: *The American journal of cardiology* 85.10, pp. 3–11 (cit. on p. 65).

Punjabi, N. M. (2008). "The epidemiology of adult obstructive sleep apnea". In: *Proceedings of the American Thoracic Society* 5.2, pp. 136–143 (cit. on pp. 104, 105, 112).

Puvanendran, K. and K. L. Goh (1999). "From snoring to sleep apnea in a Singapore population." In: *Sleep research online: SRO* 2.1, pp. 11–14 (cit. on p. 105).

Qi, X., S. Silvestrov, and T. Nazir (2017). "Data classification with support vector machine and generalized support vector machine". In: *AIP Conference Proceedings*. Vol. 1798. 1. AIP Publishing LLC, p. 020126 (cit. on p. 32).

Quan, S. F. et al. (1997). "The sleep heart health study: design, rationale, and methods". In: *Sleep* 20.12, pp. 1077–1085 (cit. on p. 113).

Radha, M. et al. (2019). "Sleep stage classification from heart-rate variability using long short-term memory neural networks". In: *Scientific reports* 9.1, pp. 1–11 (cit. on p. 111).

Raghunath, S. et al. (2021). "Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation–related stroke". In: *Circulation* 143.13, pp. 1287–1298 (cit. on p. 66).

Raghupathi, W. and V. Raghupathi (2014). "Big data analytics in healthcare: promise and potential". In: *Health information science and systems* 2.1, pp. 1–10 (cit. on p. 1).

Rai, H. M. and K. Chatterjee (2018). "A unique feature extraction using MRDWT for automatic classification of abnormal heartbeat from ECG big data with multilayered probabilistic neural network classifier". In: *Applied Soft Computing* 72, pp. 596–608 (cit. on p. 63).

Rajkumar, A, M Ganesan, and R Lavanya (2019). "Arrhythmia classification on ECG using Deep Learning". In: *2019 5th international conference on advanced computing & communication systems (ICACCS)*. IEEE, pp. 365–369 (cit. on p. 64).

Ramachandran, A. and A. Karuppiah (2021). "A survey on recent advances in machine learning based sleep apnea detection systems". In: *Healthcare*. Vol. 9. 7. MDPI, p. 914 (cit. on pp. 50, 112, 119).

Rechtschaffen, A. (1968). "A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects". In: *Brain information service* (cit. on p. 113).

Redline, S. et al. (1992). "Studies in the genetics of obstructive sleep apnea: familial aggregation of symptoms associated with sleep-related breathing disturbances". In: *American Review of Respiratory Disease* 145.2_pt_1, pp. 440–444 (cit. on p. 106).

Redmond, S. J. and C. Heneghan (2006). "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea". In: *IEEE Transactions on Biomedical Engineering* 53.3, pp. 485–496 (cit. on p. 110).

Refaeilzadeh, P., L. Tang, and H. Liu (2009). "Cross-validation." In: *Encyclopedia of database systems* 5, pp. 532–538 (cit. on p. 43).

Remmers, J. et al. (1978). "Pathogenesis of upper airway occlusion during sleep". In: *Journal of Applied Physiology* 44.6, pp. 931–938 (cit. on p. 97).

Rosenkranz, S. (2004). "TGF-β1 and angiotensin networking in cardiac remodeling". In: *Cardiovascular research* 63.3, pp. 423–432 (cit. on p. 58).

Ruder, S. (2016). "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747* (cit. on p. 27).

Ruff, K. M. and R. V. Pappu (2021). "AlphaFold and implications for intrinsically disordered proteins". In: *Journal of Molecular Biology* 433.20, p. 167208 (cit. on p. 5).

Ruiz-Garcia, A. et al. (2017). "Stacked deep convolutional auto-encoders for emotion recognition from facial expressions". In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1586–1593 (cit. on p. 5).

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536 (cit. on p. 26).

Sabbah, H. N. et al. (1995). "Progression of heart failure: a role for interstitial fibrosis". In: *Cellular Interactions in Cardiac Pathophysiology*. Springer, pp. 29–34 (cit. on p. 57).

Samuel, A. L. (1967). "Some studies in machine learning using the game of checkers. II—Recent progress". In: *IBM Journal of research and development* 11.6, pp. 601–617 (cit. on p. 9).

Sandler, M. et al. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520 (cit. on pp. 33, 34).

Sands, S. A. et al. (2009). "A model analysis of arterial oxygen desaturation during apnea in preterm infants". In: *PLoS computational biology* 5.12, e1000588 (cit. on p. 105).

Santhanakrishnan, R. et al. (2016). "Atrial fibrillation begets heart failure and vice versa: temporal associations and differences in preserved versus reduced ejection fraction". In: *Circulation* 133.5, pp. 484–492 (cit. on p. 60).

Santurkar, S. et al. (2018). "How does batch normalization help optimization?" In: *Advances in neural information processing systems* 31 (cit. on p. 33).

Sarkar, S., D. Ritscher, and R. Mehra (2008). "A detector for a chronic implantable atrial tachyarrhythmia monitor". In: *IEEE Transactions on Biomedical Engineering* 55.3, pp. 1219–1224 (cit. on p. 66).

Sateia, M. J. (2014). "International classification of sleep disorders". In: *Chest* 146.5, pp. 1387–1394 (cit. on pp. 94, 101).

Schäfer, A. M. and H. G. Zimmermann (2006). "Recurrent neural networks are universal approximators". In: *International Conference on Artificial Neural Networks*. Springer, pp. 632–640 (cit. on p. 27).

Schmidhuber, J. (2015). "Deep learning in neural networks: An overview". In: *Neural networks* 61, pp. 85–117 (cit. on pp. 32, 33).

Schnabel, R. B. et al. (2015). "50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study". In: *The Lancet* 386.9989, pp. 154–162 (cit. on p. 60).

Schotten, U. et al. (2016). "Current controversies in determining the main mechanisms of atrial fibrillation". In: *Journal of internal medicine* 279.5, pp. 428–438 (cit. on p. 58).

Schwab, R. J. (2005). "Genetic determinants of upper airway structures that predispose to obstructive sleep apnea". In: *Respiratory physiology & neurobiology* 147.2-3, pp. 289–298 (cit. on p. 104).

Scullin, M. K. and D. L. Bliwise (2015). “Sleep, cognition, and normal aging: integrating a half century of multidisciplinary research”. In: *Perspectives on Psychological Science* 10.1, pp. 97–137 (cit. on p. 111).

Selmer, C. et al. (2012). “The spectrum of thyroid disease and risk of new onset atrial fibrillation: a large population cohort study”. In: *Bmj* 345 (cit. on p. 60).

Selvaraju, R. R. et al. (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626 (cit. on pp. 36, 37).

Sforza, E. et al. (1999). “Pharyngeal critical pressure in patients with obstructive sleep apnea syndrome: clinical implications”. In: *American journal of respiratory and critical care medicine* 159.1, pp. 149–157 (cit. on p. 94).

Sharma, M. et al. (2022). “Automated detection of obstructive sleep apnea in more than 8000 subjects using frequency optimized orthogonal wavelet filter bank with respiratory and oximetry signals”. In: *Computers in Biology and Medicine* 144, p. 105364 (cit. on pp. 122, 130).

Sha’Shonda, L. R., B. H. Cameron, and R. B. Cameron (2019). “Obstructive sleep apnea and perioperative delirium among thoracic surgery intensive care unit patients: Perspective on the STOP-BANG questionnaire and postoperative outcomes”. In: *Journal of Thoracic Disease* 11.Suppl 9, S1292 (cit. on pp. 103, 104).

Shen, Y. et al. (2016). “Risk prediction for cardiovascular disease using ECG data in the China Kadoorie Biobank”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 2419–2422 (cit. on p. 66).

Shorten, C. and T. M. Khoshgoftaar (2019). “A survey on image data augmentation for deep learning”. In: *Journal of big data* 6.1, pp. 1–48 (cit. on p. 35).

Silva, H. P. da et al. (2015). “Off-the-person electrocardiography: performance assessment and clinical correlation”. In: *Health and Technology* 4.4, pp. 309–318 (cit. on p. 50).

Silver, D. et al. (2016). “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587, pp. 484–489 (cit. on p. 5).

Sleep Medicine, A. A. of et al. (2005). “The international classification of sleep disorders, revised”. In: *Chicago, IL: American Academy of Sleep Medicine*, p. 297 (cit. on pp. 96, 97).

Soll, B. and P. George (1985). “Treatment of obstructive sleep apnea with a nocturnal airway-patency appliance”. In: *The New England journal of medicine* 313.6, pp. 386–387 (cit. on p. 108).

Somani, S. et al. (2021). “Deep learning and the electrocardiogram: review of the current state-of-the-art”. In: *EP Europace* (cit. on p. 74).

Somers, V. K. et al. (2008). “Sleep apnea and cardiovascular disease: An American heart association/American college of cardiology foundation scientific statement from the American heart association council for high blood pressure research professional education committee, council on clinical cardiology, stroke council, and council on cardiovascular nursing in collaboration with the national heart, lung, and blood institute national center on sleep disorders research (national institutes of health)”. In: *Journal of the American College of Cardiology* 52.8, pp. 686–717 (cit. on p. 107).

Springenberg, J. T. et al. (2014). "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (cit. on p. 30).

Sridhar, N. et al. (2020). "Deep learning for automated sleep staging using instantaneous heart rate". In: *NPJ digital medicine* 3.1, pp. 1–10 (cit. on p. 111).

Steinberg, B. A. et al. (2015). "Higher risk of death and stroke in patients with persistent vs. paroxysmal atrial fibrillation: results from the ROCKET-AF Trial". In: *European heart journal* 36.5, pp. 288–296 (cit. on pp. 48, 55).

Stiles, M. K., P. Sanders, and D. H. Lau (2018). "Targeting the substrate in ablation of persistent atrial fibrillation: recent lessons and future directions". In: *Frontiers in physiology* 9, p. 1158 (cit. on p. 62).

Stroobandt, R. X., S. S. Barold, and A. F. Sinnaeve (2016). *ECG from Basics to Essentials: Step by Step*. John Wiley & Sons (cit. on p. 52).

Strutz, P. et al. (2018). "Obstructive sleep apnea as an independent predictor of postoperative delirium and pain: protocol for an observational study of a surgical cohort". In: *F1000Research* 7 (cit. on p. 104).

Sullivan, C. et al. (1981). "Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares". In: *The Lancet* 317.8225, pp. 862–865 (cit. on p. 107).

Sun, R. and Y. Wang (2008). "Predicting termination of atrial fibrillation based on the structure and quantification of the recurrence plot". In: *Medical Engineering & Physics* 30.9, pp. 1105–1111 (cit. on p. 72).

Sun, Y. et al. (2014). "Deep learning face representation by joint identification-verification". In: *Advances in neural information processing systems* 27 (cit. on p. 32).

Suthaharan, S. (2016). "Support vector machine". In: *Machine learning models and algorithms for big data classification*. Springer, pp. 207–235 (cit. on p. 16).

Sutherland, K. et al. (2011). "Comparative effects of two oral appliances on upper airway structure in obstructive sleep apnea". In: *Sleep* 34.4, pp. 469–477 (cit. on p. 107).

Tan, M and Q. Le (2021a). "Efficientnetv2: Smaller models and faster training". In: *arXiv preprint arXiv:2104.00298* (cit. on p. 35).

Tan, M. and Q. Le (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR, pp. 6105–6114 (cit. on pp. 34, 35, 117).

— (2021b). "Efficientnetv2: Smaller models and faster training". In: *International Conference on Machine Learning*. PMLR, pp. 10096–10106 (cit. on p. 117).

Tang, J., S. Alelyani, and H. Liu (2014). "Feature selection for classification: A review". In: *Data classification: Algorithms and applications*, p. 37 (cit. on p. 112).

Tascini, G. et al. (1996). *Machine Learning and Perception*. Vol. 23. World Scientific (cit. on p. 4).

Tateno, K and L Glass (2001). "Automatic detection of atrial fibrillation using the coefficient of variation and density histograms of RR and ΔRR intervals". In: *Medical and Biological Engineering and Computing* 39.6, pp. 664–671 (cit. on p. 66).

Thevenot, J., M. B. López, and A. Hadid (2017). "A survey on computer vision for assistive medical diagnosis from faces". In: *IEEE journal of biomedical and health informatics* 22.5, pp. 1497–1511 (cit. on p. 1).

Thong, T. et al. (2004). “Prediction of paroxysmal atrial fibrillation by analysis of atrial premature complexes”. In: *IEEE Transactions on Biomedical Engineering* 51.4, pp. 561–569 (cit. on p. 66).

Tobaldini, E, M Pecis, and N Montano (2014). “Effects of acute and chronic sleep deprivation on cardiovascular regulation”. In: *Arch Ital Biol* 152.2-3, pp. 103–110 (cit. on p. 103).

Trauth, M. H. et al. (2019). “Classifying past climate change in the Chew Bahir basin, southern Ethiopia, using recurrence quantification analysis”. In: *Climate Dynamics* 53.5, pp. 2557–2572 (cit. on p. 72).

Tung, P. and E. Anter (2016). “Atrial fibrillation and sleep apnea: considerations for a dual epidemic”. In: *Journal of atrial fibrillation* 8.6 (cit. on p. 48).

Tzou, H.-A., S.-F. Lin, and P.-S. Chen (2021). “Paroxysmal atrial fibrillation prediction based on morphological variant P-wave analysis with wideband ECG and deep learning”. In: *Computer Methods and Programs in Biomedicine* 211, p. 106396 (cit. on pp. 66, 88).

Uddin, M., C. Chow, and S. Su (2018). “Classification methods to detect sleep apnea in adults based on respiratory and oximetry signals: a systematic review”. In: *Physiological measurement* 39.3, 03TR01 (cit. on p. 109).

Uittenbogaart, S. B. et al. (2018). “Burden of atrial high-rate episodes and risk of stroke: a systematic review”. In: *EP Europace* 20.9, pp. 1420–1427 (cit. on p. 60).

Ullah, A. et al. (2020). “Classification of arrhythmia by using deep learning with 2-D ECG spectral image representation”. In: *Remote Sensing* 12.10, p. 1685 (cit. on p. 65).

Vafaie, M. H., M. Ataei, and H. R. Koofigar (2014). “Heart diseases prediction based on ECG signals’ classification using a genetic-fuzzy system and dynamical model of ECG signals”. In: *Biomedical Signal Processing and Control* 14, pp. 291–296 (cit. on p. 65).

Van Steenkiste, T. et al. (2018). “Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks”. In: *IEEE journal of biomedical and health informatics* 23.6, pp. 2354–2364 (cit. on pp. 122, 130).

Vanderveken, O. M. et al. (2013). “Evaluation of drug-induced sleep endoscopy as a patient selection tool for implanted upper airway stimulation for obstructive sleep apnea”. In: *Journal of clinical sleep medicine* 9.5, pp. 433–438 (cit. on p. 103).

Venkata Jagannath (2017). *The ROC space for a "better" and "worse" classifier* (cit. on p. 24).

Verheule, S. et al. (2004). “Increased vulnerability to atrial fibrillation in transgenic mice with selective atrial fibrosis caused by overexpression of TGF-β1”. In: *Circulation research* 94.11, pp. 1458–1465 (cit. on p. 58).

Vgontzas, A. N. et al. (2001). “Polycystic ovary syndrome is associated with obstructive sleep apnea and daytime sleepiness: role of insulin resistance”. In: *The Journal of Clinical Endocrinology & Metabolism* 86.2, pp. 517–520 (cit. on p. 105).

Vijayan, V. et al. (2021). “Review of wearable devices and data collection considerations for connected health”. In: *Sensors* 21.16, p. 5589 (cit. on p. 5).

Vocaturo, E., D. Perna, and E. Zumpano (2019). "Machine learning techniques for automated melanoma detection". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 2310–2317 (cit. on p. 16).

Voulodimos, A. et al. (2018). "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (cit. on pp. 5, 32).

Wager, S., S. Wang, and P. S. Liang (2013). "Dropout training as adaptive regularization". In: *Advances in neural information processing systems* 26 (cit. on p. 35).

Wakili, R. et al. (2011). "Recent advances in the molecular pathophysiology of atrial fibrillation". In: *The Journal of clinical investigation* 121.8, pp. 2955–2968 (cit. on p. 54).

Wang, J.-S. et al. (2013). "ECG arrhythmia classification using a probabilistic neural network with a feature reduction method". In: *Neurocomputing* 116, pp. 38–45 (cit. on p. 65).

Waring, J., C. Lindvall, and R. Umeton (2020). "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare". In: *Artificial intelligence in medicine* 104, p. 101822 (cit. on p. 12).

Watanabe, T. et al. (2002). "Contribution of body habitus and craniofacial characteristics to segmental closing pressures of the passive pharynx in patients with sleep-disordered breathing". In: *American journal of respiratory and critical care medicine* 165.2, pp. 260–265 (cit. on p. 97).

Weaver, T. E. et al. (2012). "Continuous positive airway pressure treatment of sleepy patients with milder obstructive sleep apnea: results of the CPAP Apnea Trial North American Program (CATNAP) randomized clinical trial". In: *American journal of respiratory and critical care medicine* 186.7, pp. 677–683 (cit. on p. 107).

Weber, K. and C. Brilla (1992). "Factors associated with reactive and reparative fibrosis of the myocardium". In: *Cellular and molecular alterations in the failing human heart*, pp. 291–301 (cit. on p. 57).

Weingarten, T. et al. (2011). "Obstructive sleep apnoea and perioperative complications in bariatric patients". In: *British journal of anaesthesia* 106.1, pp. 131–139 (cit. on pp. 105, 109).

Werbos, P. J. (1990). "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10, pp. 1550–1560 (cit. on p. 29).

Wijffels, M. C. et al. (1995). "Atrial fibrillation begets atrial fibrillation: a study in awake chronically instrumented goats". In: *Circulation* 92.7, pp. 1954–1968 (cit. on p. 55).

Willemen, T. et al. (2013). "An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification". In: *IEEE journal of biomedical and health informatics* 18.2, pp. 661–669 (cit. on p. 110).

Willemen, T. et al. (2015). "Probabilistic cardiac and respiratory based classification of sleep and apneic events in subjects with sleep apnea". In: *Physiological Measurement* 36.10, p. 2103 (cit. on p. 110).

Wilson, R. E. et al. (2020). "The symptom experience of patients with atrial fibrillation before their initial diagnosis". In: *Journal of Cardiovascular Nursing* 35.4, pp. 347–357 (cit. on p. 66).

Worsnop, C. et al. (2000). "Effect of age on sleep onset-related changes in respiratory pump and upper airway muscle function". In: *Journal of Applied Physiology* 88.5, pp. 1831–1839 (cit. on p. 104).

Wright, R. E. (1995). *Logistic regression.* American Psychological Association (cit. on p. 12).

Xiao, M. et al. (2013). “Sleep stages classification based on heart rate variability and random forest”. In: *Biomedical Signal Processing and Control* 8.6, pp. 624–633 (cit. on p. 110).

Xiong, Z. et al. (2018a). “ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network”. In: *Physiological measurement* 39.9, p. 094006 (cit. on p. 64).

Xiong, Z. et al. (2018b). “Fully automatic left atrium segmentation from late gadolinium enhanced magnetic resonance imaging using a dual fully convolutional neural network”. In: *IEEE transactions on medical imaging* 38.2, pp. 515–524 (cit. on p. 32).

Yamashita, R. et al. (2018). “Convolutional neural networks: an overview and application in radiology”. In: *Insights into imaging* 9.4, pp. 611–629 (cit. on p. 31).

Yang, A. and H. Yin (2001). “Prediction of paroxysmal atrial fibrillation by footprint analysis”. In: *Computers in Cardiology*. Vol. 28. IEEE, pp. 401–404 (cit. on p. 66).

Ye, J. et al. (2009). “Stochastic gradient boosted distributed decision trees”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 2061–2064 (cit. on p. 117).

Young, T., P. E. Peppard, and D. J. Gottlieb (2002). “Epidemiology of obstructive sleep apnea: a population health perspective”. In: *American journal of respiratory and critical care medicine* 165.9, pp. 1217–1239 (cit. on p. 103).

Young, T., J. Skatrud, and P. E. Peppard (2004). “Risk factors for obstructive sleep apnea in adults”. In: *Jama* 291.16, pp. 2013–2016 (cit. on p. 104).

Yumino, D. et al. (2013). “Differing effects of obstructive and central sleep apneas on stroke volume in patients with heart failure”. In: *American journal of respiratory and critical care medicine* 187.4, pp. 433–438 (cit. on p. 112).

Zhang, G.-Q. et al. (2018). “The National Sleep Research Resource: towards a sleep data commons”. In: *Journal of the American Medical Informatics Association* 25.10, pp. 1351–1358 (cit. on pp. 113, 131).

Zhang, H. et al. (2015). “A novel quality evaluation method for resistance spot welding based on the electrode displacement signal and the Chernoff faces technique”. In: *Mechanical Systems and Signal Processing* 62, pp. 431–443 (cit. on p. 12).

Zhang, H. et al. (2021). “Recurrence Plot-Based Approach for Cardiac Arrhythmia Classification Using Inception-ResNet-v2”. In: *Frontiers in Physiology* 12, p. 648950 (cit. on p. 72).

Zhao, E. et al. (2018). “Association between sleep apnea hypopnea syndrome and the risk of atrial fibrillation: a meta-analysis of cohort study”. In: *BioMed research international* 2018 (cit. on p. 60).

Zhao, J. et al. (2012). “An image-based model of atrial muscular architecture: effects of structural anisotropy on electrical activation”. In: *Circulation: Arrhythmia and Electrophysiology* 5.2, pp. 361–370 (cit. on p. 63).

Zhou, Z.-H. (2021). *Machine learning*. Springer Nature (cit. on p. 11).

Zihlmann, M., D. Perekrestenko, and M. Tschannen (2017). “Convolutional recurrent neural networks for electrocardiogram classification”. In: *2017 Computing in Cardiology (CinC)*. IEEE, pp. 1–4 (cit. on p. 64).

Zong, W, R Mukkamala, and R. Mark (2001). "A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis". In: *Computers in Cardiology 2001*. Vol. 28. IEEE, pp. 125–128 (cit. on p. 66).

Zou, R. et al. (2005). "Substrate size as a determinant of fibrillatory activity maintenance in a mathematical model of canine atrium". In: *American Journal of Physiology-Heart and Circulatory Physiology* 289.3, H1002–H1012 (cit. on pp. 55, 57).